What did we learn in Artificial Intelligence I?

2. We looked at how to choose a sequence of actions to achieve a goal using search, adversarial search (game-playing), logical inference (situation calculus), and planning.
   - All these approaches suffer in the same way as inference.
   - So all benefit from considering uncertainty.
   - All implicitly deal with time. How is this possible under uncertainty?
   - All tend to be trying to reach goals, but these may also be uncertain.

Utility theory is used to assign preferences.
Decision theory combines probability theory and utility theory.
A rational agent should act in order to maximise expected utility as time passes.

We need to look at other ways of learning.
We need to introduce uncertainty into learning.
We need to consider wider applications of learning.
Artificial Intelligence: what are we going to learn now?

What are we going to learn now?

In moving from logic to probability:

- We replace the knowledge base by a probability distribution that represents our beliefs about the world.
- We replace the task of logical inference with the task of computing conditional probabilities.

Both of these changes turn out to be considerably more complex than they sound.

Bayesian networks and Markov random fields allow us to represent probability distributions.
Various algorithms can be used to perform efficient inference.

General knowledge representation and inference: the BIG PICTURE

The current approach to uncertainty in AI can be summed up in a few sentences:
Everything of interest in the world is a random variable. The probabilities associated with RVs summarize our uncertainty.

The world: \( V = \{V_1, V_2, \ldots, V_n\} \)

If the \( n \) RVs \( V = \{V_1, V_2, \ldots, V_n\} \) represent everything of interest, then our knowledge base is the joint distribution

\[
\Pr (V) = \Pr (V_1, V_2, \ldots, V_n)
\]

General knowledge representation and inference: the BIG PICTURE

Say we have observed the values of a subset \( O = \{O_1, O_2, \ldots, O_m\} \) of \( m \) RVs.
In other words, we know that \( (O_1 = o_1, O_2 = o_2, \ldots, O_m = o_m) \).

Also, say we are interested in some subset \( Q \) of \( k \) query variables.

The world: \( V = \{V_1, V_2, \ldots, V_n\} \)

Then inference corresponds to computing a conditional distribution

\[
\Pr (Q|o_1, o_2, \ldots, o_m)
\]

General knowledge representation and inference: the BIG PICTURE

The latent variables \( L \) are all the RVs not in the sets \( Q \) or \( O \).

The world: \( V = \{V_1, V_2, \ldots, V_n\} \)

Latent variables \( L = \{L_1, L_2, \ldots\} \)

To compute a conditional distribution from a knowledge base \( \Pr (V) \) we have to sum over the latent variables

\[
\Pr (Q|o_1, o_2, \ldots, o_m) = \sum_L \Pr (Q, L|o_1, o_2, \ldots, o_m)
= \frac{1}{Z} \sum_L \Pr (Q, L, o_1, o_2, \ldots, o_m)
\]

Knowledge base

Bayes’ theorem tells us how to update an inference when new information is available.

For example, if we now receive a new observation \( O' = o' \) then

\[
\Pr(Q' | o_1, o_2, \ldots, o_m) = \frac{1}{Z} \Pr(o' | Q, o_1, o_2, \ldots, o_m) \Pr(Q | o_1, o_2, \ldots, o_m)
\]

After \( O' \) observed

Before \( O' \) observed

How can we get around this?

1. You can be clever about representing \( \Pr(V) \) to avoid storing all \( O(2^n) \) numbers.
2. You can take that a step further and exploit the structure of \( \Pr(V) \) in specific scenarios to get good time-complexity.
3. You can do approximate inference.

We’ll be looking at all three…

General knowledge representation and inference: the BIG PICTURE

By addressing AI using Bayesian Inference in this way, in addition to general methods for making inferences:

- We get rigorous methods for supervised learning.
- We get one of the most unreasonably effective ideas in computer science: the hidden Markov model.
- We get methods for unsupervised learning.

Bayesian supervised learning provides a (potentially) optimal method for supervised learning.

Hidden Markov models allow us to infer (probabilistically) the state of the world as time passes.

Mixture models form the basis of probabilistic methods for unsupervised learning.
Artificial Intelligence: what are we going to learn now?

Putting it all together...

Ideally we want an agent to be able to:

- Explore the world to see how it works.
- Use the resulting knowledge to form a plan of how to act in the future.
- Achieve both, even when the world is uncertain.

In essence reinforcement learning algorithms allow us to do this. In practice they often employ supervised learners as a subsystem.

Books

Books recommended for the course:

I suggest you make use of the recommended text for Artificial Intelligence I:


and supplement it with one of the following:


The latter is more comprehensive and goes beyond this course. Further recommended books, covering specific areas in greater detail, can be found on the course web site.

What have we done so far?

We’re going to begin with a review of the material on supervised learning from *Artificial Intelligence I*.

Evil Robot hates kittens, and consequently wants to build a kitten detector. He thinks he can do this by measuring cuteness and furryness.

Provided he has some examples labelled as kitten or not kitten...

...this seems sufficient to find a region that identifies kittens.
What have we done so far?

Of course, when put into practice…

. . . some non-kittens will be labelled as kittens.

And conversely…

. . . some kittens will be labelled as non-kittens.

Kinds of learning: supervised learning

Supervised learning:

We have $m$ vectors $x_1, x_2, \ldots, x_m$ each in $\mathbb{R}^n$.

We have corresponding labels $\{y_1, y_2, \ldots, y_m\}$ each in a set $Y$.

We wish to find a hypothesis $h: \mathbb{R}^n \rightarrow Y$ that can be used to predict $y$ from $x$.

This may itself be defined by a vector $w$ of weights.

To make the latter point clear the hypothesis will be written $h_w(x)$.

If it can do this well it generalizes.

- If $Y = \mathbb{R}$ or some other set such that the output can be regarded as continuous then we’re doing regression.
- If $Y$ has a finite number $K$ of categories, so $Y = \{c_1, c_2, \ldots, c_K\}$ then we are doing classification.
- In the case of classification, we might alternatively treat $Y$ as a random variable (RV), and find a hypothesis $h_w: \mathbb{R}^n \rightarrow [0, 1]$ of the form $h_w(x) = \Pr(Y = c_i | x)$.

The *key issue* is to choose the correct degree of complexity.
What have we done so far?

The *training data* is \( s = [(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)] \).

Fit a polynomial
\( h_w(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d \)
by choosing the weights \( w_i \) to minimize
\[
E(w) = \frac{1}{2} \sum_{i=1}^{m} (y_i - h_w(x_i))^2.
\]

The degree \( d \) sets how complex the fitted function can be.

What have we done so far?

If we use a very simple function \( \sigma(x) = x \) then we’re back to polynomials with \( d = 1 \) and now
\[
E(w) = \frac{1}{2} \sum_{i=1}^{m} (y_i - w^T x_i)^2.
\]

If we can find the gradient \( \frac{\partial E(w)}{\partial w} \) of \( E(w) \) then we can minimize the error using gradient descent
\[
w_{t+1} = w_t - \lambda \frac{\partial E(w)}{\partial w} \bigg|_{w_t}
\]

What have we done so far?

Real problems tend to have more than 1 input.
We can solve problems like this using a *perceptron*:

The trick is the same: select the weights \( w_i \) to minimize some measure of error \( E(w) \) on some training examples.

Gradient descent: the simplest possible method for minimizing such functions:

Take small steps downhill until you reach the minimum.

But remember: there might be many minima.
Some minima might be local and some global.
The step size matters.
What have we done so far?

For a perceptron with $\sigma(x) = x$ this is easy:

$$\frac{\partial E(w)}{\partial w_j} = \frac{1}{2} \sum_{i=1}^{m} (y_i - w^T x_i)^2$$

$$= \sum_{i=1}^{m} \left( y_i - w^T x_i \right) \frac{\partial}{\partial w_j} \left( -w^T x_i \right)$$

$$= -\sum_{i=1}^{m} (y_i - w^T x_i) x_j^{(i)}$$

where $x_j^{(i)}$ is the $j$th element of $x_i$. So:

$$\frac{\partial E(w)}{\partial w} = -\sum_{i=1}^{m} (y_i - w^T x_i) x_i$$

The multilayer perceptron

Real problems tend also to be nonlinear.

We can combine perceptrons to make a multilayer perceptron:

Here, each node is a perceptron and each edge has a weight attached.

The multilayer perceptron

The network computes a function $h_w(x)$.

The trick remains the same: minimize an error $E(w)$.

We do that by gradient descent

$$w_{t+1} = w_t - \lambda \frac{\partial E(w)}{\partial w} \bigg|_{w_t}$$

This can be achieved using backpropagation.

Backpropagation is just a method for computing $\frac{\partial E(w)}{\partial w}$.

It’s needed because we’re doing gradient descent

In supervised learning, you can get quite a long way using a multilayer perceptron.

If you understand backpropagation, you already know the key idea needed for stuff involving the word ‘deep’.

But this is a long way from being the full story.
Kinds of learning: unsupervised learning

What if we have no labels?

Unsupervised learning: we have $m$ vectors $x_1, x_2, \ldots, x_m$ each in $\mathbb{R}^n$...

...and we want to find some regularity.

Kinds of learning: semi-supervised learning

Semi-supervised learning: we have the same labelled data as for supervised learning, but...

...in addition a further $m'$ input vectors $x'_1, \ldots, x'_{m'}$.

We want to use the extra information to improve the hypothesis obtained.

Kinds of learning: reinforcement learning

What if we want to learn from rewards rather than labels?

Reinforcement learning works as follows.

1. We are in a state and can perform an action.
2. When an action is performed we move to a new state and receive a reward. (Possibly zero or negative.)
3. New states and rewards can be uncertain.
4. We have no knowledge in advance of how actions affect either the new state or the reward.
5. We want to learn a policy. This tells us what action to perform in any state.
6. We want to learn a policy that in some sense maximizes reward obtained over time.

Note that this can be regarded as a form of planning.

Matrix notation

We denote by $\mathbb{R}^n$ the set of $n$-dimensional vectors of reals, and by the set $\mathbb{R}^{m \times n}$ the set of $m$ (rows) by $n$ (columns) matrices of reals.

Vectors are denoted using lower-case bold and matrices in upper-case bold.

It is conventional to assume that vectors are column vectors and to denote the transpose using superscripted $T$. So for $x \in \mathbb{R}^n$ we write

$$x^T = [x_1, x_2, \cdots, x_n]$$

and for $X \in \mathbb{R}^{m \times n}$ we write

$$X = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1n} \\
  x_{21} & x_{22} & \cdots & x_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{m1} & x_{m2} & \cdots & x_{mn}
\end{bmatrix}$$

Denote by $X_{i*}$ and $X_{*j}$ the $i$th row and $j$th column of $X$ respectively.
Matrix notation

If we have \( m \) vectors \( x_1, x_2, \ldots, x_m \) then the \( j \)th element of the \( i \)th vector is \( x_{ij} \).

We may also form the matrix

\[
X = \begin{bmatrix}
  x_1^T \\
  x_2^T \\
  \vdots \\
  x_m^T
\end{bmatrix} = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1m} \\
  x_{21} & x_{22} & \cdots & x_{2m} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{m1} & x_{m2} & \cdots & x_{mm}
\end{bmatrix}
\]

Similarly we can write

\[
X^T = [x_1 \ x_2 \ \cdots \ x_m]
\]

The identity matrix is as usual

\[
I = \begin{bmatrix}
  1 & 0 & \cdots & 0 \\
  0 & 1 & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & 1
\end{bmatrix}
\]

The inverse of \( X \) is \( X^{-1} \) and its determinant is \( |X| \).

General notation

An RV can take on one of a set of values. For example, \( X \) is an RV with values \( \{x_1, x_2, \ldots, x_n\} \).

By convention random variables (RVs) are denoted using upper-case and their values using lower-case.

The probability that \( X \) takes a specific value \( x \in \{x_1, x_2, \ldots, x_n\} \) is \( \Pr(X = x) \).

This will generally be abbreviated to just \( \Pr(x) \).

Sometimes we need to sum over all possible values. We write this using the usual notation. So for example the expected value of \( X \) is

\[
E[X] = \sum_{x \in X} x \Pr(x) = \sum_{x \in X} x \Pr(X).
\]

We extend this to vector-valued RVs in the obvious way.

So for example we might define an RV \( X \) taking values in \( \mathbb{R}^n \) and refer to a specific value \( x \in \mathbb{R}^n \).

(But remember: asking about something like \( \Pr(X = x) \) now makes little sense if \( x \in \mathbb{R}^n \).

General notation for supervised learning

• Inputs are in \( n \) dimensions and are denoted by

\[
x^T = [x_1 \ x_2 \ \cdots \ x_n]
\]

Each element \( x_i \) is a feature.

• A training sequence has \( m \) inputs. The \( m \) inputs are \( x_1, \ldots, x_m \) and can be collected into the matrix

\[
X = \begin{bmatrix}
  x_1^T \\
  x_2^T \\
  \vdots \\
  x_m^T
\end{bmatrix}
\]

• The labels in the training sequence are denoted by

\[
y^T = [y_1 \ y_2 \ \cdots \ y_m]
\]

with each \( y_i \) a set \( Y \) depending on the type of problem.
Machine Learning and Bayesian Inference

Major subject number one:

Making learning probabilistic.

It will turn out that in order to talk about optimal methods for machine learning we’ll have to put it into a probabilistic context.

As a bonus, this leads to a much better understanding of what happens when we choose weights by minimizing an error function.

And it turns out that choosing weights in this way is suboptimal . . .

. . . although, intriguingly, that’s not a reason not to do it.

Probabilistic models for generating data

I’m going to start with a very simple, but very informative approach.

Typically, we can think of individual examples as being generated according to some distribution \( p(X, Y) \).

We generally make the simplifying assumption that examples are independent and identically distributed (iid). Thus the training data \( s^T = [(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)] \) represents \( m \) iid samples from the relevant distribution.

As the examples are iid we can write

\[
p(s) = \prod_{i=1}^{m} p(x_i, y_i).
\]

Example: simple regression

Here’s how I generated the regression data for the initial examples:

We have spoken of an unknown underlying function \( f \) used to generate the data. In fact, this is the hypothesis \( h_w \) that we want to identify by choosing \( w \).

I chose \( h_w \) to be a polynomial with parameters \( w \) — this is the dashed blue line.

So in fact the unknown function is \( h_w(x) \), emphasizing that \( w \) determines a specific function \( f \).

Remember: you don’t know what \( w \) is: you need to identify it by analysing \( s \).

The Normal Distribution

In 1 dimension \( \mathcal{N}(\mu, \sigma^2) \) is

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)
\]

with mean \( \mu \) and variance \( \sigma^2 \).
Example: simple regression

To make $s$:

For the $i$th example:

1. I sampled $x_i$ according to the uniform density on $[0, 3]$. So there is a distribution $p(x)$.
2. I computed the value $h_w(x_i)$.
3. I sampled $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 0.1$ and formed $y_i = h_w(x_i) + \epsilon_i$.

Combining steps 2 and 3 gives you $p(y_i|x_i, w)$.

\[
p(y_i|x_i, w) = \mathcal{N}(h_w(x_i), \sigma^2)
= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - h_w(x_i))^2\right).
\]

The likelihood function

The quantity $p(y_i|x_i, w)$ is important: it is known as the likelihood. You will sometimes see it re-arranged and written as the likelihood function

\[
L(w|x_i, y_i) = p(y_i|x_i, w).
\]

Note that its form depends on how you model the data. There are different likelihood functions depending on what assumptions you make.

Now let’s image $w$ is fixed (but hidden!) from the outset and extend the likelihood to the whole data set $s$.

The likelihood for the full data set is:

\[
p(s|w) = \prod_{i=1}^{m} p(x_i, y_i|w)
= \prod_{i=1}^{m} p(y_i|x_i, w)p(x_i|w)
= \prod_{i=1}^{m} p(y_i|x_i, w)p(x_i)
\]

The last step involves the reasonable assumption that $x_i$ itself never depends on $w$.

Maximizing likelihood

This expression, roughly translated, tells us how probable the data $s$ would be if a particular vector $w$ had been used to generate it.

This immediately suggests a way of choosing $w$:

Choose

\[
w_{\text{opt}} = \arg \max_w p(s|w).
\]

This is called (surprise surprise) a maximum likelihood algorithm.

How would we solve this maximization problem?
Maximizing likelihood

This is surprisingly easy:

\[ w_{\text{opt}} = \arg\max_w p(s|w) \]

\[ = \arg\max_w \left( \prod_{i=1}^{m} p(y_i|x_i, w) p(x_i) \right) \]

\[ = \arg\max_w \left( \sum_{i=1}^{m} \log p(y_i|x_i, w) + \sum_{i=1}^{m} \log p(x_i) \right) \]

\[ = \arg\max_w \sum_{i=1}^{m} \log p(y_i|x_i, w) \]

We’ve used three standard tricks:

1. To maximize something you can alternatively maximize its logarithm.
2. Logarithms turn products into sums.
3. You can drop parts of the expression that don’t depend on the variable you’re maximizing over

Maximizing likelihood

It’s worth reflecting on that for a moment:

- Originally, we plucked

\[ E(w) = \sum_{i=1}^{m} (y_i - h_w(x_i))^2 \]

pretty much out of thin air because it seemed to make sense.

- We’ve just shown that hidden inside it is an assumption; that noise in the data is Gaussian.

- We’ve also uncovered a second assumption: that maximizing the likelihood is the right thing to do.

Of course, assumptions such as these are open to question. . .

Maximizing likelihood

Then:

\[ w_{\text{opt}} = \arg\max_w \left( \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} (y_i - h_w(x_i))^2 \right) \]

\[ = \arg\min_w \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - h_w(x_i))^2 \]

So we’ve just shown that:

To choose \( w \) by maximizing likelihood . . .

. . . we minimize the sum of squared errors.

Result!

Maximizing the posterior

For example, what if we don’t regard \( w \) as being fixed in advance but instead make it an RV as well?

That means we need a distribution \( p(w) \), generally known as the prior on \( w \). How about our old friend the normal? In \( d \) dimensions \( w \sim \mathcal{N}(\mu, \Sigma) \) looks like

\[ p(w) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} \exp \left( -\frac{1}{2} (w - \mu)^T \Sigma^{-1} (w - \mu) \right) \]

with mean vector \( \mu \) and covariance matrix \( \Sigma \).
Maximizing the posterior

This suggests another natural algorithm for choosing a good \( w \), called the maximum a posteriori (MAP) algorithm. Let’s choose \( w \sim \mathcal{N}(0, \lambda^{-1}) \) so

\[
p(w) = \frac{1}{\sqrt{2\pi\lambda}} \exp \left( -\frac{\lambda}{2} w^T w \right)
\]

Then

\[
w_{\text{opt}} = \arg\max_w p(w|s)
= \arg\max_w \frac{p(s|w)p(w)}{p(s)}
= \arg\max_w [\log p(s|w) + \log p(w)]
\]

The maximization of \( \log p(s|w) \) proceeds as before, and we end up with

\[
w_{\text{opt}} = \arg\min_w \left[ \frac{1}{2\sigma^2} \sum_{i=1}^m \left( (y_i - h_w(x_i))^2 \right) + \frac{\lambda}{2} \|w\|^2 \right],
\]

The likelihood for classification problems

For regression problems just adding noise to the labels seems reasonable:

The likelihood \( p(y|x, w) \) is in fact a density and can take any value in \( \mathbb{R} \) as long as the density is non-negative and integrates to 1.

(Think of the Gaussian as usual . . . ).

But what about for classification problems?

Maximizing the posterior

This appears in the literature under names such as weight decay.

- It was often proposed, again on the basis that it seemed sensible, as a sensible-looking way of controlling the complexity of \( h_w \).
- The idea was to use \( \lambda \) to achieve this.
- We’ll be seeing later how to do this.

Once again, we can now see that it hides certain assumptions.

In addition to the assumptions made by maximum likelihood:

- We are assuming that some kinds of \( w \) are more likely than others.
- We are assuming that the distribution governing this is Gaussian.

And again, these assumptions may or may not be appropriate.

The likelihood for classification problems

For simplicity, let’s just consider two-class classification with labels in \( \{0, 1\} \).

For a classification problem the likelihood is now a distribution \( \Pr(Y|x, w) \). It has two non-negative values, and

\[
\Pr(Y = 1|x, w) = 1 - \Pr(Y = 0|x, w).
\]

So you can’t just add noise to the underlying \( h_w \).

Fix \( \theta \) to define the likelihood as

\[
\Pr(Y = 1|x, w) = \sigma_\theta(h_w(x))
\]

and use something like

\[
\sigma_\theta(z) = \frac{1}{1 + \exp(-\theta z)}
\]

to impose the above property.
The likelihood for classification problems

So: if we’re given a training sequence $s$, what is the probability that it was generated using some $w$?

For an example $(x, y)$

$$\Pr(Y | x, w) = \begin{cases} \sigma_\theta(h_w(x)) & \text{if } Y = 1 \\ 1 - \sigma_\theta(h_w(x)) & \text{if } Y = 0 \end{cases}$$

Consequently when $Y$ has a known value we can write

$$\Pr(Y | x, w) = [\sigma_\theta(h_w(x))]^Y [1 - \sigma_\theta(h_w(x))]^{(1-Y)}$$

If we assume that the examples are iid then the probability of seeing the labels in a training sequence $s$ is straightforward.

The likelihood is now

$$p(s|w) = \prod_{i=1}^{m} p(y_i|x_i, w)p(x_i)$$

$$= \prod_{i=1}^{m} [\sigma_\theta(h_w(x_i))]^{y_i} [1 - \sigma_\theta(h_w(x_i))]^{(1-y_i)} p(x_i)$$

where the first line comes straight from an earlier slide.

Note that:

- Whereas previously we had the noise variance $\sigma^2$ we now have the parameter $\theta$. Both serve a similar purpose.
- From this expression we can directly derive maximum-likelihood and MAP learning algorithms for classifiers.
We have so far concentrated throughout our coverage of machine learning on choosing a single hypothesis. Are we asking the right question though?

Ultimately, we want to generalise. This means finding a hypothesis that works well for previously unseen examples. That means we have to define what good generalization is and ask what method might do it the best.

Is it reasonable to expect a single hypothesis to provide the optimal answer? We need to look at what the optimal solution to this kind of problem might be...

The ability to specify losses in this way can be important, for example:

- In learning to diagnose cancer we might always assign a loss of 0 when the action is 'say the patient has cancer', assuming the patient does in fact have cancer.
- A loss of 0 is also appropriate if we take action 'say the patient is healthy' when the patient actually is healthy.
- The subtlety appears when our action is wrong. We should probably assign a bigger penalty (higher loss) if we tell a patient they're healthy when they're sick, than if we tell a patient they're sick when they're healthy.

Having extra actions can also be useful.

Also, sometimes we want the system to defer to a human.

There is essentially nothing new here. The actions can be thought of as saying 'assign $x$ to class $c_i$' and so on. We may have further actions, for example the action 'I don't know how to classify $x$'.

There is also a loss $\lambda_{ij}$ associated with taking action $a_i$ when the class is in fact $c_j$.

Sometimes we will need to write $\lambda(a_i, c_j)$ for $\lambda_{ij}$.

Say we can further model the world as follows:

- Classes have probabilities $\Pr(C)$ of occurring.
- There are probability densities $p(X|C)$ for seeing $X$ when the class is $C$.

So now we have a slightly different, though equivalent way of modelling how labelled examples are generated: nature chooses classes at random using $\Pr(C)$ and selects a vector using $p(X|C)$.

As usual Bayes rule tells us that

$$\Pr(C|X) = \frac{1}{Z}p(X|C)\Pr(C)$$

where

$$Z = p(X) = \sum_{i=1}^{K} p(X|c_i)\Pr(c_i).$$
Bayesian decision theory

Say nature shows us $x$ and we take action $a_i$.

If we always take action $a_i$ when we see $x$ then the average loss on seeing $x$ is

$$R(a_i|x) = \mathbb{E}_{c \sim p(C|x)} [\lambda_{ij} | x] = \sum_{j=1}^{K} \lambda_{ij} \Pr(c_j | x).$$

The quantity $R(a_i|x)$ is called the conditional risk.

Note that this particular $x$ is fixed.

Now say we have a decision rule $D : \mathbb{R}^d \rightarrow \{a_1, \ldots, a_L\}$ telling us what action to take on seeing any $x \in \mathbb{R}^d$.

The average loss, or risk, is

$$R = \mathbb{E}_{(x,c) \sim p(X,C)} [\lambda(D(x),c)] = \mathbb{E}_{x \sim p(X)} \left[ \mathbb{E}_{c \sim \Pr(C|x)} [\lambda(D(x),c) | x] \right] = \mathbb{E}_{x \sim p(x)} [R(D(x)|x)] = \int R(D(x)|x)p(x)dx.$$

Here we have used the standard result from probability theory that

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

(See the supplementary notes for a proof.)

Bayesian decision theory

Clearly the risk is minimised by the following decision rule:

Given any $x \in \mathbb{R}^d$: $D(x)$ outputs the action $a_i$ that minimises $R(a_i|x)$

This $D$ provides us with the minimum possible risk, or Bayes risk $R^\star$.

The rule specified is called the Bayes decision rule.

Example: minimum error rate classification

In supervised learning our aim is often to work in such a way that we minimise the probability of making an error when predicting the label for a previously unseen example.

What loss should we consider in these circumstances?

From basic probability theory, we know that for any event $E$

$$\Pr(E) = \mathbb{E} [I[E]]$$

where $I[\cdot]$ denotes the indicator function

$$I[E] = \begin{cases} 1 & \text{if } E \text{ happens} \\ 0 & \text{otherwise} \end{cases}.$$

(See the supplementary notes for a proof.)
Example: minimum error rate classification

So if we are addressing a supervised learning problem with

- $K$ classes $\{c_1, \ldots, c_K\}$.
- $L = K$ corresponding actions $\{a_1, \ldots, a_K\}$.
- We interpret action $a_i$ as meaning ‘the input is in class $c_i$’.
- The loss is defined as

$$
\lambda_{ij} = \begin{cases} 
1 & \text{if } i \neq j \\
0 & \text{otherwise}
\end{cases}
$$

then...

The risk $R$ is

$$
R = \mathbb{E}_{(x,c) \sim p(X,C)} \left[ \lambda(D(x),C) \right] \\
= \Pr(D(x) \text{ chooses the wrong class})
$$

so the Bayes decision rule minimises the probability of error.

Bayesian supervised learning

But what about the training sequence $s$?

Shouldn’t the Bayes optimal classifier depend on that as well?

- Yes, it should if there is uncertainty about the mechanism used to generate the data.
- (All of the above assumes that the mechanism is fixed, so seeing examples has no effect on the optimal classifier.)
- In our case we don’t know what underlying $h$ was used. There is a prior $p(h)$.
- If you carry through the above derivation letting the conditional risk be conditional on both $x$ and $s$ then you find that...
- ...to minimize error probability you should maximize $\Pr(C|x,s)$.

You should now work through the related exercise.

Example: minimum error rate classification

What is the Bayes decision rule in this case?

$$
R(a_i|x) = \sum_{j=1}^{K} \lambda_{ij} \Pr(c_j|x) = \sum_{i \neq j} \Pr(c_j|x) = 1 - \Pr(c_i|x)
$$

so $D(x)$ should be the class that maximises $\Pr(C|x)$.

THE IMPORTANT SUMMARY: Given a new $x$ to classify, choosing the class that maximises $\Pr(C|x)$ is the best strategy if your aim is to minimize the probability of error.

Bayesian supervised learning

But the uncertain underlying hypothesis $h$ used to assign classes still doesn’t appear!

Well, we want to maximize $\Pr(C|x,s)$:

$$
\Pr(C|x,s) = \sum_h \Pr(C,h|x,s) = \sum_h \Pr(C|h,x,s) \Pr(h|x,s) = \sum_h \Pr(C|h,x) \Pr(h|s).
$$

Here we have re-introduced $h$ using marginalisation.
Bayesian supervised learning

So our classification should be

\[ C = \arg\max_{C \in \{c_1, \ldots, c_K\}} \sum_{h} \Pr(C|h, x) \Pr(h|s) \]

Of course, when dealing with hypotheses defined by weights \( w \) the sum becomes an integral

\[ C = \arg\max_{C \in \{c_1, \ldots, c_K\}} \int_{W} \Pr(C|w, x) p(w|s) \, dw \]

where \( W \) is the number of weights. The key point:

- You can also write these equations in the form

  \[ C = \arg\max_{C \in \{c_1, \ldots, c_K\}} \mathbb{E}_{h \sim \Pr(h|s)}[\Pr(C|h, x)] \]

- We are not choosing a single \( h \).

- We are averaging the predictions of all possible functions \( h \).

- In doing this we are weighting according to how probable they are.

A word of caution

We know the optimal classifier, so we’ve solved supervised learning right?

WRONG!!!

In practice, solving

\[ C = \arg\max_{C \in \{c_1, \ldots, c_K\}} \mathbb{E}_{h \sim \Pr(h|s)}[\Pr(C|h, x)] \]

is intractible in all but the simplest of cases.

Thou shalt beware Bayesians bearing gifts.

They may well be too good to be true…

Machine Learning and Bayesian Inference

Major subject number two:

The road to Support Vector Machines (SVMs).

It is worth remembering that not all state-of-the-art machine learning is inherently probabilistic.

There is good reason for this: you can almost never actually compute

\[ C = \arg\max_{C \in \{c_1, \ldots, c_K\}} \mathbb{E}_{h \sim \Pr(h|s)}[\Pr(C|h, x)] \]

So before we go any further, let’s see how far it’s possible to get using only linear methods.

This is generally a good idea.

Why? Because linear methods are EASY!

The problem with linear classifiers

Purely linear classifiers or regressors are great for some problems but awful for others:

This example actually killed neural network research for many years.
The kernel trick

One way of getting around this problem is to employ the kernel trick:

Map the data into a bigger space and hope it’s more separable there.
Here, we’ve added one new dimension by introducing a new feature equal to $x_1x_2$.

The kernel trick

Here is a linear hypothesis learned to separate the two classes in the new space.

This was obtained using the Iterative Recursive Least Squares (IRLS) algorithm.
We’ll be deriving this in a moment.

machine Learning Commandments

Thou shalt not rely on toy data.

Resources such as the UCI Machine Learning Repository are there for a very good reason:


Thou shalt not rebrand the kernel trick.

Linear classifiers

We’ve already seen the linear classifier

$h_w(x) = \sigma \left( w_0 + \sum_{i=1}^{n} w_ix_i \right)$

Or $h_w(x) = \sigma (w^T x)$ if we add an extra element having constant value 1 to $x$.

Make it nonlinear by introducing basis functions $\phi$:

$\Phi^T(x) = [\phi_1(x) \phi_2(x) \cdots \phi_k(x)]$

$h_w(x) = \sigma \left( w_0 + \sum_{i=1}^{k} w_i \phi_i(x) \right)$

or assuming there’s a basis function $\phi(x) = 1$

$h_w(x) = \sigma (w^T \Phi(x))$. 
Linear regression

We’ve already seen linear regression. We use $\sigma(x) = x$ and we have training data
$s^T = [(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)]$.

I want to minimize
$$E(w) = \frac{1}{2} \sum_{i=1}^{m} (y_i - h_w(x_i))^2.$$ 

Last year we would have found the gradient of $E(w)$ and used gradient descent
$$w_{t+1} = w_t - \lambda \frac{\partial E(w)}{\partial w} \big|_{w_t}.$$ 

But for linear regression there is an easier way. We can directly solve the equation
$$\frac{\partial E(w)}{\partial w} = 0.$$

Write
$$\Phi = \begin{bmatrix} \Phi^T(x_1) \\ \Phi^T(x_2) \\ \vdots \\ \Phi^T(x_m) \end{bmatrix}$$
so
$$E(w) = \frac{1}{2} (y - \Phi w)^T (y - \Phi w)$$
$$= \frac{1}{2} (y^T y - 2y^T \Phi w + w^T \Phi^T \Phi w)$$
and
$$\frac{\partial E(w)}{\partial w} = \Phi^T \Phi w - \Phi^T y$$

So the optimum solution is obtained by solving
$$\Phi^T \Phi w = \Phi^T y$$
giving
$$w_{opt} = (\Phi^T \Phi)^{-1} \Phi^T y$$
This is the maximum likelihood solution to the problem, assuming noise is Gaussian.

Recall that we can also consider the maximum a posteriori (MAP) solution…

Calculus with matrices

It is much easier to handle this kind of calculation in matrix/vector format than by writing it out in full.

For example, if $a$ and $x$ are both vectors in $\mathbb{R}^n$ we can verify that
$$\frac{\partial a^T x}{\partial x} = \begin{bmatrix} \frac{\partial a^T x}{\partial x_1} \\ \frac{\partial a^T x}{\partial x_2} \\ \vdots \\ \frac{\partial a^T x}{\partial x_n} \end{bmatrix} = a$$
because for each element $x_j$
$$\frac{\partial a^T x}{\partial x_j} = \frac{\partial}{\partial x_j} \left( a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \right) = a_j$$

You should verify for yourself that most standard manipulations involving derivatives carry over directly.

Exercise: Show that if $A \in \mathbb{R}^{n \times n}$ is symmetric then
$$\frac{\partial x^T A x}{\partial x} = 2A x$$
Linear regression: the MAP solution

We saw earlier that to get the MAP solution we minimize the error
\[
E(w) = \frac{1}{2} \sum_{i=1}^{m} ((y_i - h_w(x_i))^2) + \frac{\lambda}{2} ||w||^2.
\]

It is an exercise to show that the solution is:
\[
w_{opt} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^Ty.
\]
This is regularized linear regression or ridge regression.

Iterative re-weighted least squares

What about if we’re classifying rather than doing regression?
We now need to use a non-linear \(\sigma\), typically the sigmoid function, so
\[h_w(x) = \sigma_\theta(w^T\Phi(x)).\]
We saw earlier that to get the maximum likelihood solution we should maximize the likelihood
\[p(s|w) = \prod_{i=1}^{m} [\sigma_\theta(w^T\Phi(x_i))]^{y_i} [1 - \sigma_\theta(w^T\Phi(x_i))]^{(1-y_i)} p(x_i).
\]
Assuming you’ve been completing the exercises you now know that this corresponds to minimizing the error
\[E(w) = -\sum_{i=1}^{m} y_i \log \sigma_\theta(w^T\Phi(x_i)) + (1 - y_i) \log(1 - \sigma_\theta(w^T\Phi(x_i))).\]

Linear regression: the MAP solution

This can make a huge difference.
Revisiting our earlier simple example and training using different values for \(\lambda\):

How can we choose \(\lambda\)? We’ll address this a little later…

Iterative re-weighted least squares

Introducing the extra nonlinearity means we can no longer minimize
\[E(w) = -\sum_{i=1}^{m} y_i \log \sigma_\theta(w^T\Phi(x_i)) + (1 - y_i) \log(1 - \sigma_\theta(w^T\Phi(x_i))).\]
just by computing a derivative and solving. (Sad, but I suggest you get used to it!)
We need to go back to an iterative solution: this time using the Newton-Raphson method.

Given a function \(f: \mathbb{R} \rightarrow \mathbb{R}\), to find where \(f(x) = 0\) iterate as
\[x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}\]
Obviously, to find a minimum we can iterate as
\[x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}\]
This works for 1 dimension. How about many dimensions?
Iterative re-weighted least squares

The Newton-Raphson method generalizes easily to functions of a vector:

To minimize $E: \mathbb{R}^n \rightarrow \mathbb{R}$ iterate as follows:

$$w_{t+1} = w_t - \left[H^{-1}(w_t) \frac{\partial E(w)}{\partial w}\right]w_t.$$ 

Here the Hessian is the matrix of second derivatives of $E(w)$:

$$H_{ij}(w) = \frac{\partial^2 E(w)}{\partial w_i \partial w_j}.$$ 

All we need to do now is to work out the derivatives...
The maximum margin classifier

**Suggestion:** why not drop all this probability nonsense and just do this:

\[ f(x) = w^T x + b \]

Draw the boundary as far away from the examples as possible.

The distance \( \gamma \) is the **margin**, and this is the **maximum margin classifier**.

For \( x' \) on one side of the line \( f(x) = 0 \) we have \( f(x') > 0 \) and on the other side \( f(x') < 0 \).
The maximum margin classifier

Problems:
- Given the usual training data $s$, can we now find a training algorithm for obtaining the weights?
- What happens when the data is not linearly separable?

To derive the necessary training algorithm we need to know something about constrained optimization.

We can address the second issue with a simple modification. This leads to the Support Vector Machine (SVM).

Despite being decidedly “non-Bayesian” the SVM is currently a gold-standard:


Constrained optimization

You are familiar with maximizing and minimizing a function $f(x)$. This is unconstrained optimization.

We want to extend this:

1. Minimize a function $f(x)$ with the constraint that $g(x) = 0$.
2. Minimize a function $f(x)$ with the constraints that $g(x) = 0$ and $h(x) \geq 0$.

Ultimately we will need to be able to solve problems of the form: find $x_{opt}$ such that

$$x_{opt} = \arg\min_x f(x)$$

under the constraints

$$g_i(x) = 0 \text{ for } i = 1, 2, \ldots, n$$

and

$$h_j(x) \geq 0 \text{ for } j = 1, 2, \ldots, m.$$
Constrained optimization

Step 2: solving these equations tells us that the solution is at:

\[(x, y) = (4, -\frac{3}{2})\]

With multiple constraints we follow the same approach, with a Lagrange multiplier for each constraint.

The necessary conditions now require that when \(x'\) is a solution \(\exists \lambda', \alpha'\) such that:

1. \(\frac{\partial L(x', \lambda', \alpha')}{\partial x} = 0\).
2. The equality and inequality constraints are satisfied at \(x'\).
3. \(\alpha' \geq 0\).
4. \(\alpha'_j h_j(x') = 0\) for \(j = 1, \ldots, m\).

These are called the Karush-Kuhn-Tucker (KKT) conditions.

The KKT conditions tell us some important things about the solution.

We will only need to address this problem when the constraints are all inequalities.

How about the full problem? Find

\[x_{\text{opt}} = \arg\min_x f(x) \text{ such that } g_i(x) = 0 \text{ for } i = 1, 2, \ldots, n\]
\[h_j(x) \geq 0 \text{ for } j = 1, 2, \ldots, m\]

The Lagrangian is now

\[L(x, \lambda, \alpha) = f(x) - \sum_{i=1}^{n} \lambda_i g_i(x) - \sum_{j=1}^{m} \alpha_j h_j(x)\]

and the relevant necessary conditions are more numerous.

What we’ve seem so far is called the primal problem.

There is also a dual version of the problem. Simplifying a little by dropping the equality constraints.

1. The dual objective function is

\[\hat{L}(\alpha) = \inf_x L(x, \alpha)\]

2. The dual optimization problem is

\[\max_{\alpha} \hat{L}(\alpha) \text{ such that } \alpha \geq 0\]

Sometimes it is easier to work by solving the dual problem and this allows us to obtain actual learning algorithms.

We won’t be looking in detail at methods for solving such problems, only the minimum needed to see how SVMs work.

The maximum margin classifier

It turns out that with SVMs we get particular benefits when using the kernel trick. So we work, as before, in the extended space, but now with:

\[ f_{w_0}(x) = w_0 + w^T \Phi(x) \]

where

\[ \text{sgn}(z) = \begin{cases} +1 & \text{if } z > 0 \\ -1 & \text{otherwise} \end{cases} \]

Note the following:

1. Things are easier for SVMs if we use labels \{+1, -1\} for the two classes. (Previously we used \{0, 1\}.)
2. It also turns out to be easier if we keep \(w_0\) separate rather than rolling it into \(w\).
3. We now classify using a “hard” threshold \(\text{sgn}\), rather than the “soft” threshold \(\sigma\).

Step 2:

- But we also want the examples to fall on the correct side of the line according to their label.
- Noting that for any labelled example \((x_i, y_i)\) the quantity \(y_i f_{w_0}(x_i)\) will be positive if the resulting classification is correct. . .
- . . . the aim is to solve:

\[ (w, w_0) = \arg\max_{w \neq 0} \left[ \min_i \frac{y_i f_{w_0}(x_i)}{||w||} \right] \]
The maximum margin classifier

**Solution, version 1:** convert to a constrained optimization. For any \( c \in \mathbb{R} \)

\[
fw,w0(x) = 0 \iff w0 + w^T \Phi(x) = 0
\]

\[
\iff cw0 + cw^T \Phi(x) = 0.
\]

That means you can fix \( ||w||^2 \) to be *anything you like*! (Actually, fix \( ||w||^2 \) to avoid a square root.)

**Version 1:**

\[
(w, w0, \gamma) = \arg\max_{w, w0, \gamma} \gamma
\]

subject to the constraints

\[
y_i f_{w, w0}(x_i) \geq \gamma, i = 1, 2, \ldots, m
\]

\[
||w||^2 = 1.
\]

We’ll use the second formulation. (You can work through the first as an exercise.)

The constrained optimization problem is:

Minimize \( \frac{1}{2} ||w||^2 \)

such that

\[
y_i f_{w, w0}(x_i) \geq 1 \text{ for } i = 1, \ldots, m.
\]

Referring back, this means the Lagrangian is

\[
L(w, w0, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{m} \alpha_i (y_i f_{w, w0}(x_i) - 1)
\]

and a necessary condition for a solution is that

\[
\frac{\partial L(w, w0, \alpha)}{\partial w} = 0 \quad \frac{\partial L(w, w0, \alpha)}{\partial w0} = 0.
\]

The maximum margin classifier

**Solution, version 2:** still, convert to a constrained optimization, but instead of fixing \( ||w|| \):

Fix \( \{y_i f_{w, w0}(x_i)\} \) to be *anything you like*!

Version 2:

\[
(w, w0) = \arg\min_{w, w0} \frac{1}{2} ||w||^2
\]

subject to the constraints

\[
y_i f_{w, w0}(x_i) \geq 1, i = 1, 2, \ldots, m.
\]

(This works because maximizing \( \gamma \) now corresponds to *minimizing* \( ||w|| \).)

Working these out is easy:

\[
\frac{\partial L(w, w0, \alpha)}{\partial w} = \frac{\partial}{\partial w} \left( \frac{1}{2} ||w||^2 - \sum_{i=1}^{m} \alpha_i (y_i f_{w, w0}(x_i) - 1) \right)
\]

\[
= w - \sum_{i=1}^{m} \alpha_i y_i \frac{\partial}{\partial w} (w^T \Phi(x_i) + w0)
\]

\[
= w - \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i)
\]

and

\[
\frac{\partial L(w, w0, \alpha)}{\partial w0} = -\frac{\partial}{\partial w0} \left( \sum_{i=1}^{m} \alpha_i y_i f_{w, w0}(x_i) \right)
\]

\[
= -\frac{\partial}{\partial w0} \left( \sum_{i=1}^{m} \alpha_i y_i (w^T \Phi(x_i) + w0) \right)
\]

\[
= -\sum_{i=1}^{m} \alpha_i y_i.
\]
The maximum margin classifier

Equating those to 0 and adding the KKT conditions tells us several things:

1. The weight vector can be expressed as
   \[ w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i) \]
   with \( \alpha \geq 0 \). This is important: we’ll return to it in a moment.

2. There is a constraint that
   \[ \sum_{i=1}^{m} \alpha_i y_i = 0. \]
   This will be needed for working out the dual Lagrangian.

3. For each example
   \[ \alpha_i [y_i f_w(x_i) - 1] = 0. \]

The fact that for each example
\[ \alpha_i [y_i f_w(x_i) - 1] = 0 \]
means that:

Either \( y_i f_w(x_i) = 1 \) or \( \alpha_i = 0 \).

This means that examples fall into two groups.

1. Those for which \( y_i f_w(x_i) = 1 \).
   As the constraint used to maximize the margin was \( y_i f_w(x_i) \geq 1 \) these are the examples that are closest to the boundary.
   They are called support vectors and they can have non-zero weights.

2. Those for which \( y_i f_w(x_i) \neq 1 \).
   These are non-support vectors and in this case it must be that \( \alpha_i = 0 \).

Support vectors:

1. Circled examples: support vectors with \( \alpha_i > 0 \).
2. Other examples: have \( \alpha_i = 0 \).

Remember that
\[ w = \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i). \]

so the weight vector \( w \) only depends on the support vectors.

ALSO: the dual parameters \( \alpha \) can be used as an alternative set of weights. The overall classifier is
\[ h_{w,v_0}(x) = \text{sgn} \left( v_0 + w^T \Phi(x) \right) \]
\[ = \text{sgn} \left( v_0 + \sum_{i=1}^{m} \alpha_i y_i \Phi(x_i) \Phi(x) \right) \]
\[ = \text{sgn} \left( v_0 + \sum_{i=1}^{m} \alpha_i y_i K(x_i, x) \right) \]
where \( K(x_i, x) = \Phi^T(x_i) \Phi(x) \) is called the kernel.
The maximum margin classifier

Remember where this process started:

The kernel is computing

$$K(x, x') = \Phi^T(x)\Phi(x') = \sum_{i=1}^k \phi_i(x)\phi_i(x')$$

This is generally called an inner product.

If it’s a hard problem then you’ll probably want lots of basis functions so $k$ is BIG:

$$h_w, w_0(x) = \text{sgn}(w_0 + w^T\Phi(x)) = \text{sgn}(w_0 + \sum_{i=1}^k w_i\phi_i(x))$$

$$= \text{sgn}(w_0 + \sum_{i=1}^m \alpha_i y_i \Phi^T(x_i)\Phi(x))$$

$$= \text{sgn}(w_0 + \sum_{i=1}^m \alpha_i y_i K(x_i, x))$$

What if $K(x, x')$ is easy to compute even if $k$ is HUGE? (In particular $k >> m$.)

1. We get a definite computational advantage by using the dual version with weights $\alpha$.
2. Mercer’s theorem tells us exactly when a function $K$ has a corresponding set of basis functions $\{\phi_i\}$.

The last is particularly prominent. Interestingly, the corresponding set of basis functions is infinite. (So we get an improvement in computational complexity from infinite to linear in the number of examples!)

Maximum margin classifier: the dual version

Collecting together some of the results up to now:

1. The Lagrangian is

$$L(w, w_0, \alpha) = \frac{1}{2}||w||^2 - \sum_i \alpha_i (y_i f_{w, w_0}(x_i) - 1).$$

2. The weight vector is

$$w = \sum_i \alpha_i y_i \Phi(x_i).$$

3. The KKT conditions require

$$\sum_i \alpha_i = 0.$$ 

It’s easy to show (this is an exercise) that the dual optimization problem is to maximize

$$L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

such that $\alpha \geq 0$. 

Designing good kernels $K$ is a subject in itself.

Luckily for the majority of the time you will tend to see one of the following:

1. Polynomial:

$$K_{c,d}(x, x') = (c + x^T x')^d$$

where $c$ and $d$ are parameters.

2. Radial basis function (RBF):

$$K_{\sigma^2}(x, x') = \exp\left(\frac{1}{2\sigma^2}||x - x'||^2\right)$$

where $\sigma^2$ is a parameter.

The last is particularly prominent. Interestingly, the corresponding set of basis functions is infinite. (So we get an improvement in computational complexity from infinite to linear in the number of examples!)
Support Vector Machines

There is one thing still missing:

Problem: so far we’ve only covered the linearly separable case.
Even though that means linearly separable in the extended space it’s still not enough.
By dealing with this we get the Support Vector Machine (SVM).

Fortunately a small modification allows us to let some examples be misclassified.
We introduce the slack variables $\xi_i$, one for each example.
Although $y' f_{w, w_0}(x') < 0$ we have $y' f_{w, w_0}(x') \geq 1 - \xi_i$ and we try to force $\xi_i$ to be small.

Once again, the theory of constrained optimization can be employed:

1. We get the same insights into the solution of the problem, and the same conclusions.
2. The development is exactly analogous to what we’ve just seen.
However as is often the case it is not straightforward to move all the way to having a functioning training algorithm.
For this some attention to good numerical computing is required. See:

Supervised learning in practice

We now look at several issues that need to be considered when applying machine learning algorithms in practice:

- We often have more examples from some classes than from others.
- The obvious measure of performance is not always the best.
- Much as we’d love to have an optimal method for finding hyperparameters, we don’t have one, and it’s unlikely that we ever will.
- We need to exercise care if we want to claim that one approach is superior to another.

This part of the course has an unusually large number of Commandments. That’s because so many people get so much of it wrong!

Supervised learning

As usual, we want to design a classifier.

\[ h_\theta(x) \]

It should take an attribute vector

\[ x^T = [x_1, x_2, \ldots, x_n] \]

and label it.

We now denote a classifier by \( h_\theta(x) \) where \( \theta^T = (w, p) \) denotes any weights \( w \) and (hyper)parameters \( p \).

To keep the discussion and notation simple we assume a classification problem with two classes labelled +1 (positive examples) and −1 (negative examples).

Supervised learning

Previously, the learning algorithm was a box labelled \( L \).
Machine Learning Commandments

We’ve already come across the Commandment:

**Thou shalt try a simple method.** Preferably many simple methods.

Now we will add:

**Thou shalt use an appropriate measure of performance.**

---

Measuring performance

How do you assess the performance of your classifier?

1. That is, after training, how do you know how well you’ve done?

2. In general, the only way to do this is to divide your examples into a smaller training set $s$ of $m$ examples and a test set $s'$ of $m'$ examples.

The **GOLDEN RULE:** data used to assess performance must NEVER have been seen during training.

This might seem obvious, but it was a major flaw in a lot of early work.

---

Measuring performance

How do we choose $m$ and $m'$? Trial and error!

Assume the training is complete, and we have a classifier $h_\theta$ obtained using only $s$. How do we use $s'$ to assess our method’s performance?

The obvious way is to see how many examples in $s'$ the classifier classifies correctly:

$$\hat{er}_{s'}(h_\theta) = \frac{1}{m'} \sum_{i=1}^{m'} I[h_\theta(x'_i) \neq y'_i]$$

where

$$s' = [(x'_1, y'_1) \ (x'_2, y'_2) \ \cdots \ (x'_m, y'_m)]^T$$

and

$$I[z] = \begin{cases} 1 & \text{if } z = \text{true} \\ 0 & \text{if } z = \text{false} \end{cases}$$

This is just an estimate of the *probability of error* and is often called the *accuracy*.

---

Unbalanced data

Unfortunately it is often the case that we have unbalanced data and this can make such a measure misleading. For example:

If the data is naturally such that almost all examples are negative (medical diagnosis for instance) then simply classifying everything as negative gives a high performance using this measure.

We need more subtle measures.

For a classifier $h$ and any set $s$ of size $m$ containing $m^+$ positive examples and $m^-$ negative examples...
Unbalanced data

Define

1. The true positives
   \[ P^+ = \{(x, +1) \in s | h(x) = +1\} \], and \( p^+ = |P^+| \)

2. The false positives
   \[ P^- = \{(x, -1) \in s | h(x) = +1\} \], and \( p^- = |P^-| \)

3. The true negatives
   \[ N^+ = \{(x, -1) \in s | h(x) = -1\} \], and \( n^+ = |N^+| \)

4. The false negatives
   \[ N^- = \{(x, +1) \in s | h(x) = -1\} \], and \( n^- = |N^-| \)

Thus \( \hat{\epsilon}_R(h) = (p^+ + n^+)/m \).

This allows us to define more discriminating measures of performance.

Performance measures

Some standard performance measures:

1. Precision/Positive predictive value (PPV) \( \frac{p^+}{p^+ + p^-} \).
2. Negative predictive value (NPR) \( \frac{n^+}{n^+ + n^-} \).
3. Recall/Sensitivity/True positive rate (TPR) \( \frac{p^+}{p^+ + n^-} \).
4. Specificity/True negative rate (TNR) \( \frac{n^+}{n^+ + p^-} \).
5. False positive rate (FPR) \( \frac{p^-}{p^- + n^+} \).
6. False negative rate (FNR) \( \frac{n^-}{n^- + p^+} \).
7. False discovery rate \( \frac{p^-}{p^- + p^+} \).

In addition, plotting sensitivity (true positive rate) against the false positive rate while a parameter is varied gives the receiver operating characteristic (ROC) curve.

Performance measures

The following specifically take account of unbalanced data:

1. Matthews Correlation Coefficient (MCC)
   \[ \text{MCC} = \frac{p^+ n^+ - p^- n^-}{\sqrt{(p^+ + p^-)(n^+ + n^-)(p^+ + n^-)(n^+ + p^-)}} \]

2. F1 score
   \[ F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]

When data is unbalanced these are preferred over the accuracy.

Machine Learning Commandments

Thou shalt not use default parameters.
Thou shalt not use parameters chosen by an unprincipled formula.
Thou shalt not avoid this issue by clicking on ‘Learn’ and hoping it works.
Thou shalt either choose them carefully or integrate them out.
Bad hyperparameters give bad performance

Validation and crossvalidation

The next question: how do we choose hyperparameters?

Answer: try different values and see which values give the best (estimated) performance.

There is however a problem:

If I use my test set $s'$ to find good hyperparameters, then I can’t use it to get a final measure of performance. (See the Golden Rule above.)

Solution 1: make a further division of the complete set of examples to obtain a third, validation set:

Now, to choose the value of a hyperparameter $p$:

For some range of values $p_1, p_2, \ldots, p_n$

1. Run the training algorithm using training data $s$ and with the hyperparameter set to $p$.
2. Assess the resulting $h_\theta$ by computing a suitable measure (for example accuracy, MCC or F1) using $v$.

Finally, select the $h_\theta$ with maximum estimated performance and assess its actual performance using $s'$. 
Validation and crossvalidation

This was originally used in a similar way when deciding the best point at which to stop training a neural network.

The figure shows the typical scenario.

Crossvalidation

The method of crossvalidation takes this a step further.

We our complete set into training set $s$ and testing set $s'$ as before.

But now instead of further subdividing $s$ just once we divide it into $n$ folds $s^{(i)}$ each having $m/n$ examples.

Typically $n = 10$ although other values are also used, for example if $n = m$ we have leave-one-out cross-validation.

Crossvalidation

Let $s_{-i}$ denote the set obtained from $s$ by removing $s^{(i)}$.

Let $e_{r}^{d}(h)$ denote any suitable error measure, such as accuracy, MCC or F1, computed for $h$ using fold $i$.

Let $g_{k-1}^{d}$ be the classifier obtained by running learning algorithm $L$ on examples $s_{-i}$ using hyperparameters $p$.

Then,

$$\frac{1}{n} \sum_{i=1}^{n} e_{r}^{d}(g_{k-1}^{d}(L_{k-1}^{d}, p))$$

is the $n$-fold crossvalidation error estimate.

For example, let $s^{(i)}_{j}$ denote the $j$th example in the $i$th fold. Then using accuracy as the error estimate we have

$$\frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{m/n} \left[ L_{s_{-i}}^{d}(x_{j}^{(i)}) \neq y_{j}^{(i)} \right]$$

Two further points:

1. What if the data are unbalanced? Stratified crossvalidation chooses folds such that the proportion of positive examples in each fold matches that in $s$.
2. Hyperparameter choice can be done just as above, using a basic search.

What happens however if we have multiple hyperparameters?

1. We can search over all combinations of values for specified ranges of each parameter.
2. This is the standard method in choosing parameters for support vector machines (SVMs).
3. With SVMs it is generally limited to the case of only two hyperparameters.
4. Larger numbers quickly become infeasible.
Crossvalidation

This is what we get for an SVM applied to the two spirals:

Using crossvalidation to optimize the hyperparameters $C$ and $\sigma^2$.

Machine Learning Commandments

Thou shalt *provide evidence* before claiming that *thy method is the best*.

Thou shalt take extra notice of this Commandment if *thou considers thyself a True And Pure Bayesian*.

Thou shalt even take notice of this Commandment if *thou considers thyself to be DEEP*.

Comparing classifiers

Imagine I have compared the AIMLBlockChain Classificator and the DeepHype Discriminotron and found that:

1. The Classificator has estimated accuracy 0.981 on the test set.
2. The Discriminotron has estimated accuracy 0.982 on the test set.

Can I claim that the Discriminotron is the better classifier?

Answer:

NO! NO! NO! NO! NO! NO! NO! NO! NO!!!!!!!!!!!!!!

Note for next year: include photo of grumpy-looking cat.
Assessing a single classifier

From *Mathematical Methods for Computer Science*:

The Central Limit Theorem: If we have independent identically distributed (iid) random variables $X_1, X_2, \ldots, X_n$ with mean $E[X] = \mu$ and variance $E[(X - \mu)^2] = \sigma^2$, then as $n \to \infty$ $\frac{\hat{X}_n - \mu}{\sigma/\sqrt{n}} \to N(0,1)$ where $\hat{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. 

Rearranging this using the equation from the previous slide we have that with probability $p$ $\mu \in \left[ \hat{X}_n \pm z_p \frac{\sigma}{\sqrt{n}} \right]$. 

We don’t know $\sigma^2$ but it can be estimated using $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{X}_n)^2$. Alternatively, when $X$ takes only values 0 or 1 $\sigma^2 = E[(X - \mu)^2] = E[X^2] - \mu^2 = \mu(1 - \mu) \approx \hat{X}_n(1 - \hat{X}_n)$.

Typically we are interested in a 95% confidence interval, for which $z_p = 1.96$. Thus, when $m > 30$ (so that the central limit theorem applies) we know that, with probability 0.95 $\hat{\text{er}}(h) = \frac{1}{m} \sum_{i=1}^{m} I[h(x_i) \neq y_i]$ for a test set $m$.

We can find a confidence interval for this estimate using precisely the derivation above, simply by noting that the $X_i$ are the random variables $X_i = I[h(x_i) \neq y_i]$.

Example: I have 100 test examples and my classifier makes 18 errors. With probability 0.95 I know that $\hat{\text{er}}(h) = \frac{1}{m} \sum_{i=1}^{m} I[h(x_i) \neq y_i] 
\text{er}(h) = \hat{\text{er}}(h) \pm 1.96 \sqrt{\frac{\hat{\text{er}}(h)(1 - \hat{\text{er}}(h))}{m}}$.

Thus, $\text{er}(h) = 0.18 \pm 1.96 \sqrt{\frac{0.18(1 - 0.18)}{100}} = 0.18 \pm 0.075$.

This should perhaps *raise an alarm* regarding our suggested comparison of classifiers above.
Assessing a single classifier

There is an important distinction to be made here:
1. The mean of $X$ is $\mu$ and the variance of $X$ is $\sigma^2$.
2. We can also ask about the mean and variance of $X_n$.
3. The mean of $X_n$ is
   \[
   \mathbb{E}[\hat{X}_n] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \mu.
   \]
4. It is left as an exercise to show that the variance of $\hat{X}_n$ is
   \[
   \sigma^2_{\hat{X}_n} = \frac{\sigma^2}{n}.
   \]

Comparing classifiers

Now say I have classifiers $h_1$ (Bloggs Classificator 2000) and $h_2$ (CleverCorp Discriminotron) and I want to know something about the quantity $d = \hat{e}_s(h_1) - \hat{e}_s(h_2)$.

I estimate $d$ using
   \[
   \hat{d} = \hat{e}_s(h_1) - \hat{e}_s(h_2)
   \]
where $s_1$ and $s_2$ are two independent test sets.

Notice:
1. The estimate of $d$ is a sum of random variables, and we can apply the central limit theorem.
2. The estimate is unbiased
   \[
   \mathbb{E}[\hat{e}_s(h_1) - \hat{e}_s(h_2)] = d.
   \]

Comparing classifiers

We are using the values $z_p$ such that if $x \sim N(0, 1)$ then
   \[
   \Pr(-z_p \leq x \leq z_p) > p.
   \]

There is an alternative way to think about this.
1. Say we have a random variable $Y$ with variance $\sigma^2_Y$ and mean $\mu_Y$.
2. The random variable $Y = \mu_Y$ has variance $\sigma^2_Y$ and mean 0.
3. It is a straightforward exercise to show that dividing a random variable having variance $\sigma^2$ by $\sigma$ gives us a new random variable with variance 1.
4. Thus the random variable $\frac{Y - \mu_Y}{\sigma_Y}$ has mean 0 and variance 1.

So: with probability $p$
   \[
   Y = \mu_Y \pm z_p \sigma_Y
   \]
   \[
   \mu_Y = Y \pm z_p \sigma_Y.
   \]

Compare this with what we saw earlier. You need to be careful to keep track of whether you are considering the mean and variance of a single RV or a sum of RVs.
Comparing classifiers—hypothesis testing

This still doesn’t tell us directly about whether one classifier is better than another—whether \( h_1 \) is better than \( h_2 \).

What we actually want to know is whether
\[
d = \text{er}(h_1) - \text{er}(h_2) > 0.
\]

Say we’ve measured \( \hat{D} = \hat{d} \). Then:

- Imagine the actual value of \( d \) is \( 0 \).
- Recall that the mean of \( \hat{D} \) is \( \hat{d} \).
- So larger measured values \( d \) are less likely, even though some random variation is inevitable.
- If it is highly unlikely that when \( d = 0 \) a measured value of \( \hat{d} \) would be observed, then we can be confident that \( d > 0 \).
- Thus we are interested in
\[
\Pr(\hat{D} > \hat{d} + \hat{d}).
\]

This is known as a one-sided bound.

One-sided bounds

Given the two-sided bound
\[
\Pr(-z \leq x \leq z) = 1 - \epsilon
\]
we actually need to know the one-sided bound
\[
\Pr(x \leq z) = 1 - \epsilon / 2.
\]

Clearly, if our random variable is Gaussian then \( \Pr(x \leq z) = 1 - \epsilon / 2 \).

Comparing algorithms: paired t-tests

We now know how to compare hypotheses \( h_1 \) and \( h_2 \).

But we still haven’t properly addressed the comparison of algorithms.

- Remember, a learning algorithm \( L \) maps training data \( s \) to hypothesis \( h \).
- So we really want to know about the quantity
\[
d = \mathbb{E}_{s \in S} [\text{er}(L_1(s)) - \text{er}(L_2(s))].
\]
- This is the expected difference between the actual errors of the two different algorithms \( L_1 \) and \( L_2 \).

Unfortunately, we have only one set of data \( s \) available and we can only estimate errors \( \text{er}(h) \)—we don’t have access to the actual quantities.

We can however use the idea of crossvalidation.

Comparing algorithms: paired t-tests

Recall, we subdivide \( s \) into \( n \) folds \( s^{(i)} \) each having \( m/n \) examples

and denote by \( s_{-i} \) the set obtained from \( s \) by removing \( s^{(i)} \). Then
\[
\frac{1}{n} \sum_{i=1}^{n} \text{er}_{s^{(i)}}(L(s_{-i}))
\]
is the \( n \)-fold crossvalidation error estimate. Now we estimate \( d \) using
\[
\hat{d} = \frac{1}{n} \sum_{i=1}^{n} [\text{er}_{s^{(i)}}(L_1(s_{-i})) - \text{er}_{s^{(i)}}(L_2(s_{-i}))].
\]
Comparing algorithms: paired t-tests

As usual, there is a statistical test allowing us to assess how likely this estimate is to mislead us.

We will not consider the derivation in detail. With probability $p$

$$d \in \left[ \hat{d} \pm t_{p,n-1} \sigma_{\hat{d}} \right].$$

This is analogous to the equations seen above, however:

- The parameter $t_{p,n-1}$ is analogous to $z_p$.
- The parameter $t_{p,n-1}$ is related to the area under the Student’s $t$-distribution whereas $z_p$ is related to the area under the normal distribution.
- The relevant estimate of standard deviation is

$$\sigma_{\hat{d}} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (d_i - \hat{d})^2}$$

where

$$d_i = \text{er}_{\theta(i)}(L_1(s-i)) - \text{er}_{\theta(i)}(L_2(s-i)).$$

Where now?

There are some simple take-home messages from the study of SVMs:

You can get state-of-the-art performance.
You can do this using the kernel trick to obtain a non-linear model.
You can do this without invoking the full machinery of the Bayes-optimal classifier.

BUT:

You don’t have anything keeping you honest regarding which assumptions you’re making.
As we shall see, by using the full-strength probabilistic framework we gain some useful extras.
In particular, the ability to assign confidences to our predictions.

The Bayesian approach to neural networks

We’re now going to see how the idea of the Bayes-optimal classifier can be applied to neural networks.

We have:

- A neural network computing a function $h_{\theta}(x)$. (In fact this can be pretty much any parameterized function we like.)
- A training sequence $s^T = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, split into

$$y = (y_1 y_2 \cdots y_m)$$

and

$$X = (x_1 x_2 \cdots x_m).$$

Machine Learning and Bayesian Inference

Dr Sean Holden
Computer Laboratory, Room FC06
Telephone extension 63725
Email: sbh11@cl.cam.ac.uk
www.cl.cam.ac.uk/~sbh11/

Part III: back to Bayes
Bayesian neural networks
Gaussian processes

Copyright © Sean Holden 2002-18.
The Bayesian approach to neural networks

We’re only going to consider regression. Classification can also be done this way, but it’s a bit more complicated.

For classification we derived the Bayes-optimal classifier as the maximizer of:

$$\Pr(C|x, s) = \int \Pr(C|w, x) p(w|s) \, dw$$

For regression the Bayes-optimal classifier ends up having the same expression as we’ve already seen. We want to compute:

$$p(Y|x, s) = \int p(Y|w, x) p(w|s) \, dw$$

$s$ is the training set.

$x$ is a new example to be classified.

$Y$ is the RV representing the prediction for $x$.

What’s going on? Turning prior into posterior

Let’s make a brief sidetrack into what’s going on with the posterior density

$$p(w|y, X) \propto p(y|w, X)p(w).$$

Typically, the prior starts wide and as we see more data the posterior narrows.

It turns out that if you try to incorporate the density $p(x)$ modelling how feature vectors are generated, things can get complicated. So:

1. We regard all input vectors as fixed; they are not treated as random variables.
2. This means that, strictly speaking, they should no longer appear in expressions like $p(Y|w, x)$.
3. However, this seems to be uniformly disliked—writing $p(Y|w)$ for an expression that still depends on $x$ seems confusing.
4. Solution: write $p(Y|w, x)$ instead. Note the semi-colon!

So we’re actually going to look at

$$p(Y|y, x, X) = \int p(Y|w, x) p(w|y, X) \, dw$$

NOTE: this is a notational hack. There’s nothing new, just an attempt at clarity.

What’s going on? Turning prior into posterior

This can be seen very clearly if we use real numbers:
So now we have three things to do:

1. **STEP 1**: remind ourselves what $p(Y|w; x)$ is.
2. **STEP 2**: remind ourselves what $p(w|y; X)$ is.
3. **STEP 3**: do the integral. *(This is the fun bit...)*

The first two steps are straightforward as we’ve already derived them when looking at maximum-likelihood and MAP learning.

**STEP 1**: assuming Gaussian noise is added to the labels so $y = h_w(x) + \epsilon$

where $\epsilon \sim \mathcal{N}(0, \sigma^2_n)$ we have the usual likelihood

$$p(Y|w; x) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left( -\frac{1}{2\sigma_n^2} (Y - h_w(x))^2 \right).$$

Here, the subscript in $\sigma^2_n$ reminds us that it’s the variance of the noise. Traditionally this is re-written using the hyperparameter

$$\beta = \frac{1}{\sigma^2_n}$$

so the likelihood is

$$p(Y|w; x) \propto \exp \left( -\frac{\beta}{2} (Y - h_w(x))^2 \right).$$

**STEP 2**: the posterior is also exactly as it was when we derived the MAP learning algorithms.

$$p(w|y; X) \propto p(y|w; X)p(w)$$

and as before, the likelihood is

$$p(y|w; X) \propto \exp \left( -\frac{\beta}{2} \sum_{i=1}^{m} (y_i - h_w(x_i))^2 \right)$$

$$= \exp \left( -\beta E(w) \right)$$

and using a Gaussian prior with mean 0 and covariance $\Sigma = \sigma^2 I$ gives

$$p(w) \propto \exp \left( -\frac{\alpha}{2}||w||^2 \right)$$

where traditionally the second hyperparameter is $\alpha = 1/\sigma^2$. Combining these

$$p(w|y; X) = \frac{1}{Z(\alpha, \beta)} \exp \left( -\left( \frac{\alpha}{2}||w||^2 + \beta E(w) \right) \right).$$

What’s going on? Turning prior into posterior

Considering the central part of $p(w|y; X)$:

$$\frac{\alpha}{2}||w||^2 + \beta E(w).$$

What happens as the number $m$ of examples increases?

- The first term *corresponding to the prior* remains fixed.
- The second term *corresponding to the likelihood* increases.

So for small training sequences the prior dominates, but for large ones $w_{ML}$ is a good approximation to $w_{MAP}$. 

The Bayesian approach to neural networks

**Step 3:** putting together steps 1 and 2, the integral we need to evaluate is:

\[
I \propto \int \exp \left( -\frac{\beta}{2} (Y - h_w(x))^2 \right) \exp \left( -\frac{\alpha ||w||^2}{2} + \beta E(w) \right) \, dw.
\]

Obviously this gives us all a sad face because there is no solution.

So what can we do now…?

---

Method 1: approximation to \( p(w|y; X) \)

\[
I \propto \int \exp \left( -\frac{\beta}{2} (Y - h_w(x))^2 \right) \exp \left( -\frac{\alpha ||w||^2}{2} + \beta E(w) \right) \, dw.
\]

The first approach introduces a Gaussian approximation to \( p(w|y; X) \) by using a Taylor expansion of

\[
S(w) = \frac{\alpha ||w||^2}{2} + \beta E(w)
\]

at the maximum a posteriori weights \( w_{MAP} \).

This allows us to use a standard integral.

The result will be approximate but we hope it’s good!

Let’s recall how Taylor series work...

---

Reminder: Taylor expansion

In one dimension the Taylor expansion about a point \( x_0 \in \mathbb{R} \) for a function \( f: \mathbb{R} \to \mathbb{R} \) is

\[
f(x) \approx f(x_0) + \frac{1}{1!}(x - x_0)f'(x_0) + \frac{1}{2!}(x - x_0)^2f''(x_0) + \cdots + \frac{1}{k!}(x - x_0)^k f^k(x_0).
\]

What does this look like for the kinds of function we’re interested in? As an example we can try to approximate

\[
\exp(-f(x))
\]

where

\[
f(x) = x^4 - \frac{1}{3}x^3 - 7x^2 + \frac{5}{3}x + 22.
\]

This has a form similar to \( S(w) \), but in one dimension.

---

The Bayesian approach to neural networks

In order to make further progress it’s necessary to perform integrals of the general form

\[
\int F(w)p(w|y; X) \, dw
\]

for various functions \( F \) and this is generally not possible.

There are two ways to get around this:

1. We can use an approximate form for \( p(w|y; X) \).
2. We can use Monte Carlo methods.

We’ll be taking a look at both possibilities.
Reminder: Taylor expansion

The functions of interest look like this:

By replacing \(-f(x)\) with its Taylor expansion about its maximum, which is at \(x_{\text{max}} \approx 2.1437\), we can see what the approximation to \(\exp(−f(x))\) looks like. Note that the \(\exp\) hugely emphasises peaks.

Reminder: Taylor expansion

In multiple dimensions the Taylor expansion for \(k=2\) is

\[
f(x) \approx f(x_0) + \frac{1}{1!} (x - x_0)^T \nabla f(x) \big|_{x_0} (x - x_0) + \frac{1}{2!} (x - x_0)^T \nabla^2 f(x) \big|_{x_0} (x - x_0)
\]

where \(\nabla\) denotes gradient

\[
\nabla f(x) = \left( \frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)
\]

and \(\nabla^2 f(x)\) is the matrix with elements

\[
M_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}
\]

(Looks complicated, but it’s just the obvious extension of the 1-dimensional case.)

Method 1: approximation to \(p(w|y; X)\)

Applying this to \(S(w)\) and expanding around \(w_{\text{MAP}}\)

\[
S(w) = \frac{\alpha |w|^2}{2} + \beta E(w) \approx S(w_{\text{MAP}}) + \frac{1}{2} (w - w_{\text{MAP}})^T A (w - w_{\text{MAP}}).
\]

- As \(w_{\text{MAP}}\) minimises the function the first derivatives are zero and the corresponding term in the Taylor expansion disappears.
- The quantity \(A = \nabla \nabla S(w)|_{w_{\text{MAP}}}\) can be simplified.

This is because

\[
A = \nabla \nabla \left( \frac{\alpha |w|^2}{2} + \beta E(w) \right) \big|_{w_{\text{MAP}}} = \alpha I + \beta \nabla \nabla E(w_{\text{MAP}}).
\]
Method 1: approximation to $p(w|y; X)$

We actually already know something about how to get $w_{\text{MAP}}$:

1. A method such as backpropagation can be used to compute $\nabla S(w)$.
2. The vector $w_{\text{MAP}}$ can then be obtained using any standard optimisation method (such as gradient descent).

It’s also likely to be straightforward to compute $\nabla \nabla E(w)$:

The quantity $\nabla \nabla E(w)$ can be evaluated using an extended form of backpropagation.

A useful integral

Dropping for this slide only the special meaning usually given to the vector $x$,

here is a useful standard integral:

If $A \in \mathbb{R}^{n \times n}$ is symmetric then for $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$

$$
\int_{\mathbb{R}^n} \exp \left( -\frac{1}{2} (x^T A x + x^T b + c) \right) \, dx = (2\pi)^{n/2} |A|^{-1/2} \exp \left( -\frac{1}{2} \left( c - b^T A^{-1} b \right) \right)
$$

You’re not expected to know how to evaluate this, but see the handout on the course web page if you’re curious1.

To make this easy to refer to, let’s call it the BIG INTEGRAL.

Method 1: approximation to $p(w|y; X)$

Defining

$$
\Delta w = w - w_{\text{MAP}},
$$

we now have an approximation

$$
p(w|y; X) \approx \frac{1}{Z} \exp \left( -S(w_{\text{MAP}}) - \frac{1}{2} \Delta w^T A \Delta w \right).
$$

Using the BIG INTEGRAL

$$
Z = (2\pi)^{W/2} |A|^{-1/2} \exp(-S(w_{\text{MAP}}))
$$

where $W$ is the number of weights.

Let’s plug this approximation back into the expression for the Bayes-optimum and see what we get...

Method 1: approximation to $p(w|y; X)$

$I \propto \int \exp \left( -\frac{\beta}{2} (Y - h_w(x))^2 \right) \exp \left( -\frac{1}{2} \Delta w^T A \Delta w \right) \, dw.$

Likelihood $p(Y|x;w)$ Approximation to $p(w|y; X)$

There is still no solution! We need another approximation...

We can introduce a linear approximation2 of $h_w(x)$ at $w_{\text{MAP}}$:

$$
h_w(x) \approx h_{w_{\text{MAP}}}(x) + g^T \Delta w
$$

where $g = \nabla h_w(x)_{|w_{\text{MAP}}}.$

(By linear approximation we just mean the Taylor expansion for $k = 1$.)

---

1 No, I won’t ask you to evaluate it in the exam.
2 We really are making assumptions here—this is OK if we assume that $p(w|y; X)$ is narrow, which depends on $A$. 
**Method 1: second approximation**

This leads to

\[
p(Y|\mathbf{x}, \mathbf{X}) \propto \int \exp \left( -\frac{\beta}{2} (Y - h_{\text{MAP}}(\mathbf{x}) - \mathbf{g}^T \Delta \mathbf{w})^2 - \frac{1}{2} \Delta \mathbf{w}^T \mathbf{A} \Delta \mathbf{w} \right) d\mathbf{w}.
\]

**SUCCESS!!!**

This integral can be evaluated (this is an exercise) using the BIG INTEGRAL to give THE ANSWER...

\[
p(Y|\mathbf{x}, \mathbf{X}) \simeq \frac{1}{\sqrt{2\pi \sigma^2_Y}} \exp \left( -\frac{(Y - h_{\text{MAP}}(\mathbf{x}))^2}{2\sigma^2_Y} \right)
\]

where

\[
\sigma^2_Y = \frac{1}{\beta} + \mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}.
\]

**Method 1: final expression**

**Hooray! But what does it mean?**

This is a Gaussian density, so we can now see that:

\[
p(Y|\mathbf{y}; \mathbf{x}, \mathbf{X}) \text{ peaks at } h_{\text{MAP}}(\mathbf{x}).
\]

That is, the MAP solution.

The variance \(\sigma^2_Y\) can be interpreted as a measure of certainty:

The first term of \(\sigma^2_Y\) is \(1/\beta\) and corresponds to the noise.

The second term of \(\sigma^2_Y\) is \(\mathbf{g}^T \mathbf{A}^{-1} \mathbf{g}\) and corresponds to the width of \(p(\mathbf{w}|\mathbf{y}; \mathbf{X})\).

**Method II: Markov chain Monte Carlo (MCMC) methods**

The second solution to the problem of performing integrals

\[
I = \int F(\mathbf{w}) p(\mathbf{w}|\mathbf{y}; \mathbf{X}) d\mathbf{w}
\]

is to use Monte Carlo methods. The basic approach is to make the approximation

\[
I \approx \frac{1}{N} \sum_{i=1}^{N} F(\mathbf{w}_i)
\]

where the \(\mathbf{w}_i\) have distribution \(p(\mathbf{w}|\mathbf{y}; \mathbf{X})\). Unfortunately, generating \(\mathbf{w}_i\) with a given distribution can be non-trivial.
A simple technique is to introduce a random walk, so 
\[ w_{i+1} = w_i + \epsilon \]
where \( \epsilon \) is zero mean spherical Gaussian and has small variance. Obviously the sequence \( w_i \) does not have the required distribution. However, we can use the Metropolis algorithm, which does not accept all the steps in the random walk:

1. If \( p(w_{i+1}|y; X) > p(w_i|y; X) \) then accept the step.
2. Else accept the step with probability \( \frac{p(w_{i+1}|y; X)}{p(w_i|y; X)} \).

In practice, the Metropolis algorithm has several shortcomings, and a great deal of research exists on improved methods, see:


The prior and likelihood depend on \( \alpha \) and \( \beta \) respectively so we now make this clear and write

\[ p(w|y, \alpha, \beta) = \frac{p(y|w, \beta)p(w|\alpha)}{p(y|\alpha, \beta)}. \]

Don’t worry about recalling the actual expressions for the prior and likelihood—we’re not going to delve deep enough to need them.

Let’s write down directly something that might be useful to know:

\[ p(\alpha, \beta|y) = \frac{p(y|\alpha, \beta)p(\alpha, \beta)}{p(y)}. \]

Hierarchical Bayes and the evidence

If we know \( p(\alpha, \beta|y) \) then a straightforward approach is to use the values for \( \alpha \) and \( \beta \) that maximise it:

\[ \arg\max_{\alpha, \beta} p(\alpha, \beta|y). \]

Here is a standard trick: assume that the prior \( p(\alpha, \beta) \) is flat, so that we can just maximise

\[ p(y|\alpha, \beta). \]

This is called type II maximum likelihood and is one common way of doing the job.

A (very) brief introduction to how to learn hyperparameters

So far in our coverage of the Bayesian approach to neural networks, the hyperparameters \( \alpha \) and \( \beta \) were assumed to be known and fixed.

- But this is not a good assumption because...
- ...\( \alpha \) corresponds to the width of the prior and \( \beta \) to the noise variance.
- So we really want to learn these from the data as well.
- How can this be done?

We now take a look at one of several ways of addressing this problem.

Note: from now on I’m going to leave out the dependencies on \( x \) and \( X \) as leaving them in starts to make everything cluttered.
Hierarchical Bayes and the evidence

The quantity
\[ p(y|\alpha, \beta) \]

is called the **evidence** or **marginal likelihood**.

When we re-wrote our earlier equation for the posterior density of the weights, making \( \alpha \) and \( \beta \) explicit, we found

\[ p(w|y, \alpha, \beta) = \frac{p(y|w, \beta)p(w|\alpha)}{p(y|\alpha, \beta)}. \]

So the evidence is the denominator in this equation.

This is the **common pattern** and leads to the idea of **hierarchical Bayes**: the evidence for the hyperparameters at one level is the denominator in the relevant application of Bayes' theorem.

---

Unsupervised learning

Can we find **regularity in data** without the aid of **labels**?

- We can learn from **unlabelled** data. This kind of learning is often known as **clustering**.
- We can do this using a simple, obvious algorithm known as **K-means**.
- We can also approach it **probabilistically** using **maximum likelihood**.
- This is less straightforward, but there is a general algorithm called **Expectation Maximization (EM)** that can be applied.
The \textit{K}-means algorithm

The example on the last slide was obtained using the classical \textit{K}-means algorithm. Given a set \( \{x_i\} \) of \( m \) points, guess that there are \( K \) clusters. Here \( K = 3 \).

Chose at random \( K \) centre points \( c_j \) for the clusters. Then iterate as follows:

1. Divide \( \{x_i\} \) into \( K \) clusters, so each point is associated with the closest centre:
   \[ x_i \in C_j \iff \forall k \ ||x_i - c_j|| \leq ||x_i - c_k||. \]
   Call these clusters \( C_1, \ldots, C_K \).

2. Update the cluster centres to be the average of the associated points:
   \[ c_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i. \]

Clustering as maximum-likelihood

We saw in the introductory lectures that data from \( K \) clusters can be modelled probabilistically as
\[
p(x|\theta) = \sum_{k=1}^{K} \pi_k p(x|\mu_k, \Sigma_k)
\]
where \( \theta = \{\pi, \mu_1, \Sigma_1, \ldots, \mu_K, \Sigma_K\} \) and typically \( p(x|\mu, \Sigma) = N(\mu, \Sigma) \).

This leads to a log-likelihood for \( m \) points of
\[
\log p(X|\theta) = \log \prod_{i=1}^{m} p(x_i|\theta) \\
= \sum_{i=1}^{m} \log p(x_i|\theta) \\
= \sum_{i=1}^{m} \log \sum_{k=1}^{K} \pi_k p(x_i|\mu_k, \Sigma_k)
\]
which tends to be hard to maximise directly. (You can find stationary points but they depend on one-another.)
We can however introduce some latent variables.

For each \( x_i \) introduce the latent variable \( z_i \) where

\[
z_i^T = \left[ z_i^{(1)}, \ldots, z_i^{(K)} \right]
\]

and

\[
z_i^{(j)} = \begin{cases} 
1 & \text{if } x_i \text{ was generated by cluster } j \\
0 & \text{otherwise}
\end{cases}
\]

Having introduced the \( z_i \) we can use the marginalization trick and write

\[
\log p(X|\theta) = \log \sum_{Z} p(X, Z|\theta) = \log \sum_{Z} p(X|Z, \theta)p(Z|\theta)
\]

where the final step has given us probabilities that are reasonably tractable.

Why is this?

First, if I know which cluster generated \( x_i \) then its probability is just that for the corresponding Gaussian

\[
p(x_i|z_i, \theta) = \prod_{k=1}^{K} \left[ p(x_i|\mu_k, \Sigma_k) \right]^{z_i^{(k)}}
\]

and similarly

\[
p(z_i|\theta) = \prod_{k=1}^{K} \left[ \pi_k \right]^{z_i^{(k)}}
\]

Consequently

\[
\log p(X, Z|\theta) = \sum_{i=1}^{m} \sum_{k=1}^{K} z_i^{(k)} (\log p(x_i|\mu_k, \Sigma_k) + \log \pi_k)
\]

What have we achieved so far?

1. We want to maximize the log-likelihood \( \log p(X|\theta) \) but this is intractable.
2. We introduce some latent variables \( Z \).
3. That gives us a tractable log-likelihood \( \log p(X, Z|\theta) \).

But how do we link them together?
The EM algorithm

The Expectation Maximization (EM) algorithm provides a general way of maximizing likelihood for problems like this.

Let’s do something a little strange. Let \( q(Z) \) be any distribution on the latent variables. Write

\[
\sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} = \sum_Z q(Z) \log \frac{p(Z|X, \theta)}{q(Z)} p(X|\theta)
\]

\[
= \sum_Z q(Z) \left( \log \frac{p(Z|X, \theta)}{q(Z)} + \log p(X|\theta) \right)
\]

\[
= -D_{KL}[q(Z)||p(Z|X, \theta)] + \sum_Z q(Z) \log p(X|\theta)
\]

\[
= -D_{KL}[q(Z)||p(Z|X, \theta)] + \log p(X|\theta)
\]

\( D_{KL} \) is the Kullback-Leibler (KL) distance.

The Kullback-Leibler (KL) distance

The Kullback-Leibler (KL) distance measures the distance between two probability distributions. For discrete distributions \( p \) and \( q \) it is

\[
D_{KL}[p||q] = \sum_x p(x) \log \frac{p(x)}{q(x)}
\]

It has the important properties that:

1. It is non-negative \( D_{KL}(p||q) \geq 0 \).
2. It is 0 precisely when the distributions are equal \( D_{KL}[p||q] = 0 \) if and only if \( p = q \).

The EM algorithm

If we also define

\[
L[q, \theta] = \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)}
\]

then we can re-arrange the last expression to get

\[
\log p(X|\theta) = L[q, \theta] + D_{KL}[q||p]
\]

and we know that \( D_{KL}[q||p] \geq 0 \) so that gives us an upper bound

\[
L[q, \theta] \leq \log p(X|\theta).
\]

The EM algorithm works as follows:

- We iteratively maximize \( L[q, \theta] \).
- We do this by alternately maximizing with respect to \( q \) and \( \theta \) while keeping the other fixed.
- Maximizing with respect to \( q \) is the E step.
- Maximizing with respect to \( \theta \) is the M step.

Let’s look at the two steps separately.

Say we have \( \theta_t \) at time \( t \) in the iteration.

For the E step, we have \( \theta_t \) fixed and

\[
\log p(X|\theta_t) = L[q, \theta_t] + D_{KL}[q||p]
\]

so this is easy!

1. As \( \theta_t \) is fixed, so is \( \log p(X|\theta_t) \).
2. So to maximize \( L[q, \theta_t] \) we must minimize \( D_{KL}[q||p] \).
3. And we know that \( D_{KL}[q||p] \) is minimized and equal to 0 when \( q = p \).

So in the E step we just choose

\[
q_{t+1}(Z) = p(Z|X, \theta_t).
\]
The EM algorithm

For the \textit{M step} we have
\[
L[q, \theta] = \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)
\]
where the second term (the entropy of \(q(Z)\)) doesn’t depend on \(\theta\).

We fix \(q_{t+1}(Z) = p(Z|X, \theta_t)\). We now choose \(\theta_{t+1}\) as
\[
\theta_{t+1} = \arg\max_{\theta} \sum_Z p(Z|X, \theta_t) \log p(X, Z|\theta_t)
= \arg\max_{\theta} \mathbb{E}_Z [\log p(X, Z|\theta_t)]
\]

It’s not as bad as it looks:
- Take the expected value inside the sums.
- The only part of the expression that depends on \(Z\) is \(z_i^{(k)}\).
- So we only have to compute \(\mathbb{E}_Z [z_i^{(k)}]\).

Thus
\[
\mathbb{E}_Z [z_i^{(k)}] = \sum_Z z_i^{(k)} p(Z|X, \theta_t)
= \sum_{z_1} \cdots \sum_{z_m} z_i^{(k)} p(z_1, \ldots, z_m|X, \theta_t)
= \sum_{z_i} z_i^{(k)} p(z_i|X, \theta_t) \text{ (marginalizing)}
= \sum_{z_i \in \{0,1\}} z_i^{(k)} p(z_i|X, \theta_t) \text{ (marginalizing again)}
= p(z_i^{(k)} = 1|X, \theta_t)
\]

We saw earlier that
\[
\log p(X, Z|\theta_t) = \sum_{i=1}^m \sum_{k=1}^K z_i^{(k)} \left( \log p(x_i|\mu_k, \Sigma_k) + \log \pi_k \right)
\]
where \(\theta\) collects all the parameters
\[
\theta_t = \{ \pi, \mu_1, \ldots, \mu_K, \Sigma_K \}.
\]

Note that the parameters \(\pi, \mu_k\) and \(\Sigma_k\) all have an implicit time \(t\) attached, but we avoid writing it to keep the notation manageable.

So: this step looks a little trickier; we need to maximize the expected value of this expression for the distribution \(p(Z|X, \theta_t)\).
The EM algorithm for a mixture model summarized:

- We want to find $\theta$ to maximize $\log p(X|\theta)$.
- But that’s not tractable.
- So we introduce an arbitrary distribution $q$ and obtain a lower bound $L(q, \theta) \leq \log p(X|\theta)$.
- We maximize the lower bound iteratively in two steps:
  1. **E step**: keep $\theta$ fixed and maximize with respect to $q$. This always results in $q(Z) = p(Z|X, \theta)$.
  2. **M step**: keep $q$ fixed and maximize with respect to $\theta$. For the mixture model this is

$$
\theta_{t+1} = \arg\max_{\theta} m \sum_{i=1}^{m} K \sum_{k=1}^{K} \gamma_k^{(i)} (\log p(x_i|\mu_k, \Sigma_k) + \log \pi_k)
$$
Uncertainty: Probability as Degree of Belief

At the start of the course, I presented a uniform approach to knowledge representation and reasoning using probability.

The world: $V = \{V_1, V_2, \ldots, V_n\}$. These are partitioned:

1. Query variables $Q = \{Q_1, Q_2, \ldots, Q_q\}$. We want to compute a distribution over these.
2. Observed variables $O = \{o_1, o_2, \ldots, o_m\}$. We know the values of these.
3. Latent variables $L = \{L_1, L_2, \ldots, L_l\}$. Everything else.

General knowledge representation and inference: the BIG PICTURE

Bayes’ theorem tells us how to update an inference when new information is available.

For example, if we now receive a new observation $O' = o'$ then

$$
\Pr(Q|o', o_1, o_2, \ldots, o_m) = \frac{1}{Z} \Pr(o'|Q, o_1, o_2, \ldots, o_m) \Pr(Q|o_1, o_2, \ldots, o_m)
$$

To compute a conditional distribution from a knowledge base $\Pr(V)$ we have to sum over the latent variables

$$
\Pr(Q|o_1, o_2, \ldots, o_m) = \sum_L \Pr(Q, L|o_1, o_2, \ldots, o_m)
$$

Simple eh?

HAH!!! No chance...

Even if all your RVs are just Boolean:

- For $n$ RVs knowing the knowledge base $\Pr(V)$ means storing $2^n$ numbers.
- So it looks as though storage is $O(2^n)$.
- You need to establish $2^n$ numbers to work with.
- Look at the summations. If there are $n$ latent variables then it appears that time complexity is also $O(2^n)$.
- In reality we might well have $n > 1000$, and of course it’s even worse if variables are non-Boolean.

And it really is this hard. The problem in general is #P-complete.
Even getting an approximate solution is provably intractable.
Bayesian Networks

Having seen that in principle, if not in practice, the full joint distribution alone can be used to perform any inference of interest, we now examine a practical technique.

- We introduce the Bayesian Network (BN) as a compact representation of the full joint distribution.
- We examine the way in which a BN can be constructed.
- We examine the semantics of BNs.
- We look briefly at how inference can be performed.
- We briefly introduce the Markov random field (MRF) as an alternative means of representing a distribution.

Conditional probability—a brief aside...

A brief aside on the dangers of interpreting implication versus conditional probability:

- \( \Pr(X = x | Y = y) = 0.1 \) does not mean that if \( Y = y \) then \( \Pr(X = x) = 0.1 \).
- \( \Pr(X) \) is a prior probability. It applies when you haven’t seen the value of \( Y \).
- The notation \( \Pr(X | Y = y) \) is for use when \( y \) is the entire evidence.
- \( \Pr(X | Y = y \land Z = z) \) might be very different.

Conditional probability is **not** analogous to logical implication.

Implication and conditional probability

In general, it is difficult to relate implication to conditional probability.

\[
\begin{align*}
\Pr(A \rightarrow B) &= \Pr(\neg A \lor B) \\
\Pr(A | B) &= \frac{\Pr(A \land B)}{\Pr(B)}
\end{align*}
\]

Imagine that fish are very rare, and most fish can swim.

With implication,
\[
\Pr(\text{fish} \rightarrow \neg \text{swim}) = \Pr(\neg \text{fish} \lor \neg \text{swim}) = \text{LARGE!}
\]

With conditional probability,
\[
\Pr(\neg \text{swim} | \text{fish}) = \frac{\Pr(\neg \text{swim} \land \text{fish})}{\Pr(\text{fish})} = \text{SMALL!}
\]

Bayesian networks: exploiting independence

One of the key reasons for the introduction of Bayesian networks is to let us exploit independence.

The initial pay-off is that this makes it easier to represent \( \Pr(V) \).

A further pay-off is that it introduces structure that can lead to more efficient inference.

Here is a very simple example.

If I toss a coin and roll a die, the full joint distribution of outcomes requires \( 2 \times 6 = 12 \) numbers to be specified.

<table>
<thead>
<tr>
<th></th>
<th>( \square )</th>
<th>( \oslash )</th>
<th>( \land )</th>
<th>( \neg )</th>
<th>( \ominus )</th>
<th>( \square )</th>
</tr>
</thead>
</table>
\( H \) | 0.014 | 0.028 | 0.042 | 0.057 | 0.071 | 0.086 |
\( T \) | 0.033 | 0.067 | 0.1 | 0.133 | 0.167 | 0.2 |

Here \( \Pr(\text{Coin} = H) = 0.3 \) and the die has probability \( i/21 \) for the \( i \)th outcome.
**Exploiting independence**

*BUT*: if we assume the outcomes are independent then

\[ \Pr(Coin, Dice) = \Pr(Coin) \Pr(Dice) \]

where \( \Pr(Coin) \) has two numbers and \( \Pr(Dice) \) has six.

So instead of 12 numbers we only need 8.

<table>
<thead>
<tr>
<th>CF</th>
<th>~CF</th>
</tr>
</thead>
<tbody>
<tr>
<td>SB</td>
<td>SB</td>
</tr>
<tr>
<td>HD</td>
<td>0.024</td>
</tr>
<tr>
<td>~HD</td>
<td>0.0019</td>
</tr>
</tbody>
</table>

- HD = Heart disease
- CP = Chest pain
- SB = Shortness of breath

Similarly, say instead of just considering HD, SB and CP we also consider the outcome of the *Oxford versus Cambridge tiddlywinks competition* TC:

\[ TC = \{\text{Oxford, Cambridge, Draw}\}. \]

**Conditional independence**

However although in this case we might not be able to exploit independence directly we *can* say that

\[ \Pr(CP, SB | HD) = \Pr(CP | HD) \Pr(SB | HD) \]

which simplifies matters.

**Conditional independence**: \( A \perp B | C \)

- \( A \) is conditionally independent of \( B \) given \( C \), written \( A \perp B | C \), if
  \[ \Pr(A, B | C) = \Pr(A | C) \Pr(B | C). \]
- If we know that \( C \) is the case then \( A \) and \( B \) are independent.
- Equivalently \( \Pr(A | B, C) = \Pr(A | C) \). (Prove this!)

Although CF and SB are *not* independent, they do not directly influence one another in a patient known to have heart disease.

This is much nicer!

\[ \Pr(HD | CP, SB) \propto \Pr(CP | HD) \Pr(SB | HD) \Pr(HD) \]
Bayesian networks

After a regrettable incident involving an inflatable gorilla, a famous College has decided to install an alarm for the detection of roof climbers.

- The alarm is very good at detecting climbers.
- Unfortunately, it is also sometimes triggered when one of the extremely fat geese that lives in the College lands on the roof.
- One porter’s lodge is near the alarm, and inhabited by a chap with excellent hearing and a pathological hatred of roof climbers: he always reports an alarm. His hearing is so good that he sometimes thinks he hears an alarm, even when there isn’t one.
- Another porter’s lodge is a good distance away and inhabited by an old chap with dodgy hearing who likes to listen to his collection of DEATH METAL with the sound turned up.

Bayesian networks

Also called probabilistic/belief/causal networks or knowledge maps.

- Each node is a random variable (RV).
- Each node $N_i$ has a distribution $\Pr(N_i|\text{parents}(N_i))$.
- A Bayesian network is a directed acyclic graph.
- Roughly speaking, an arrow from $N$ to $M$ means $N$ directly affects $M$.

Note that:
- In the present example all RVs are discrete (in fact Boolean) and so in all cases $\Pr(N_i|\text{parents}(N_i))$ can be represented as a table of numbers.
- Climber and Goose have only prior probabilities.
- All RVs here are Boolean, so a node with $p$ parents requires $2^p$ numbers.

A BN with $n$ nodes represents the full joint probability distribution for those nodes as

$$ \Pr(N_1 = n_1, N_2 = n_2, \ldots, N_n = n_n) = \prod_{i=1}^{n} \Pr(N_i = n_i|\text{parents}(N_i)). $$

For example

$$ \Pr(\neg C, \neg G, A, L_1, L_2) = \Pr(L_1|A) \Pr(L_2|A) \Pr(A|\neg C, \neg G) \Pr(\neg C) \Pr(\neg G) $$

$$ = 0.99 \times 0.6 \times 0.08 \times 0.95 \times 0.8.$$
In general \( \Pr(A, B) = \Pr(A|B) \Pr(B) \) so
\[
\Pr(N_1, \ldots, N_n) = \Pr(N_n|N_{n-1}, \ldots, N_1) \Pr(N_{n-1}, \ldots, N_1) = \prod_{i=1}^{n} \Pr(N_i|N_{i-1}, \ldots, N_1).
\]
Repeating this gives
\[
\Pr(N_1, \ldots, N_n) = \Pr(N_n|N_{n-1}, \ldots, N_1) \Pr(N_{n-1}, \ldots, N_1) \cdots \Pr(N_1)
\]
Now compare equations. We see that BNs make the assumption
\[
\Pr(N_i|N_{i-1}, \ldots, N_1) = \Pr(N_i|\text{parents}(N_i))
\]
for each node, assuming that \( \text{parents}(N_i) \subseteq \{N_{i-1}, \ldots, N_1\} \).

Each \( N_i \) is conditionally independent of its predecessors given its parents.

But it's not quite that straightforward: what if we want to talk about nodes other than predecessors and parents?

For example, it is possible to show:

Any node \( A \) is conditionally independent of the \( N_i \)—its non-descendants—given the \( P_i \)—its parents.

It is also possible to show:

Any node \( A \) is conditionally independent of all other nodes given the Markov blanket \( M_i \)—that is, its parents, its children and its children's parents.
There is a general method for inferring exactly what conditional independences are implied by a Bayesian network.

Let \( X, Y \) and \( Z \) be disjoint subsets of the RVs. Consider a path \( p \) consisting of directed (in any orientation) edges from some \( x \in X \) to some \( y \in Y \). For example:

The path \( p \) is said to be blocked by \( Z \) if one of three conditions holds.

Finally:

1. \( X \) and \( Y \) are d-separated by \( Z \) if all paths \( p \) from some \( x \in X \) to some \( y \in Y \) are blocked.
2. If \( X \) and \( Y \) are d-separated by \( Z \) then \( X \perp Y \mid Z \).

How do we represent \( \Pr(N_i \mid \text{parents}(N_i)) \) when nodes can denote general discrete and/or continuous RVs?

- BNs containing both kinds of RV are called hybrid BNs.
- Naive discretisation of continuous RVs tends to result in both a reduction in accuracy and large tables.
- \( O(2^p) \) might still be large enough to be unwieldy.
- We can instead attempt to use standard and well-understood distributions, such as the Gaussian.
- This will typically require only a small number of parameters to be specified.
More complex nodes

Example: a continuous RV with one continuous and one discrete parent.

Pr(Speed of car|Throttle position, Tuned engine)

where SC and TP are continuous and TE is Boolean.

- For a specific setting of ET = true it might be the case that SC increases with TP, but that some uncertainty is involved
  \[ \Pr(\text{SC}|\text{TP}, \text{et}) = N(g_\text{et}\text{TP} + c_\text{et}, \sigma_\text{et}^2). \]

- For an un-tuned engine we might have a similar relationship with a different behaviour
  \[ \Pr(\text{SC}|\text{TP}, \neg\text{et}) = N(g_{\neg\text{et}}\text{TP} + c_{\neg\text{et}}, \sigma_{\neg\text{et}}^2). \]

There is a set of parameters \( \{g, c, \sigma\} \) for each possible value of the discrete RV.

More complex nodes

Example: a discrete RV with a continuous parent

Pr(Go roofclimbing|Size of fine).

We could for example use the probit distribution

\[ \Pr(\text{Go roofclimbing} = \text{true}|\text{size}) = \Phi\left(\frac{t - \text{size}}{s}\right) \]

where

\[ \Phi(x) = \int_{-\infty}^{x} N(y)dy \]

and \( N \) is the Gaussian density with zero mean and variance 1.

Basic inference

We saw earlier that the full joint distribution can be used to perform all inference tasks:

\[ \Pr(Q|o_1, o_2, \ldots, o_m) = \frac{1}{Z} \sum_L \Pr(Q, L, o_1, o_2, \ldots, o_m) \]

where

- \( Q \) is the query.
- \( o_1, o_2, \ldots, o_m \) are the observations.
- \( L \) are the latent variables.
- \( 1/Z \) normalises the distribution.
- The query, observations and latent variables are a partition of the set \( V = \{V_1, V_2, \ldots, V_n\} \) of all variables.
Basic inference

As the BN fully describes the full joint distribution

\[ Pr(Q, L, o_1, o_2, \ldots, o_m) = \prod_{i=1}^{n} Pr(V_i|\text{parents}(V_i)) \]

it can be used to perform inference in the *obvious way*

\[ Pr(Q|o_1, o_2, \ldots, o_m) \propto \sum_{L} \prod_{i=1}^{n} Pr(V_i|\text{parents}(V_i)) \]

but this is *in practice problematic* for obvious reasons.

- More sophisticated algorithms aim to achieve this *more efficiently*.
- For complex BNs we resort to *approximation techniques*.

Performing exact inference

Consider the computation of the query \( Pr(C|l_1, l_2) \)

We have

\[ Pr(C|l_1, l_2) \propto \sum_{A} \sum_{G} Pr(C) Pr(G) Pr(A|C, G) Pr(l_1|A) Pr(l_2|A). \]

Here there are \( 5 \) multiplications for each set of values that appears for summation, and there are \( 4 \) such values.

In general this gives time complexity \( O(n2^n) \) for \( n \) Boolean RVs.

The naive implementation of this approach yields the *Enumerate-Joint-Ask* algorithm, which unfortunately requires \( O(2^n) \) time and space for \( n \) Boolean RVs.

The *enumeration-ask* algorithm improves matters to \( O(2^n) \) time and \( O(n) \) space by performing the computation *depth-first*.

However matters can be improved further by avoiding *duplication of computations*.

Performing exact inference

\( Pr(Q, L, o_1, \ldots, o_m) \) has a particular form expressing conditional independences:

\[
Pr(C, G, A, L_1, L_2) = Pr(C) \cdot Pr(G) \cdot Pr(A|C, G) \cdot Pr(L_1|A) \cdot Pr(L_2|A)
\]

Performing exact inference

Looking more closely we see that

\[
Pr(C|l_1, l_2) \propto \sum_{A} \sum_{G} Pr(C) \cdot Pr(G) \cdot Pr(A|C, G) \cdot Pr(l_1|A) \cdot Pr(l_2|A).
\]

\[
= \frac{1}{Z} Pr(C) \sum_{A} \sum_{G} Pr(l_1|A) \cdot Pr(l_2|A) \cdot Pr(G) \cdot Pr(A|C, G)
\]

\[
= \frac{1}{Z} Pr(C) \sum_{G} Pr(G) \sum_{A} Pr(A|C, G) \cdot Pr(l_1|A) \cdot Pr(l_2|A).
\]

There is some freedom in terms of how we *factorize* the expression.

This is a result of introducing *assumptions about conditional independence*.
Performing exact inference: variable elimination

Taking the second possibility:

$$
\text{Pr}(C) \sum_{G} \text{Pr}(G) \sum_{A} \text{Pr}(A|C,G) \text{Pr}(l1|A) \text{Pr}(l2|A)
$$

where $C, G, A, L1, L2$ denote the relevant factors.

The basic idea is to evaluate this from right to left (or in terms of the tree, bottom up) storing results as we progress and re-using them when necessary.

$\text{Pr}(l1|A)$ depends on the value of $A$. We store it as a table $F_{L1}(A)$. Similarly for $\text{Pr}(l2|A)$.

$$
F_{L1}(A) = \begin{pmatrix}
0.99 & 0.8 \\
0.08 & 0.001
\end{pmatrix}
$$

as $\text{Pr}(l1|a) = 0.99$, $\text{Pr}(l1|\neg a) = 0.08$ and so on.

Yes, provided multiplication of factors is defined correctly. Looking at

$$
\text{Pr}(C) \sum_{G} \text{Pr}(G) \sum_{A} \text{Pr}(A|C,G) \text{Pr}(l1|A) \text{Pr}(l2|A)
$$

note that:

1. The values of the product

$$
\text{Pr}(A|C,G) \text{Pr}(l1|A) \text{Pr}(l2|A)
$$

in the summation over $A$ depend on the values of $C$ and $G$ external to it, and the values of $A$.

2. So

$$
F_{A}(A,C,G)F_{L1}(A)F_{L2}(A)
$$

should be a table collecting values where correspondences between RVs are maintained.

This leads to a definition for multiplication of factors best given by example.
Performing exact inference: variable elimination

This process gives us

$$F_A(A, C, G)F_{L1}(A)F_{L2}(A) =
\begin{array}{c|cc|c}
A & C & G \\
\hline
\top & \top & \top & 0.98 \times 0.99 \times 0.6 \\
\top & \top & \bot & 0.96 \times 0.99 \times 0.6 \\
\top & \bot & \top & 0.2 \times 0.99 \times 0.6 \\
\top & \bot & \bot & 0.08 \times 0.99 \times 0.6 \\
\bot & \top & \top & 0.02 \times 0.98 \times 0.001 \\
\bot & \top & \bot & 0.04 \times 0.98 \times 0.001 \\
\bot & \bot & \top & 0.8 \times 0.98 \times 0.001 \\
\bot & \bot & \bot & 0.02 \times 0.98 \times 0.001 \\
\end{array}$$

Performing exact inference: variable elimination

Now, say for example we have \( \neg c, g \). Then doing the calculation explicitly would give

$$\sum_A \Pr(A|\neg c, g)\Pr(1|A)\Pr(2|A)$$

$$= \Pr(a|\neg c, g)\Pr(1|a)\Pr(2|a) + \Pr(\neg a|\neg c, g)\Pr(1|\neg a)\Pr(2|\neg a)$$

$$= (0.2 \times 0.99 \times 0.6) + (0.8 \times 0.08 \times 0.001)$$

which matches!

Continuing in this manner form

$$F_{G,\pi,\lambda_1,\lambda_2}(C, G) = F_{C}(G)F_{\pi,\lambda_1,\lambda_2}(C, G)$$

sum out \( G \) to obtain

$$F_{G,\pi,\lambda_1,\lambda_2}(C) = \sum_G F_{G}(G)F_{\pi,\lambda_1,\lambda_2}(C, G)$$

and normalise.
Performing exact inference: variable elimination

What’s the computational complexity now?

- For Bayesian networks with suitable structure we can perform inference in linear time and space.
- However in the worst case it is still \#P-hard.

Consequently, we may need to resort to approximate inference.

Approximate inference for Bayesian networks

Markov chain Monte Carlo (MCMC) methods also provide a method for performing approximate inference in Bayesian networks.

Say a system can be in a state \( S \) and moves from state to state in discrete time steps according to a probabilistic transition

\[
Pr(S \rightarrow S').
\]

Let \( \pi_t(S) \) be the probability distribution for the state after \( t \) steps, so

\[
\pi_{t+1}(S') = \sum_s Pr(s \rightarrow S') \pi_t(s).
\]

If at some point we obtain \( \pi_{t+1}(s) = \pi_t(s) \) for all \( s \) then we have reached a stationary distribution \( \pi \). In this case

\[
\forall s' \pi(s') = \sum_s Pr(s \rightarrow s') \pi(s).
\]

There is exactly one stationary distribution for a given \( Pr(S \rightarrow S') \) provided the latter obeys some simple conditions.

Approximate inference for Bayesian networks

The condition of detailed balance

\[
\forall s, s' \pi(s) Pr(s \rightarrow s') = \pi(s') Pr(s' \rightarrow s)
\]

is sufficient to provide a \( \pi \) that is a stationary distribution. To see this simply sum:

\[
\sum_s \pi(s) Pr(s \rightarrow s') = \sum_s \pi(s') Pr(s' \rightarrow s)
= \pi(s') \sum_s Pr(s' \rightarrow s)
= \pi(s')
\]

If all this is looking a little familiar, it’s because we now have another excellent application for the material in Mathematical Methods for Computer Science.

That course used the alternative term local balance.

Approximate inference for Bayesian networks

Recalling once again the basic equation for performing probabilistic inference

\[
Pr(Q|o_1, o_2, \ldots, o_m) \propto \sum_L Pr(Q, L, o_1, o_2, \ldots, o_m)
\]

where

- \( Q \) is the query.
- \( o_1, o_2, \ldots, o_m \) are the observations.
- \( L \) are the latent variables.
- \( 1/Z \) normalises the distribution.
- The query, observations and latent variables are a partition of the set \( V = \{V_1, V_2, \ldots, V_n\} \) of all variables.

We are going to consider obtaining samples from the distribution

\[
Pr(Q, L|o_1, o_2, \ldots, o_m).
\]
Approximate inference for Bayesian networks

The observations are fixed. Let the \textit{state} of our system be a specific set of values for a \textit{query variable} and the \textit{latent variables}

\[
S = (S_1, S_2, \ldots, S_{k1}) = (Q, L_1, L_2, \ldots, L_l)
\]

and define \(S_i\) to be the state vector with \(S_i\) removed

\[
\tilde{S}_i = (S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_{k1}).
\]

To move from \(s\) to \(s'\) we replace one of its elements, say \(s_i\), with a new value \(s_i'\) sampled according to

\[
s_i' \sim \Pr(S_i | \tilde{S}_i, o_1, \ldots, o_m)
\]

This has detailed balance, and has \(\Pr(Q, L | o_1, \ldots, o_m)\) as its stationary distribution.

It is known as \textit{Gibbs sampling}.

---

Approximate inference for Bayesian networks

To see that \(\Pr(Q, L | o)\) is the stationary distribution we just demonstrate \textit{detailed balance}:

\[
\pi(s) \Pr(s \rightarrow s') = \Pr(s | o) \Pr(s' | s, o)
\]

\[
= \Pr(s_i | \tilde{S}_i, o) \Pr(s'_i | \tilde{S}_i, o)
\]

\[
= \Pr(s_i | \tilde{S}_i, o) \Pr(s_i' | \tilde{S}_i, o) \Pr(s'_i | \tilde{S}_i, o)
\]

\[
= \Pr(s'_i | \tilde{S}_i, o) \Pr(s_i | \tilde{S}_i, o)
\]

\[
= \Pr(s' | s) \pi(s').
\]

As a further simplification we can exploit \textit{conditional independence}.

For example, sampling from \(\Pr(S_i | \tilde{S}_i, o)\) may be equivalent to sampling \(S_i\) conditional on some smaller set.

---

Approximate inference for Bayesian networks

So:

- We successively sample the query variable and the unobserved variables, conditional on the remaining variables.
- This gives us a sequence \(s_1, s_2, \ldots\) sampled according to \(\Pr(Q, L | o)\).

Finally, note that as

\[
\Pr(Q | o) = \sum_I \Pr(Q, L | o)
\]

we can just ignore the values obtained for the unobserved variables. This gives us \(q_1, q_2, \ldots\) with

\[
q_i \sim \Pr(Q | o).
\]
Markov random fields

Markov random fields (MRFs) (sometimes called undirected graphical models or Markov networks) provide an alternative approach to representing a probability distribution while expressing conditional independence assumptions.

We now have:

1. An undirected graph \( G = (N, E) \).
2. \( G \) has a node \( N_i \) for each RV.
3. For each maximal clique \( c \) in \( G \) there is a clique potential \( \phi_c(N_c) > 0 \) where \( N_c \) is the set of nodes in \( c \).
4. The probability distribution expressed by \( G \) is

\[
Pr(N) \propto \prod_c \phi_c(N_c).
\]

Example: 3 maximal cliques of size 2, 2 of size 3 and 1 of size 4.

\[
\begin{align*}
&\phi_1
&\phi_2
&\phi_3
&\phi_5
&\phi_6
&N_9
&N_1
&N_2
&N_4 \quad N_5
&N_6 \quad N_7
&N_3
&N_8

&Pr(N_1, \ldots, N_9) \propto \phi_1(N_1, N_4) \times \phi_2(N_3, N_6) \times \phi_3(N_7, N_8) \times \phi_4(N_1, N_2, N_3) \\
&\quad \times \phi_5(N_3, N_8, N_9) \times \phi_6(N_4, N_5, N_6, N_7).
\end{align*}
\]

Markov random fields—conditional independence

The test for conditional independence is now simple: if \( X, Y \) and \( Z \) are disjoint subsets of the RVs then:

1. Remove the nodes in \( Z \) and any attached edges from the graph.
2. If there are no paths from any variable in \( X \) to any variable in \( Y \) then

\( X \perp Y | Z \).

Final things to note:

1. MRFs have their own algorithms for inference.
2. They are an alternative to BNs for representing a probability distribution.
3. There are trade-offs that might make a BN or MRF more or less favourable.
4. For example: potentials offer flexibility because they don’t have to represent conditional distributions...
5. . . . BUT you have to normalize the distribution you’re representing.