Midway summary

Using a probability model to describe data → Models that depend on linear combinations of features → Parameter interpretation and identifiability

Supervised learning/ Generative modelling

- Too much complexity can lead to overfit
- Too little complexity can lead to underfit
- Not all the data (features) are equally informative/representative

Fitting a model's unknown parameters using MLE → Fitting via least squares (when appropriate)

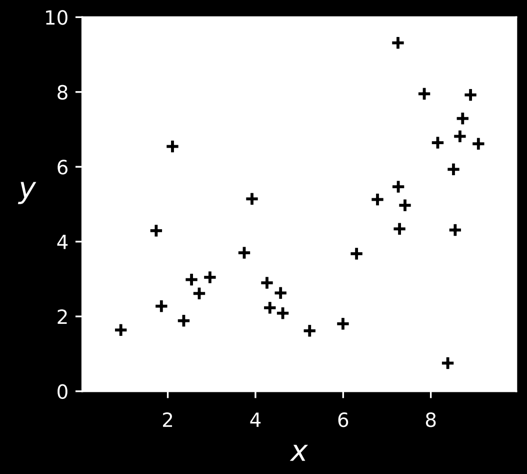
MATHS SKILLS

Reasoning about parameter uncertainty → Bayesianism

Bayesianism ← Bayes's rule

Bayesianism ← Monte Carlo

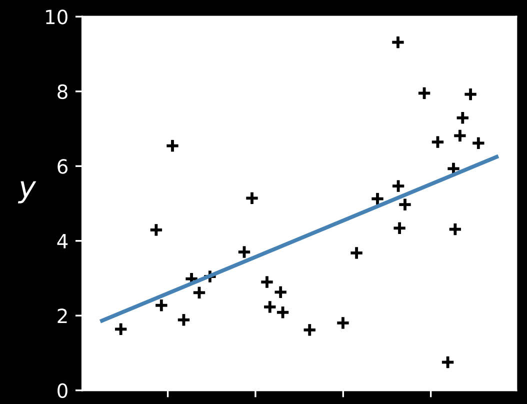
How should we compare models?



dataset of (x_i, y_i) pairs

MSE = $n^{-1} \sum_{i=1}^{n} (y_i - \text{pred}_i)^2$
measures how well a model fits...

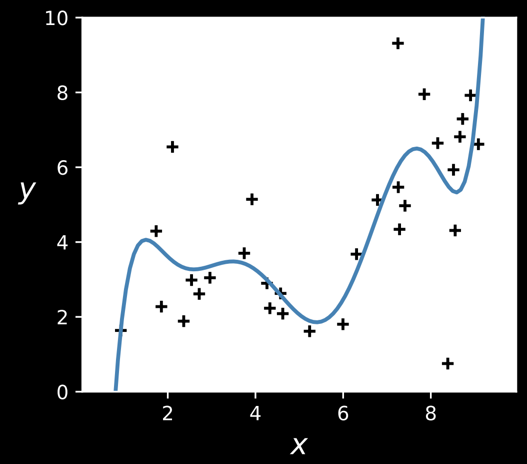
Or does it?



Model A:

$$Y_i \sim 1.62 + 0.49\, x_i + \text{Normal}(0, 2.39^2)$$

MSE large



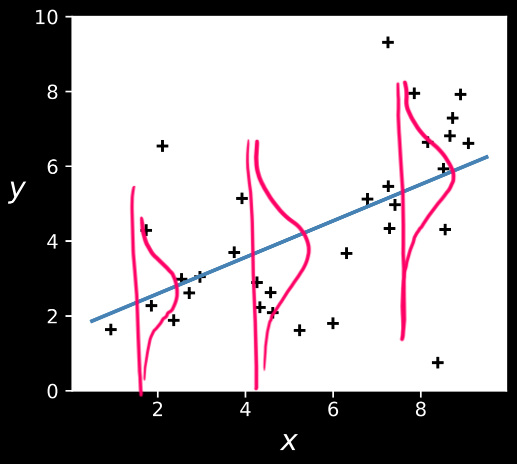
Model B:

$$Y_i \sim -38.5 + 95.7\, x_i - 84.8\, x_i^2 + 38.3\, x_i^3 - 9.5\, x_i^4 + 1.3\, x_i^5 - 0.09\, x_i^6 + 0.003\, x_i^7 + \text{Normal}(0, 0.31^2)$$

MSE small

This model doesn't just predict a value for y.

It predicts a distribution Y, at every x.

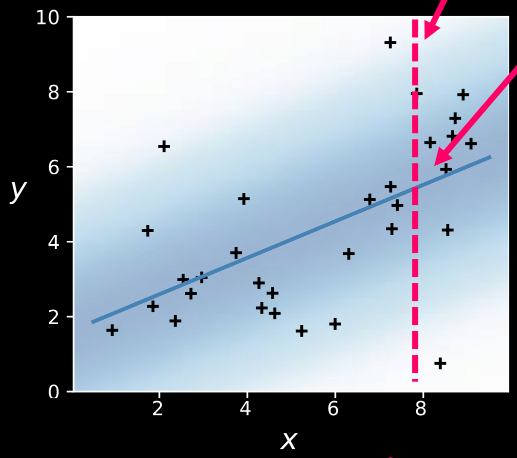


Model A:

$$Y_i \sim 1.62 + 0.49\, x_i + \text{Normal}(0, 2.39^2)$$

Area of
low likelihood

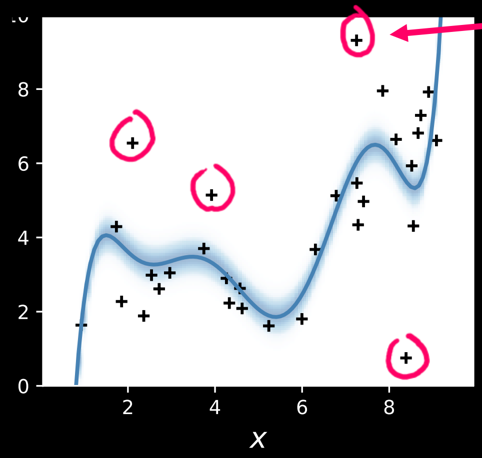
Area of
high likelihood



Model A:

$$Y_i \sim 1.62 + 0.49\, x_i + \mathrm{Normal}(0, 2.39^2)$$

These points are very unlikely to have been generated by this model



Model B:

$$Y_i \sim -38.5 + 95.7\, x_i - 84.8\, x_i^2 + 38.3\, x_i^3 - 9.5\, x_i^4 + 1.3\, x_i^5 - 0.09\, x_i^6 + 0.003\, x_i^7 + \mathrm{Normal}(0, 0.31^2)$$

There are several datapoints y_i where model B says "The likelihood of this y_i is vanishingly small." But these y_i did appear in the dataset. So model B is a bad explanation.

MODEL EVALUATION AND COMPARISON

After we fit a model, how do we decide if it's a good fit?

1. Evaluate the ~~mean square error~~ log likelihood of the dataset
2. Plot the ~~residuals~~ log likelihood of each datapoint, and look for systematic patterns.

“Bayesianism”?!

Reverend Thomas
Bayes, 1701–1761

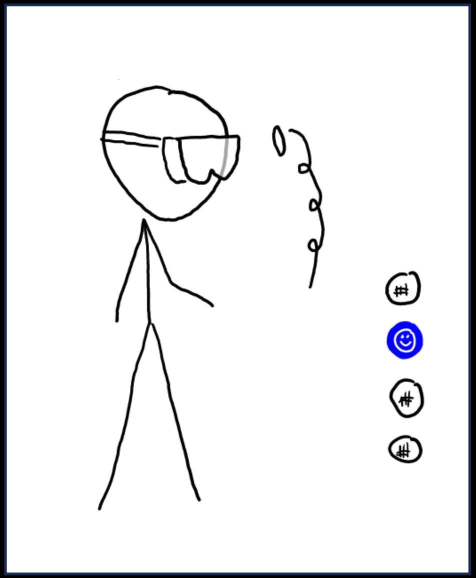
Bayes's rule for random variables

$$\mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x)\,\mathbb{P}(Y = y \mid X = x)}{\mathbb{P}(Y = y)}$$

$$\Pr_X(x \mid Y=y) = \frac{\Pr_X(x)\ \Pr_Y(y \mid X=x)}{\Pr_Y(y)}$$

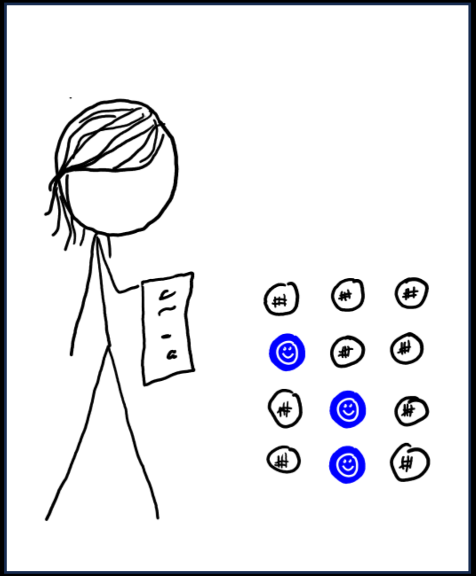
Bayesianism

Whenever there's an unknown parameter,
you should express your uncertainty about it
by treating it as a random variable.



I tossed four coins
and got one head.

Using a $\mathrm{Bin}(n, p)$ model, I estimate
the probability of heads is $\hat{p} = 25\%$

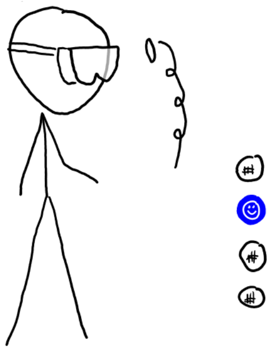


I tossed twelve coins
and got three heads.

Using a $\mathrm{Bin}(n, p)$ model, I estimate
the probability of heads is $\hat{p} = 25\%$

But surely, the more data we have, the more confident we should be!

By using random variables for unknown quantities, we can reason about confidence.



probability of heads, unknown

$$\Theta$$
$$\downarrow$$
$$X$$

number of heads from 4 coin tosses

We don't know the value of Θ, but we'll assume we know its distribution.

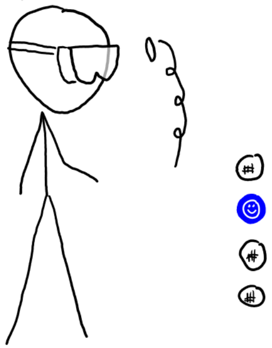
e.g. to express complete ignorance, $\Theta \sim \mathrm{Uniform}[0,1]$

We observed $X = 1$

We can use Bayes's rule to work out how confident we are about the unknown parameter's value …

$\mathbb{P}(\Theta \in [20\%, 30\%] \mid X = 1) = 21\%$

A more sophisticated way to reason about confidence is by using likelihood functions.

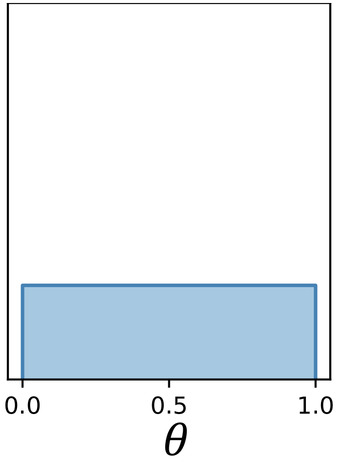


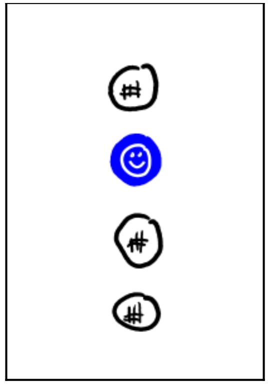
$$\Theta \sim U[0,1]$$

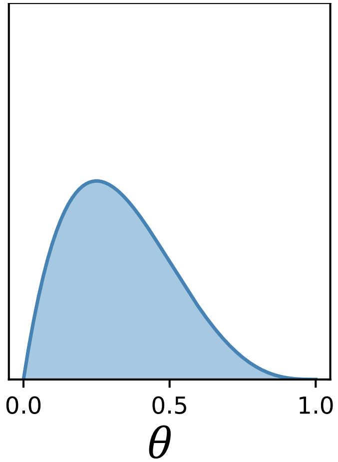
$$\downarrow$$

$$X \sim \mathrm{Bin}(n, \Theta)$$

prior belief $\Pr_\Theta(\theta)$ + data x → posterior belief $\Pr_\Theta(\theta|X=x)$

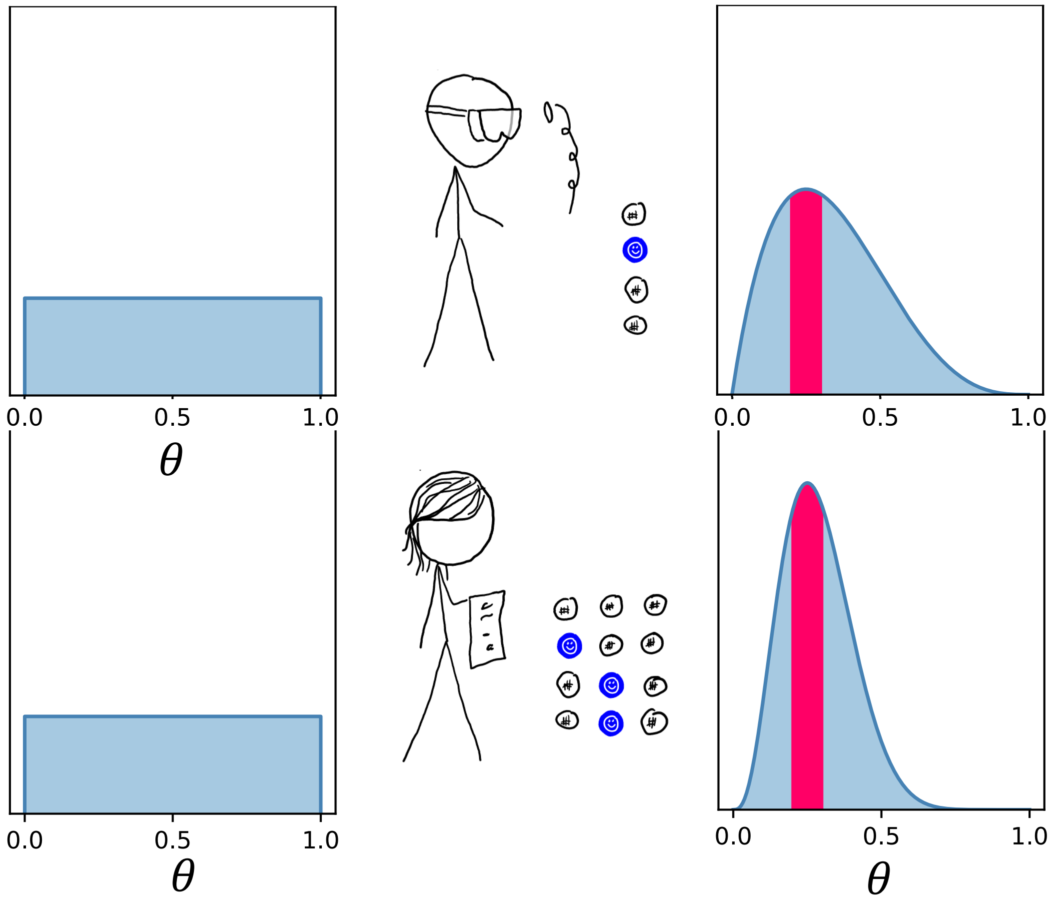






The data you see will affect your posterior belief about the parameter.

prior belief $\Pr_\Theta(\theta)$ + data x → posterior belief $\Pr_\Theta(\theta | X = x)$

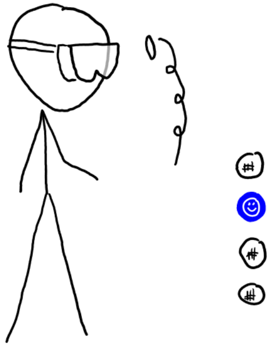


$\mathbb{P}(\Theta \in [.2, .3] \mid \text{data}) = 21\%$

$\mathbb{P}(\Theta \in [.2, .3] \mid \text{data}) = 33\%$

A tighter posterior distribution for Θ means we are more confident about its value.

By using random variables for unknown quantities, we can reason about confidence.



probability of
heads, unknown

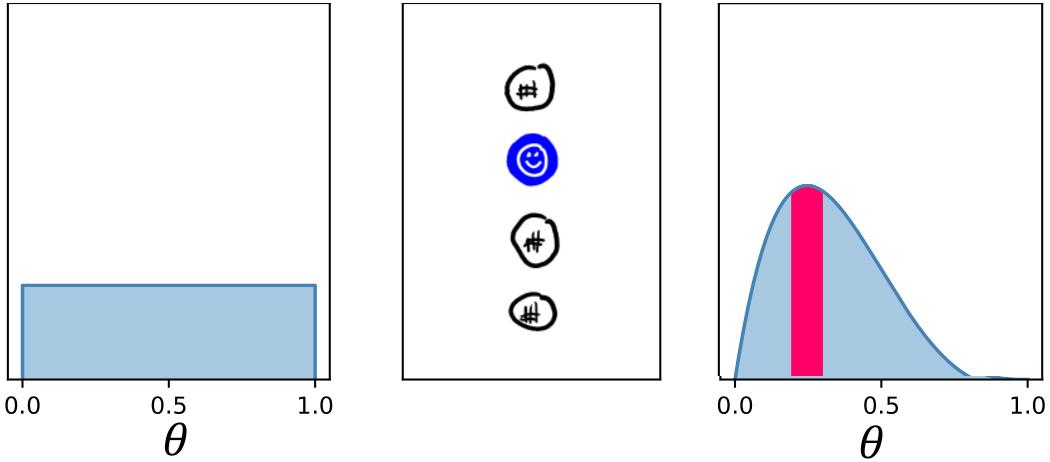
$$\Theta \sim U[0,1]$$

$$\downarrow$$

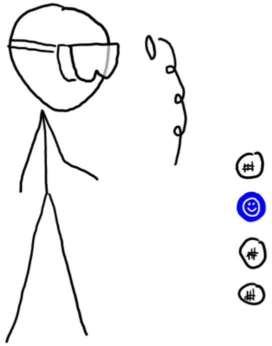
$$X \sim \text{Bin}(n, \Theta)$$

number of heads
from 4 coin tosses

prior belief $\Pr_\Theta(\theta)$ + data x → posterior belief $\Pr_\Theta(\theta | X = x)$



By using random variables for unknown quantities, we can reason about confidence.



probability of
heads, unknown

$$\Theta \sim U[0,1]$$

$$\downarrow$$

$$X \sim \mathrm{Bin}(n, \Theta)$$

number of heads
from 4 coin tosses

0. First write out our probability model for the data $\Pr_X(x|\Theta = \theta)$
1. Write out $\Pr_\Theta(\theta)$
2. Use the formula
   $\Pr_\Theta(\theta|X = x) = \kappa \Pr_\Theta(\theta) \Pr_X(x|\Theta = \theta)$
   then find κ to make this integrate to 1

This lets us calculate probabilities:

$$\mathbb{P}(\Theta \in \text{range}|X = x) = \int_{\theta \in \text{range}} \Pr_\Theta(\theta|X = x)\, d\theta$$

Exercise.

Consider the pair of random variables (Θ, X) where

$$\Theta \sim U[0,1], \qquad X \sim \text{Bin}(4, \Theta)$$

Find the distribution of $(\Theta | X = 1)$.

$\Pr_\Theta(\theta) = 1 \quad \text{for } \theta \in [0,1]$

$\Pr_X(x|\Theta = \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = 4\,\theta\,(1-\theta)^3 \quad \text{for } n=4, x=1$

$\Pr_\Theta(\theta|X = 1) = \kappa \Pr_\Theta(\theta) \Pr_X(1|\Theta = \theta)$

a function of θ

$= \kappa \times 1 \times 4\,\theta\,(1-\theta)^3$

$= \kappa'\,\theta\,(1-\theta)^3 \qquad \kappa'$ amalgamates non-θ terms.

$\int_0^1 \kappa'\,\theta(1-\theta)^3\,d\theta = 1 \quad \Rightarrow \quad \kappa' = \dfrac{1}{\int_0^1 \theta(1-\theta)^3\,d\theta}.$

Exercise.

Consider the pair of random variables (Θ, X) where

$$\Theta \sim U[0,1], \qquad X \sim \mathrm{Bin}(4,\Theta)$$

Find the distribution of $(\Theta|X=1)$.

Beta

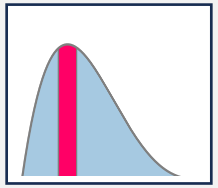
Probability density function	
Notation	$\mathrm{Beta}(\alpha,\beta)$
PDF	$\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ where $B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function.

$$\Pr_\Theta(\theta|X=1) = \kappa \Pr_\Theta(\theta)\,\Pr_X(1|\Theta=\theta)$$

$$= \kappa\,\theta\,(1-\theta)^3 = \underbrace{\kappa\,B(\alpha,\beta)}_{\text{so this constant must be 1 (otherwise this pdf wouldn't integrate to 1 wr.t. }\theta)} \overbrace{\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}^{\text{this is a standard pdf}} \quad \text{where } \alpha=2,\ \beta=4$$

Thus $(\Theta|X=1) \sim \mathrm{Beta}(\alpha=2, \beta=4)$

What is $\mathbb{P}(\Theta \in [.2,3]\mid X=1)$?



```
D = scipy.stats.beta(a=2,b=4)
D.cdf(.3) - D.cdf(.2)
```

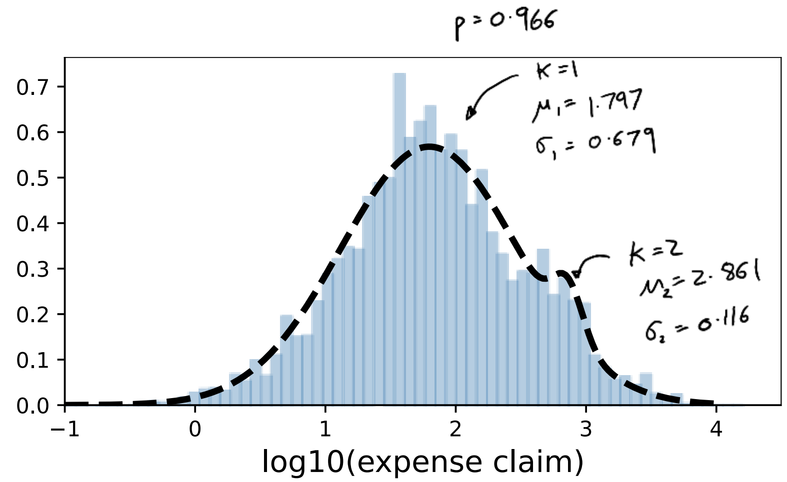
Exercise 5.2.3 (classification)

In a dataset of MP expense claims, let y_i be $\log_{10}$ of the claim amount in record i. A histogram of the y_i suggests we use a Gaussian mixture model with two components,

$$C = \begin{cases} 1 & \text{with prob } p \\ 2 & \text{with prob } 1-p \end{cases}$$

$$Y \sim \text{Normal}(\mu_C, \sigma_C^2)$$

Find the probability that a claim amount £5000 belongs to the component $c = 2$.



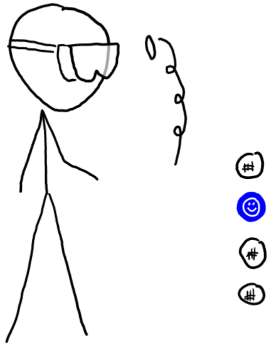
$$\Pr_C(c) =$$

$$\Pr_Y(y|C = c) =$$

$$\Pr_C(c|Y = y) = \kappa \Pr_C(c) \Pr_Y(y|C = c)$$

Exercise.

By using random variables for unknown quantities, we can reason about confidence.



probability of
heads, unknown

$\Theta \sim U[0,1]$

$\downarrow$

$X \sim \mathrm{Bin}(n, \Theta)$

number of heads
from 4 coin tosses

0. First write out our probability model for the data $\Pr_X(x|\Theta = \theta)$

1. Write out $\Pr_\Theta(\theta)$

2. Use the formula
$\Pr_\Theta(\theta|X = x) = \kappa \Pr_\Theta(\theta) \Pr_X(x|\Theta = \theta)$
then find κ to make this integrate to 1

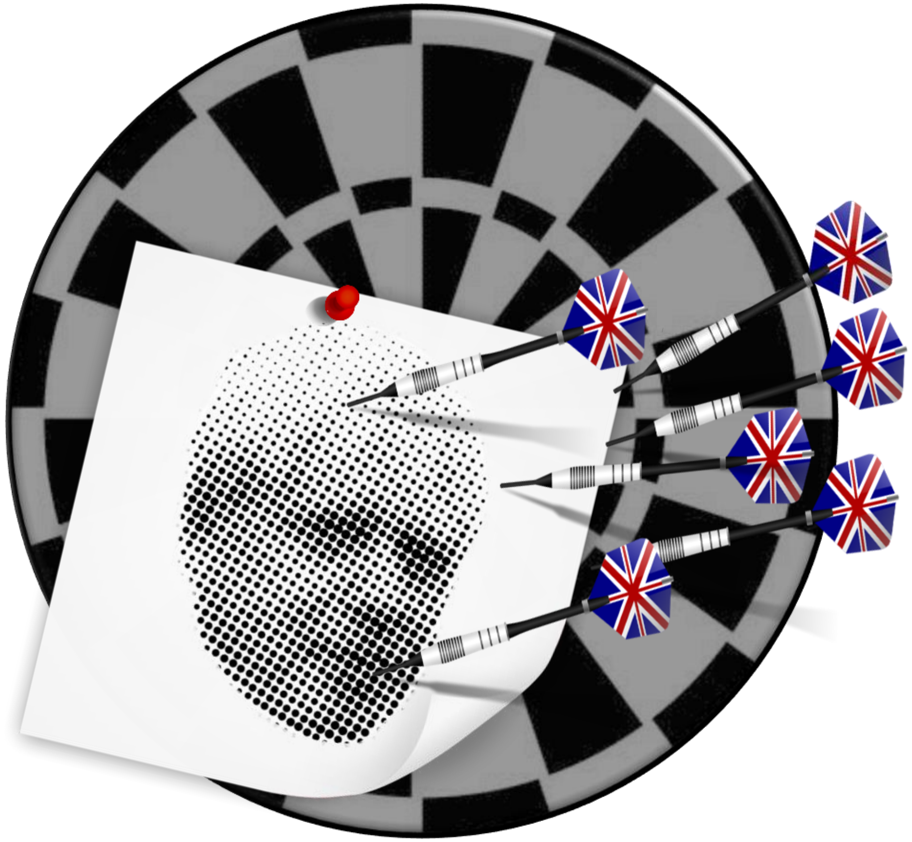
… but these are usually intractable

This lets us calculate probabilities:

$$\mathbb{P}(\Theta \in \text{range}|X = x) = \int_{\theta \in \text{range}} \Pr_\Theta(\theta|X = x)\, d\theta$$

§6. Computational methods

What's the chance that a randomly thrown dart will hit the mystery object A?



Let X be the location of a randomly thrown dart, and let $x_1, \dots, x_n$ be some throws.

The probability of hitting A is

$$\mathbb{P}(X \in A) \approx \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \in A}$$

$$1_{x_i \in A} = \begin{cases} 1 & \text{if } x_i \in A \\ 0 & \text{else} \end{cases}$$

```
# Let X ~ N(μ = 1, σ = 3). What is ℙ(X > 5)?
x = np.random.normal(loc=1, scale=3, size=10000)
i = (x > 5)   10,000 Booleans
np.mean(i)
```

typecast bool to int.

Expectation

For a real-valued random variable X

$$\mathbb{E}X = \begin{cases} \sum_x x \Pr_X(x)\,, & \text{if } X \text{ is discrete} \\ \int_x x \Pr_X(x)\, dx\,, & \text{if } X \text{ is continuous} \end{cases}$$

Law of the Unconscious Statistician

For a random variable X and a real-valued function h

$$\mathbb{E}h(X) = \begin{cases} \sum_x h(x)\,\Pr_X(x)\,, & \text{if } X \text{ is discrete} \\ \int_x h(x)\,\Pr_X(x)\,dx\,, & \text{if } X \text{ is continuous} \end{cases}$$

If we want to know the average properties of a rich random variable (random images, random texts), we have to use real-valued property readout functions h(X) so that we can take averages.

Monte Carlo integration

$$\mathbb{E}h(X) \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

where $x_1, \ldots, x_n$ is a sample drawn from X

let $h(x) = 1_{x \in A}$

$x_1 \dots x_n$ sampled from X

By Monte Carlo,

$$\mathbb{E}\, h(X) \approx \frac{1}{n} \sum_{i=1}^{n} h(x_i)$$

$$\frac{1}{n} \sum_{i} 1_{x_i \in A}$$

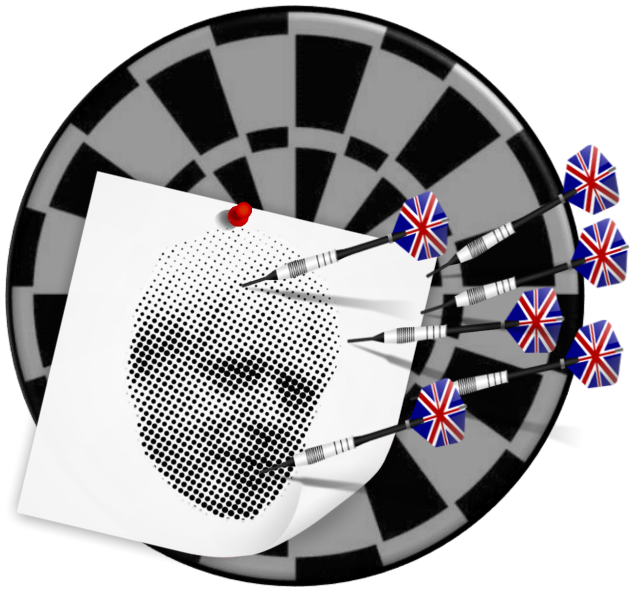
let $Y = h(X)$

$$\begin{aligned} \mathbb{E}\, Y &= 0 \times \mathbb{P}(Y=0) + 1 \times \mathbb{P}(Y=1) \\ &= \mathbb{P}(Y=1) \\ &= \mathbb{P}(h(X)=1) \\ &= \mathbb{P}(1_{X \in A} = 1) \\ &= \mathbb{P}(X \in A) \end{aligned}$$

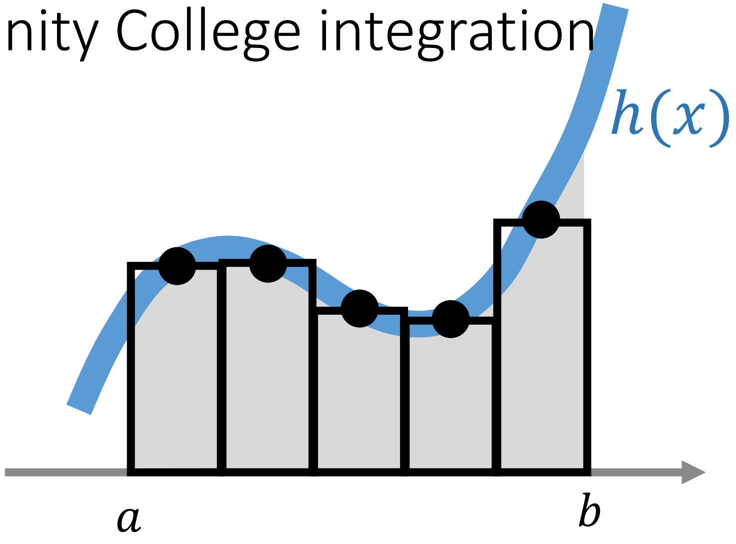
Let X be the location of a randomly thrown dart, and let $x_1, \dots, x_n$ be some throws.

The probability of hitting A is

$$\mathbb{P}(X \in A) \approx \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \in A}$$



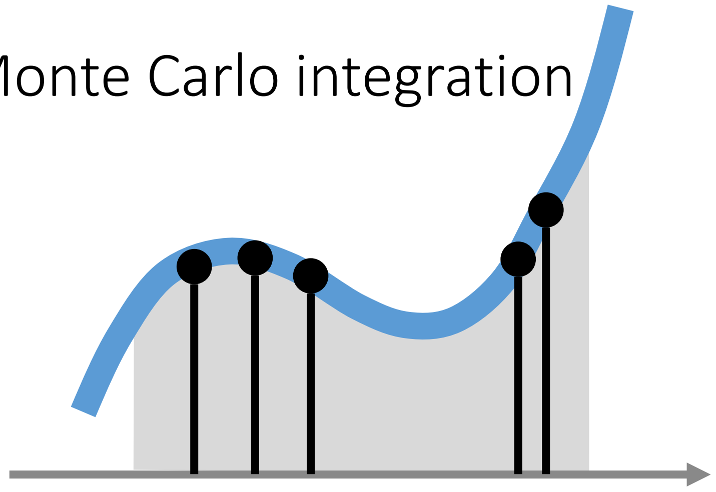
Trinity College integration



$$\int_{x=a}^{b} h(x)\,dx \approx \sum_{i=1}^{n} h(x_i)\,\frac{b-a}{n}$$

where x_i is the midpoint of interval i

Monte Carlo integration



Let's instead approximate this integral using Monte Carlo. Let $X \sim U[a,b]$.

By Monte Carlo,

$$\underbrace{\mathbb{E}h(X)}_{\int_{x=a}^{b} h(x)\,\Pr_X(x)\,dx \;=\; \int_{x=a}^{b} h(x)\,\frac{1}{b-a}\,dx} \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i) \quad \text{where } x_1,\dots,x_n \text{ sampled from } X$$

Thus,

$$\int_{x=a}^{b} h(x)\,dx \approx \frac{b-a}{n}\sum_{i=1}^{n} h(x_i)$$

COMPUTATIONAL METHODS

- If we want $\mathbb{E}h(X)$ but the maths is too complicated, we can approximate it using $x_1, \dots, x_n$ sampled from X
- The approximation for $\mathbb{E}h(X)$ also tells us how to estimate probabilities, since $\mathbb{P}(X \in A) = \mathbb{E}1_{X \in A}$
- For computational Bayes, we need something a bit fancier: *weighted samples*

Bayes classification rule is optimal w.r.t. minimising the classification error probability

Is it always the best choice?

Hint

Try to perform classification on a dataset used to determine whether a landslide is occurring or not

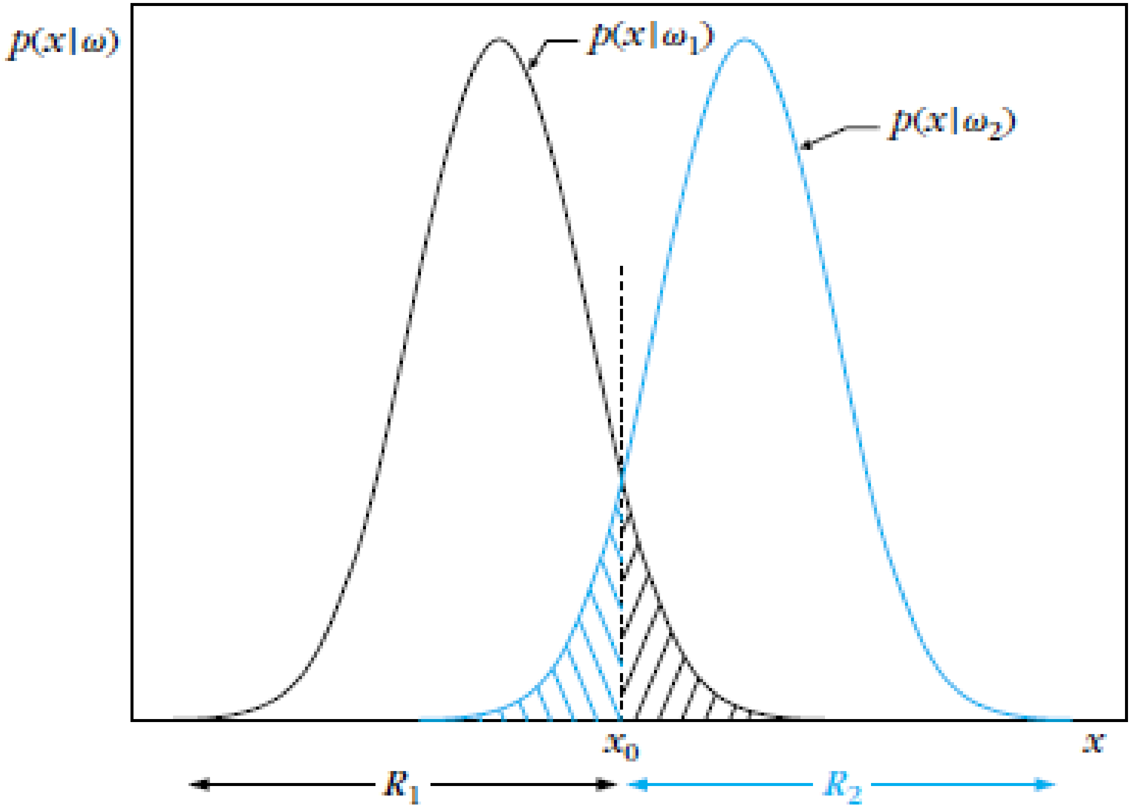
I.e., sometimes not all the errors have the same consequences

In that case, it would be good to have a different metric that could take into account this information → from *error* to *risk*

Good thing about Bayes' theory: it can incorporate risk

The next slides (OPTIONAL) explain how

Hint



S. Theodoridis, K. Koutroumbas, "Pattern Recognition", Academic press, 2009 – chapter 2

2-class classification

FIGURE 2.1

Example of the two regions R_1 and R_2 formed by the Bayesian classifier for the case of two equiprobable classes.

$x \in R_1 \Rightarrow$ assign ω_1

$x \in R_2 \Rightarrow$ assign ω_2

Classification errors : $x \in R_2$ but belongs to ω_1 plus $x \in R_1$ but belongs to ω_2 (unavoidable).

probability of error $\quad P_e = P(\omega_2) \underbrace{\int_{-\infty}^{x_0}}_{R_1} p(x|\omega_2)\,dx + P(\omega_1) \underbrace{\int_{x_0}^{+\infty}}_{R_2} p(x|\omega_1)\,dx$

Bayes classification rule is optimal with respect to minimising the classification error probability.

Proof: $$P_e = \int_{-\infty}^{x_0} p(x|\omega_2)P(\omega_2)\,dx + \int_{x_0}^{+\infty} p(x|\omega_1)P(\omega_1)\,dx$$

[Bayes rule] $$P_e = \int_{R_1} P(\omega_2|x)\,p(x)\,dx + \int_{R_2} P(\omega_1|x)\,p(x)\,dx$$

[$\int_{-\infty}^{\infty} p(x)\,dx \equiv 1$] we know that for pdf's

$$\int_{R_1} P(\omega_1|x)\,p(x)\,dx + \int_{R_2} P(\omega_1|x)\,p(x)\,dx \equiv P(\omega_1)$$

hence $\quad P_e = P(\omega_1) - \int_{R_1} \left(P(\omega_1|x) - P(\omega_2|x) \right) p(x)\, dx$

$\Rightarrow$ P_e minimum when R_1 is chosen such that

$$P(\omega_1|x) > P(\omega_2|x)$$

$\blacksquare$

Bayes classification rule is optimal with respect to minimising the classification error probability.

Is it always the best choice?

Now Consider a new problem where some errors have far worse consequences than others, eg. often in medical problems.

Let’s consider warning systems for landslides

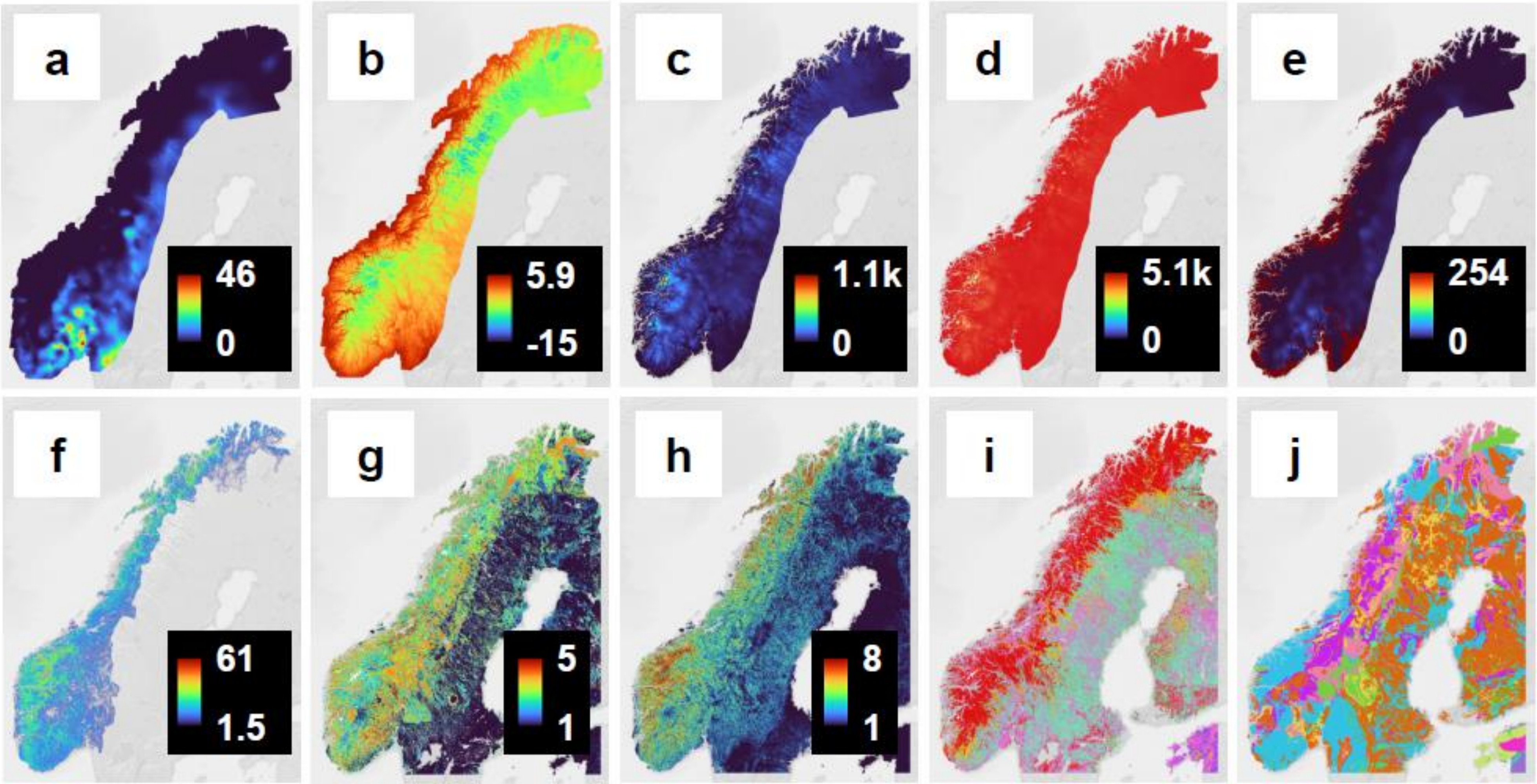
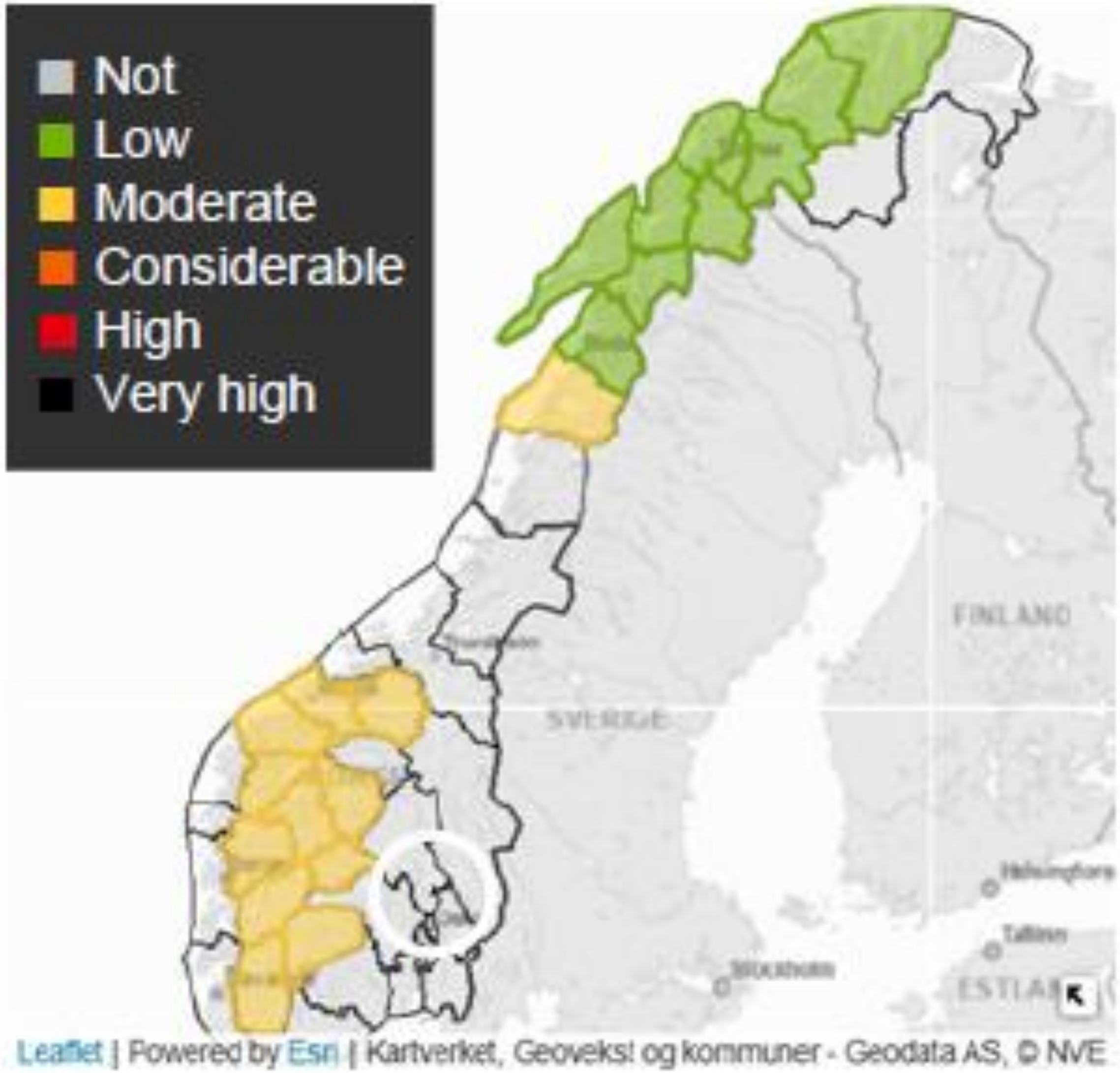
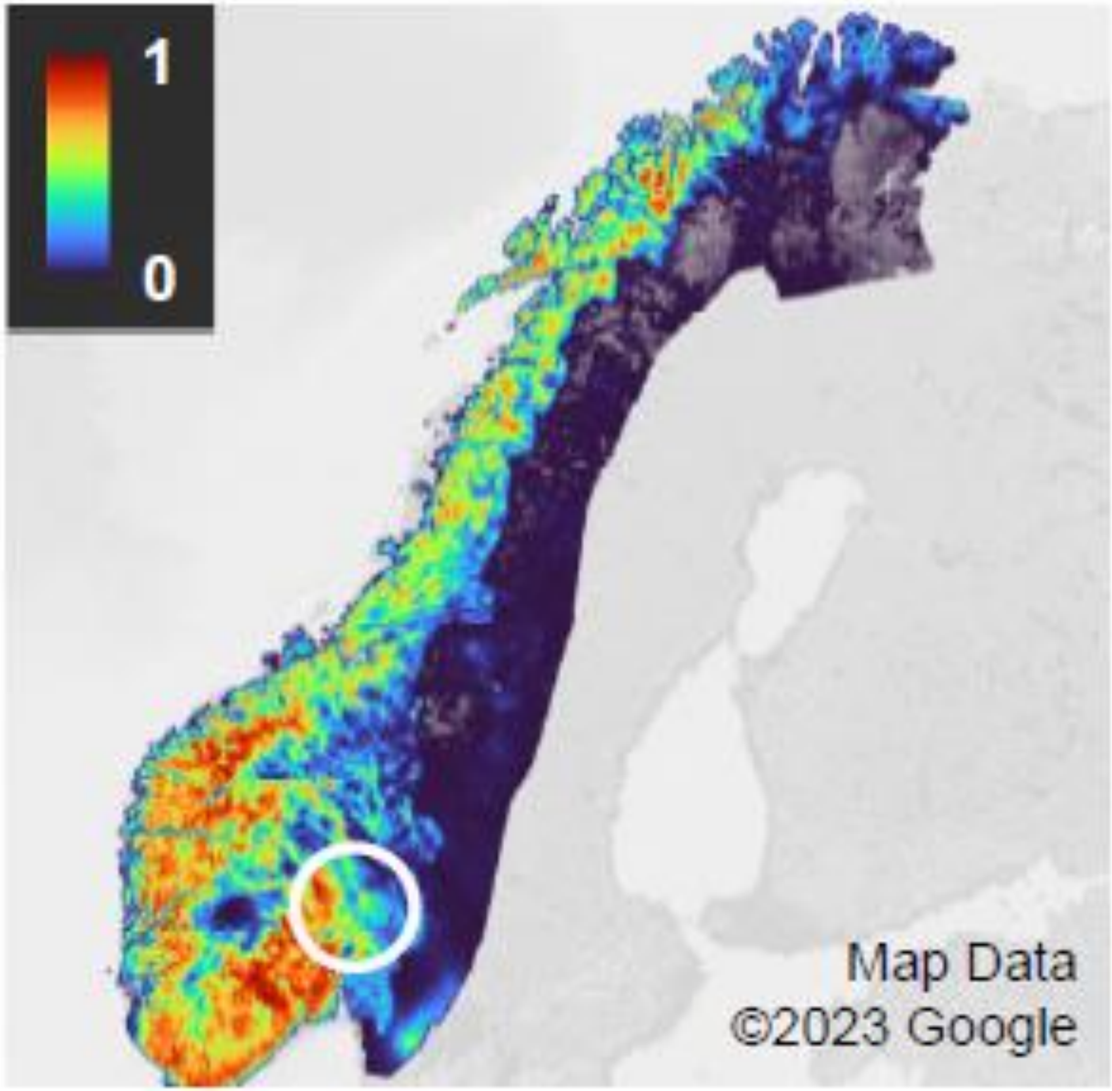


Figure 4.2: Date-specific maps of **(a)** total rainfall (mm/day), **(b)** mean temperature (Celsius degrees), **(c)** snow depth (cm/day), **(d)** snow water equivalent (mm/day), and **(e)** fresh snow water equivalent (mm/day) for December 30, 2020. Static maps of **(f)** steepness (degrees), **(g)** ELSUS susceptibility (categorical integer), **(h)** slope angle class (categorical integer), **(i)** land cover class (categorical integer), and **(j)** lithology class (categorical integer).

Credits: J. Dimasaka





Credits: J. Dimasaka

Now Consider a new problem where some errors have far worse consequences than others, eg. often in medical problems.

$\Rightarrow$ minimise the average risk.

- M class problem
- R_j, $j = 1, \ldots, M$ are M regions in feature space connected/associated to each class ω_j.
- Assume that $\underline{x}$ belongs to ω_j, but lies in $R_{i \neq j}$
- A 'penalty' or 'loss' λ_{ji} is associated with each wrong class ω_i

- The Risk associated with w_j is

$$r_j = \sum_{i=1}^{M} \lambda_{ji} \boxed{\int_{R_i} p(x|w_j)\,dx}$$

Probability of feature vector $\underline{x}$ from *j*-th class being classified in *i*-th class.

Aim is to choose R_j such that the average risk

$$r = \sum_{k=1}^{M} r_k P(\omega_k) = \sum_{i=1}^{M} \int_{R_i} \underbrace{\left(\sum_{k=1}^{M} \lambda_{ki} \, p(x|\omega_k) P(\omega_k) \right)}_{l_i} dx$$

is minimised.

This is achieved by minimising each integral, (with positive arguments in l_i)

⇒ Choose regions such that

$$\underline{x} \in R_i \quad \text{if} \quad l_i = \sum_{k=1}^{M} \lambda_{ki} \, p(\underline{x}|\omega_k) P(\omega_k) < l_j \quad \forall i \neq j$$

"Bayes classification rule with loss."

Note that if $\lambda_{ki} = 1 - \delta_{ki}$; $\delta_{ki} = \begin{cases} 1 & k = i \\ 0 & \text{else} \end{cases}$

then $r = P_e$.

"Kronecker's delta"

<u>2 class example</u>: with Loss matrix $L = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$

$$\lambda_{12} \, P(\omega_1) \, p(\underline{x} | \omega_1) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \lambda_{21} \, P(\omega_2) \, p(\underline{x} | \omega_2)$$