Computer Vision

Computer Science Tripos Part II

Dr Christopher Town

10. Bayesian inference. Classifiers; probabilistic methods.

UNIVERSITY OF CAMBRIDGE

Vision as *Going beyond the data*

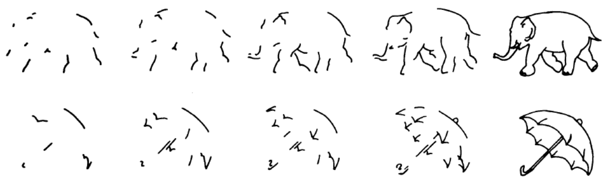
Vision as inference
- Incorporating Prior Knowledge
- Knowledge-based approaches risk being brittle or underspecified:

SYMBOL GROUNDING PROBLEM
and
FRAME PROBLEM

Vision as *Going beyond the data*

Intractable problems can be made tractable using priors such as "objects cannot just disappear, they more likely occlude each other" or "head like objects are usually found on top of body like objects".

Bayesian priors provide one means to do this, since the learning (or specification) of metaphysical principles (truths about the nature of the world) can steer the integration of evidence appropriately, making an intractable problem soluble.



Decisions under uncertainty

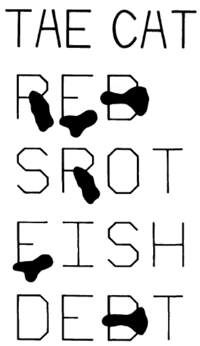
- the nature of the data or signals available
- the inherent problem of classifying or recognising them
- the unpredictability of the future
- the fact that objects and events have associated likelihoods of occurrence (depending on context)
- the uncertainty of causation
- the inherent incompleteness or imperfection of processing
- possible undecidability of a problem, given all available data
- the "ill-posed" nature of many tasks
- inherent trade-offs such as speed versus accuracy

Decisions under uncertainty

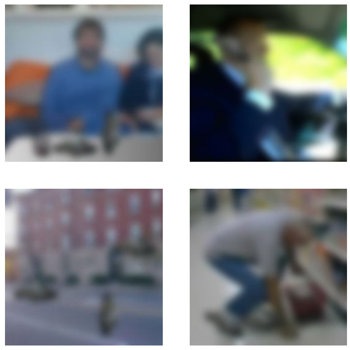
Examples of decisions-under-uncertainty in vision:

- Medical diagnosis; radiology: Is this a tumour? Does the cost of a possible False Alarm (taking a biopsy, frightening the patient unnecessarily) exceed the cost of possibly missing an early diagnosis? What should you do if the odds are 99% that it is just a benign cyst; but if it is a tumour, missing it now could be fatal?

- Military decision-making: a plane is seen approaching your aircraft carrier very low on the horizon and at high speed. Is it friend or foe? How should the costs of the two possible types of error (shooting down one of your own planes, vs allowing the whole aircraft carrier to be sunk) be balanced against their relative probabilities, when making your decision?



The multiple personalities of a blob



The multiple personalities of a blob

Probabilities: Review

- *relative frequency:* sample the random variable a great many times and tally up the fraction of times that each of its different possible values occurs, to arrive at the probability of each.
- *degree-of-belief:* probability is the plausibility of a proposition or the likelihood that a particular state (or value of a random variable) might occur, even if its outcome can only be decided once (e.g. the outcome of a particular horse-race).

Probabilities: Review

$\boxed{\text{Product Rule:}}$

$$p(A,B) = \text{“joint probability of } \textit{both A and B}\text{”}$$
$$= p(A|B)p(B)$$

or equivalently,

$$= p(B|A)p(A)$$

$\boxed{\text{Sum Rule:}}$

If event A is conditionalized on a number of other events B, then the total probability of A is the sum of its joint probabilities with all B:

$$p(A) = \sum_B p(A,B) = \sum_B p(A|B)p(B)$$

Dr Chris Town

The Bayesian view

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}$$

$$posterior = \frac{likelihood * prior}{evidence}$$

Iterative integration of new evidence: posteriors become new priors

Dr Chris Town

The Bayesian view

Examples of useful priors in vision:

- Some objects and events are far more likely than others
- Matter cannot just disappear, but does routinely become occluded
- Objects rarely change their surface colour
- Uniform texturing on a complex surface shape is more likely than highly non-uniform texturing on a simple shape
- Rigid rotation in three dimensions is a ``better explanation" for deforming boundaries than actual boundary deformations in the object itself

Dr Chris Town

Statistical decision theory

Some classification tasks don't have useable priors, e.g. iris recognition

-> Need to decide which class a feature vector is more likely to belong to, even if we don't have any useful priors about the relative likelihoods of the possible object classes or interpretations.

Dr Chris Town

Statistical decision theory

The degree of match between two feature vectors must be computed and formally evaluated to make a decision of "same" or "different." Almost always, there is some similarity between "different" patterns, and some dissimilarity between "same" patterns. This creates a <u>decision environment</u> with four possible outcomes:

1. <u>Hit</u> (True accept): Actually same; decision "same".
2. <u>Miss</u> (False reject): Actually same; decision "different".
3. <u>False Alarm</u> (False accept): Actually different; decision "same".
4. <u>Correct Reject</u> (True reject): Actually different; decision "different".

We would like to maximise the probability of outcomes 1 and 4, because these are correct decisions. We would like to minimise the probability of outcomes 2 and 3, because these are incorrect decisions ("Type II" and "Type I" errors).

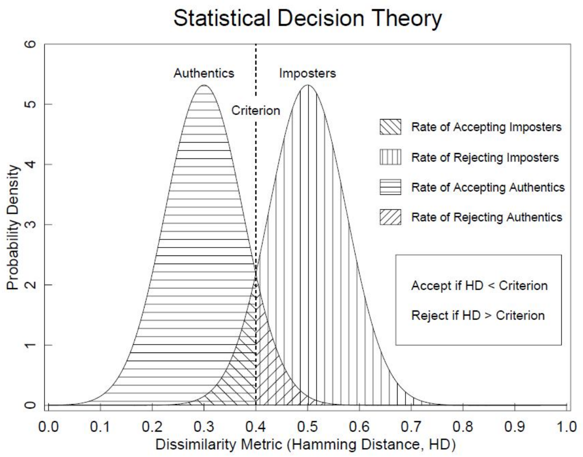
Dr Chris Town

Statistical decision theory

We can adjust our decision threshold (become more liberal or more conservative) to reflect the costs and benefits of the four possible outcomes. But adjusting the decision threshold has <u>coupled</u> effects on the four outcomes:

- Increasing the "Hit" rate will also increase the "False Alarm" rate.
- Decreasing the "Miss" rate will also decrease the "Correct Reject" rate.

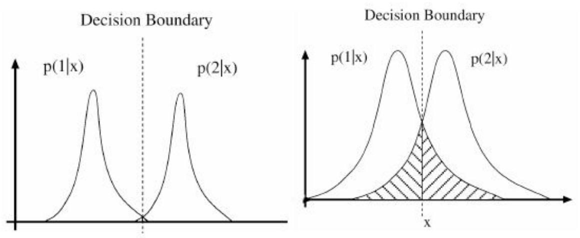
Dr Chris Town



Dr Chris Town

Bayes Risk

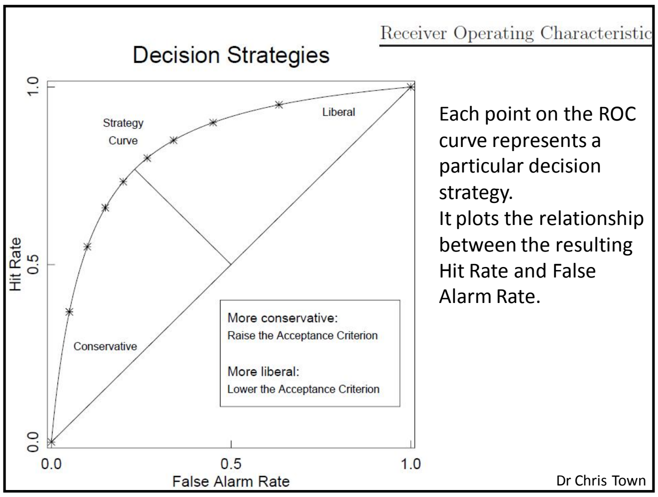
Some errors may be inevitable: the shaded area is called the Bayes risk



Probability density functions (area under each curve sums to 1)

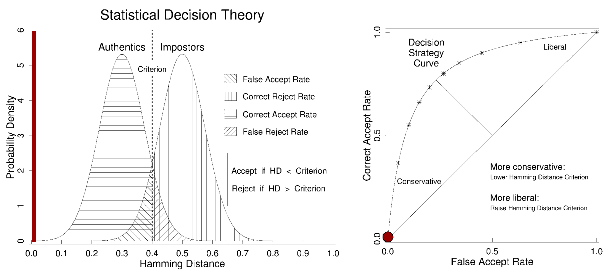
Slide credit: David Lowe

Dr Chris Town



Each point on the ROC curve represents a particular decision strategy.
It plots the relationship between the resulting Hit Rate and False Alarm Rate.

Dr Chris Town



Generating ROC (or DET) curves requires moving the decision threshold, from conservative to liberal, to see the trade-off between False Reject Rate and False Positive Rate.

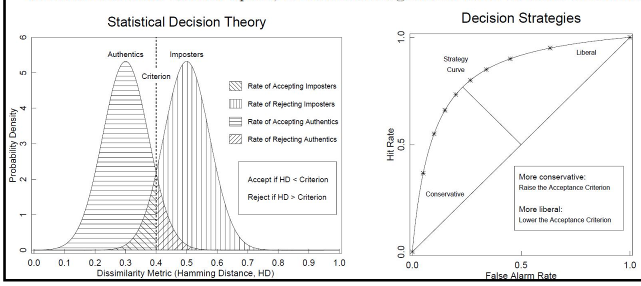
The slope of the ROC curve is the likelihood ratio: ratio of the two density distributions at a given decision threshold criterion.

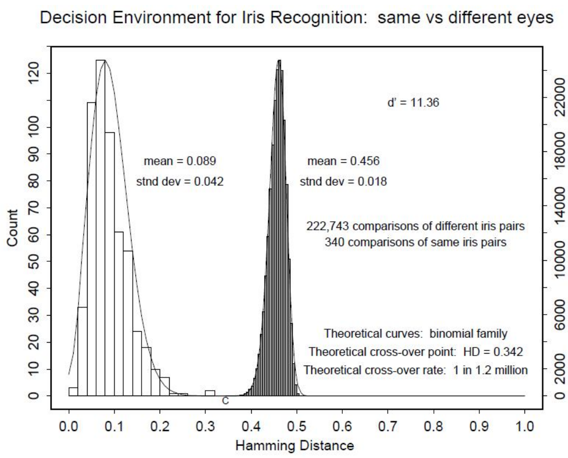
Dr Chris Town

Decidability

$$d' = \frac{|\mu_2 - \mu_1|}{\sqrt{\frac{1}{2}(\sigma_2^2 + \sigma_1^2)}}$$

where the two distributions are characterised by means μ_1 and μ_2 and standard deviations σ_1 and σ_2. An improvement in d' can result either from pushing the two distributions further apart, or from making one or both of them narrower.





Dr Chris Town

These considerations illustrate what might be called the "Primary Law of Pattern Recognition":

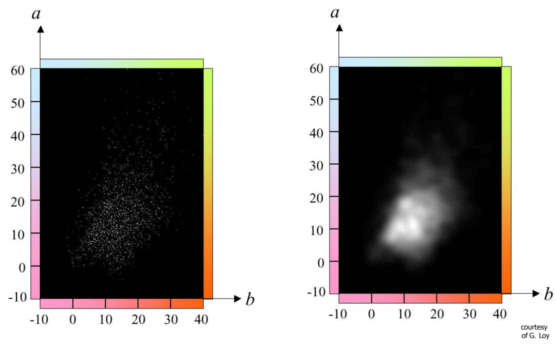
The key factor is the relation between within-class variability and between-class variability. Pattern recognition can be performed reliably only when the between-class variability is larger than the within-class variability.

Application: Skin Colour Histograms

- Skin has a very small range of (intensity independent) colours, and little texture
  - Compute colour measure, check if colour is in this range, check if there is little texture (median filter)
  - Get class conditional densities (histograms), priors from data (counting)
- Classifier is
  - if $p(\text{skin}|\boldsymbol{x}) > \theta$, classify as skin
  - if $p(\text{skin}|\boldsymbol{x}) < \theta$, classify as not skin

Slide credit: David Lowe

Skin Colour Models



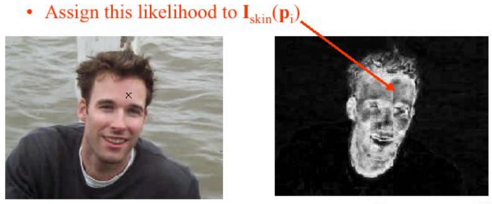
Skin chrominance points   Smoothed, [0,1]-normalized

Slide credit: David Lowe

Skin Colour Classification

For every pixel $\mathbf{p}_i$ in $\mathbf{I}_{test}$

- Determine the chrominance values (a_i, b_i) of $\mathbf{I}_{test}(\mathbf{p}_i)$
- Lookup the skin likelihood for (a_i, b_i) using the skin chrominance model.
- Assign this likelihood to $\mathbf{I}_{skin}(\mathbf{p}_i)$



Slide credit: David Lowe

Results

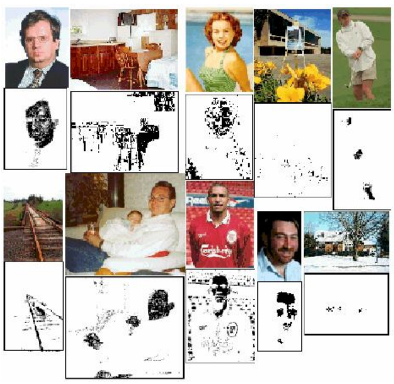


Figure from "Statistical color models with application to skin detection," M.J. Jones and J. Rehg, Proc. Computer Vision and Pattern Recognition, 1999 copyright 1999, IEEE

Slide credit: David Lowe

ROC Curves

(Receiver operating characteristics)

Plots trade-off between false positives and false negatives for different values of a threshold

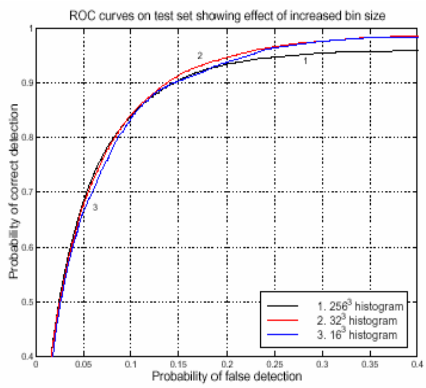
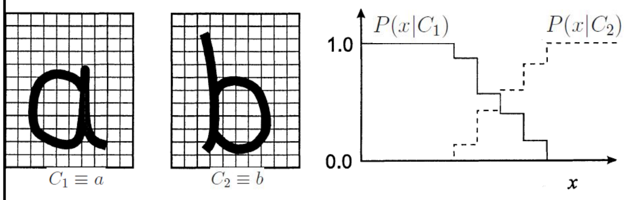


Figure from "Statistical color models with application to skin detection," M.J. Jones and J. Rehg, Proc. Computer Vision and Pattern Recognition, 1999 copyright 1999, IEEE

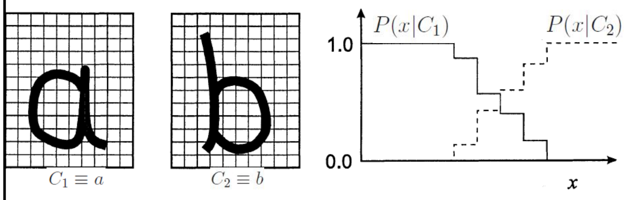
Slide credit: David Lowe

Bayesian pattern classifiers



Define the prior probabilities $P(C_1)$ and $P(C_2)$ as their relative proportions (summing to 1). If we had to guess which character had appeared without our even seeing it, we would always just guess the one with the higher prior probability. Thus since in fact an 'a' is about 4 times more frequent than a 'b' in English, and these are the only two cases in this two-class inference problem, we would set $P(a) = 0.8$ and $P(b) = 0.2$.

Bayesian pattern classifiers



For each class separately, we can measure how likely any particular feature sample value x will be, by empirical observation of instances from each class. This gives us $P(x|C_1)$ and $P(x|C_2)$.

Finally, we need to know the unconditional probability $P(x)$ of any measurement value x. We can calculate this by the probability "sum rule:"

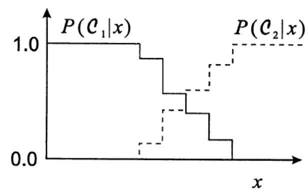
$$P(x) = \sum_{k=1}^{2} P(x|C_k)P(C_k)$$

Bayesian pattern classifiers

$$P(x) = \sum_{k=1}^{2} P(x|C_k)P(C_k)$$

Now we have everything we need to apply Bayes' Rule to calculate the likelihood of either class membership, given some observation x, factoring in the prior probabilities $P(C_k)$, the unconditional probability $P(x)$ of the observed data, and the likelihood of the data given either of the classes, $P(x|C_k)$. The likelihood of class C_k given the data x, is the posterior probability $P(C_k|x)$:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{17}$$

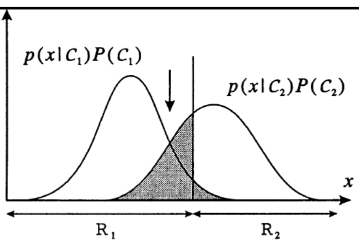


We minimise the probability of misclassification if we assign each new input x to the class with the highest posterior probability. Assign x to class C_k if:

$$P(C_k|x) > P(C_j|x) \qquad \forall j \neq k$$

Since the denominator in Bayes' Rule (equation 17) is independent of C_k, we can rewrite this *minimum misclassification criterion* simply as:

$$P(x|C_k)P(C_k) > P(x|C_j)P(C_j) \qquad \forall j \neq k$$



If the decision boundary that we choose is as indicated by the vertical line above, then the total error is equal to the total shaded area. Let R1 and R2 be the regions of x on either side of our decision boundary. Then the total probability of error is:

$$\begin{aligned} P(error) &= P(x \in R_2, C_1) + P(x \in R_1, C_2) \\ &= P(x \in R_2|C_1)P(C_1) + P(x \in R_1|C_2)P(C_2) \\ &= \int_{R_2} P(x|C_1)P(C_1)dx + \int_{R_1} P(x|C_2)P(C_2)dx \end{aligned}$$

Discriminant functions and decision boundaries

If some set of functions $y_k(x)$ of the data x are constructed, one function for each class C_k, such that classification decisions are made by assigning an observation x to class C_k if

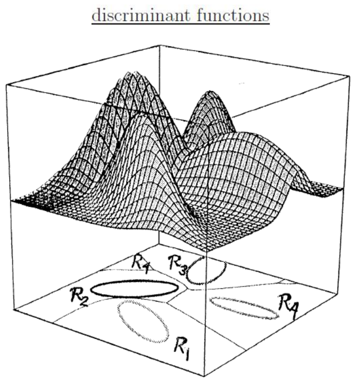
$$y_k(x) > y_j(x) \quad \forall j \neq k,$$

those functions $y_k(x)$ are called discriminant functions. The decision boundaries between data regions R_j and R_k are defined by those loci in the (normally multi-dimensional) data x at which $y_k(x) = y_j(x)$. A natural choice for discriminant functions would be the posterior probabilities:

$$y_k(x) = P(C_k|x)$$

Equivalently since the denominator $P(x)$ in Bayes' Rule is independent of k, we could choose

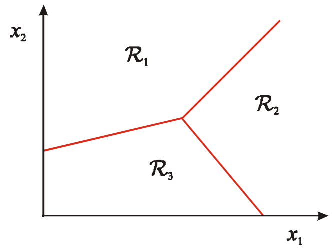
$$y_k(x) = P(x|C_k)P(C_k)$$



Dr Chris Town

Classification

- Assign input vector to one of two or more classes
- Any decision rule divides input space into *decision regions* separated by *decision boundaries*

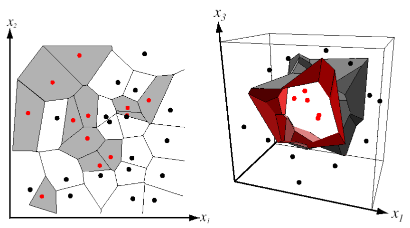


Slide credit: Svetlana Lazebnik

Dr Chris Town

Nearest Neighbour Classifier

- Assign label of nearest training data point to each test data point.



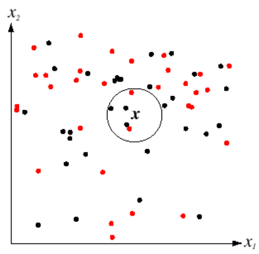
Voronoi partitioning of feature space
for 2-category 2-D and 3-D data

Slide credit: David Lowe

Dr Chris Town

K-Nearest Neighbours

- For a new point, find the k closest points from training data
- Labels of the k points "vote" to classify
- Works well provided there is lots of data and the distance function is good



Slide credit: David Lowe

Dr Chris Town

The Naïve Bayes Model

- Assume that each feature is conditionally independent given the class

$$p(w_1,\ldots,w_N \mid c) = \prod_{i=1}^{N} p(w_i \mid c)$$

Slide credit: Li Fei-Fei

Dr Chris Town

The Naïve Bayes Model

- Assume that each feature is conditionally independent given the class

$$c^* = \arg\max_c \; p(c) \prod_{i=1}^{N} p(w_i \mid c)$$

MAP decision

Prior prob. of the object classes

Likelihood of i-th feature given the class

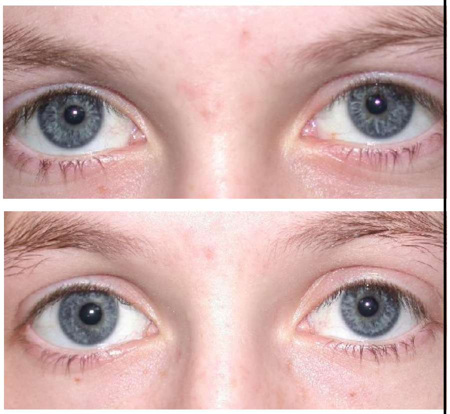
Estimated by empirical frequencies of features in images for a given class

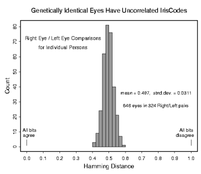
Csurka et al. 2004

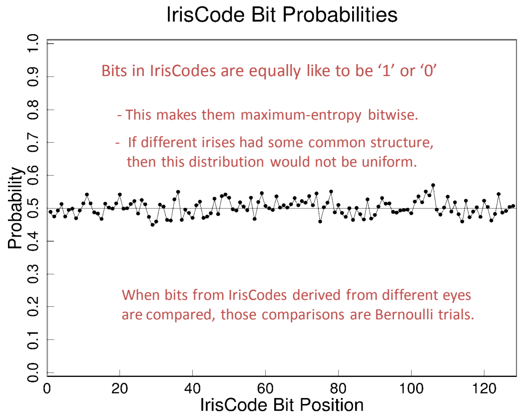
Dr Chris Town

Genetically identical eyes have iris patterns that are uncorrelated in detail:

Monozygotic Twins B
(18 year-old women)







IrisCode Bit Comparisons are Bernoulli Trials

Jacob Bernoulli (1645-1705) analyzed *coin-tossing* and derived the binomial distribution. If the probability of "heads" is p, then the likelihood that a fraction $x = m/N$ out of N tosses will turn up "heads" is:

University of Groningen

$$P(x) = \frac{N!}{m!(N-m)!}\, p^m \, (1-p)^{(N-m)}$$