PROBABILITY THEORY

Consider $P(A, B)$ the joint probability of two events A and B:

* When the events are independent: $P(A, B) = P(A)P(B)$
* Otherwise, conditional probabilities should be used ($P(A \mid B)$ means the probability of A given the occurrence of B):
  + $P(A, B) = P(A \mid B)P(B)$
  + $P(A, B) = P(B \mid A)P(A)$
* From these, Bayes' Rule follows: $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$

MAXIMUM LIKELIHOOD CLASSIFICATION

* Suppose that there is a (model of a) physical process that produces some *outcome* M.
* One measures some data D related to the outcome.
* One wants to know which outcome has produced D.
* The *maximum likelihood* principle states: $\max_M P(M \mid D)$.
* With the application of Bayes' Rule: $\max_M \dfrac{P(D \mid M)P(M)}{P(D)}$

PROBABILITY DISTRIBUTIONS

Basics:

* The set of all possible outcomes of an experiment is the *sample space.*
* A *random variable* X is a function from the sample space to the real numbers.
* X may be discrete or continuous.
* *Distribution function* of a random variable: $\Phi(x) = P(X \le x)$
* *Density function:* $p(x) = \dfrac{d\Phi(x)}{dx}$.

Well-known distributions:

* binomial, Poisson
* Gaussian

MEAN AND VARIANCE

* Discrete case:
  + Expected value or mean: $E[X] = m = \sum_i x_i P(x_i)$
  + Variance: $\sigma^2 = E[(X - m)^2] = \sum_i (x_i - m)^2 P(x_i)$
* Continuous case:
  + Expected value or mean: $E[X] = m = \int_{-\infty}^{\infty} p(x)dx$
  + Variance: $\sigma^2 = E[(X - m)^2] = \int_{-\infty}^{\infty} (x - m)^2 p(x)dx$

MEAN AND VARIANCE ESTIMATION

* Suppose that n measurements have been made: $x_1, \ldots, x_n$.
* The estimated mean is then: $m = \frac{1}{n}\sum_i x_i$
* And the estimated variance: $\sigma^2 = \frac{1}{n}\sum_i (x_i - m)^2$

INFORMATION THEORY

* Deals with issues like efficiency and redundancy in encoding.
* Consider e.g. the retina: it has 10^8 cells, but there are only 10^6 cells in the optic nerve. Hence some kind of data compression takes place to be more efficient in the transport of information.
* *Redundancy* is necessary to recover the information in received messages in the presence of noise.

CHANNEL CAPACITY AND ENTROPY

* *Channel capacity* for a channel with m locations with n symbols per location: $C_m = m \log_2 n$.
* Suppose that a source can generate N different messages $x_1, \ldots, x_N$. The lower the probability for the occurrence of some message, the higher its information content. If the probability of x_i is p_i, $(1 \le i \le N)$, then: $I_i = \log_2 \frac{1}{p_i}$.
* The *entropy* H is the expected value of the information content: $H = \sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{N} p_i \log_2 p_i$.
* Requirements for channel: $C_m \ge H$.

MAXIMAL ENTROPY

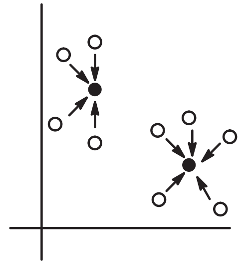
* Entropy: $H = \sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{N} p_i \log_2 p_i$
* It can be shown that $0 \le H \le \log_2 N$.
* The lower bound is reached when one of the messages has probability one and the rest probability zero.
* The upper bound is reached when all messages are equally probable: $p_i = \frac{1}{N}$.

REVERSIBLE CODES

* The theory can be used for the design of *reversible codes*, codes from which the original messages can be exactly recovered.
* Suppose that the messages x_i $(1 \le i \le N)$ have a length l_i. The average message length is then: $\sum_{i=1}^{N} p_i l_i$.
* It holds: $\sum_{i=1}^{N} p_i l_i \ge H = -\sum_{i=1}^{N} p_i \log p_i$.
* The optimum situation (equality) occurs when: $l_i = -\log p_i$.
* An example of a reversible code is *Huffman coding*.

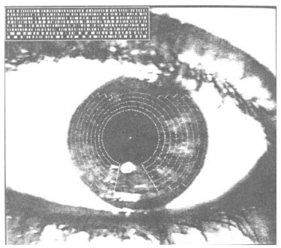
IRREVERSIBLE CODES

* In many biological systems codes do not need to be reversible. *Irreversible* codes are more efficient.
* The use of *prototypes*, also called *vector quantization*, leads to irreversible codes.



IRIS RECOGNITION EXAMPLE (1)

* Based on the work of Daugman [1].
* Image-processing techniques localize the iris in the image and apply 2-D Gabor transforms on the iris at different scales.
* The most significant bits of the coefficients obtained are collected into a 256 byte (2048 bit) code, the *feature vector.* These vectors are the prototypes.



[1] Daugman, J.G., High Confidence Visual Recognition of Persons by a Test of Statistical Independence, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.15(11), pp.1148–1161, (November 1993).

IRIS RECOGNITION EXAMPLE (2)

* *Question:* how much information do these 256 bytes of the feature vector contain?
* Tests reveal that, for each bit position, the average bit value is close to 0.5.
* Consider the *normalized Hamming distance* (HD) of two bit strings $a_1, \ldots, a_B$ and $b_1, \ldots, b_B$ with the same length B:

$$HD = \frac{1}{B} \sum_{i=1}^{B} a_i \oplus b_i$$

* One expects a binomial distribution for the HDs (the probability of a 1 is p, the probability of a 0 is $q = 1 - p$, the fraction of bits equal to 1 is $x = \frac{n}{B}$):

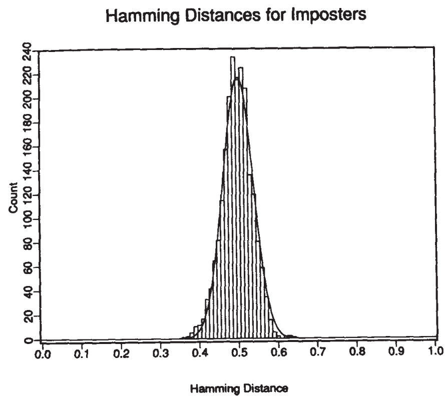
$$p(x) = \frac{B!}{n!(B-n)!} p^n q^{(B-n)}$$

* A binomial distribution has a variance of:

$$\sigma^2 = \frac{pq}{B}$$

IRIS RECOGNITION EXAMPLE (3)

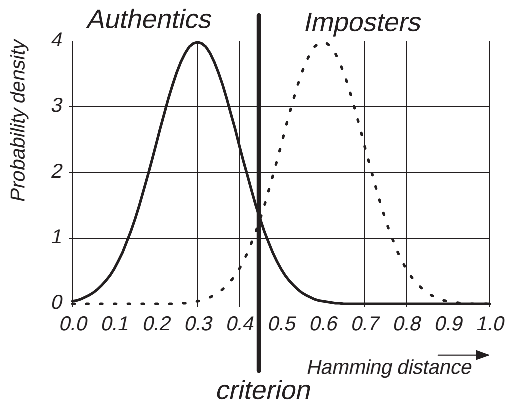
* Computing the HDs for the “imposters”, the feature vectors originating from different persons gives the next distribution.



IRIS RECOGNITION EXAMPLE (4)

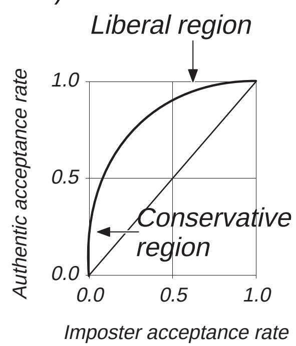
* From the distribution it can be derived that $B = 173$.
* So, the feature vector is highly redundant.

STATISTICAL DETECTION THEORY (1)



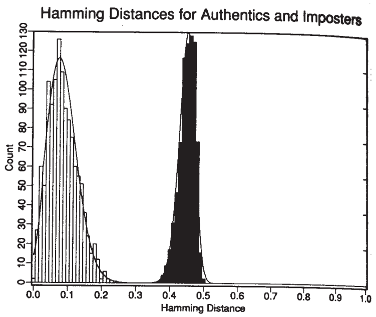
STATISTICAL DETECTION THEORY (2)

* Four outcomes:
  + Acceptance of authentic
  + Acceptance of imposter (false acceptance)
  + Rejection of authentic (false rejection)
  + Rejection of imposter
* The choice of decision criterion affects the probabilities of each outcome, from very conservative to very liberal.
* This is visualized in a receiver operating characteristic (ROC curve).



IRIS RECOGNITION EXAMPLE (5)

* It turns out that the two distributions are fully disjoint:



* This high level of reliability is a consequence of the long feature vector.
* The imposters curve is centered around 0.45 rather than 0.5 because of a “best of k” strategy to compensate for rotations.

MINIMUM DESCRIPTION LENGTH (1)

* When the goal is to learn a message D, one can store the message as such or one can try to find a compression method M for a more efficient storage.
* The most efficient situation corresponds to a minimal description of the method itself and compressed data.

$$L(M, D) = L(M) + L(D \text{ encoded using } M)$$

* Suppose that the possible models have a probability distribution. Then there is also a probability distribution of the models given the data and Bayes’ Rule can be used:

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)}$$

* The goal is to maximize $P(M \mid D)$ or to determine $\max_M P(D \mid M)P(M)$.

MINIMUM DESCRIPTION LENGTH (2)

* To maximize a quantity also means to maximize its logarithm:

$$\arg\max_M P(D \mid M)P(M) = \arg\max_M [\log P(D \mid M) + \log P(M)]$$

* or to minimize its negative:

$$\arg\min_M [-\log P(D \mid M) - \log P(M)]$$

* As the minimum length for a message that has a probability P is $-\log P$, it follows that choosing the best model according to Bayes’ Rule amounts to applying the *minimum description length* (MDL) principle.

RESIDUALS (1)

* Suppose that a model M has been chosen. It maps data points x_i $(1 \le i \le N)$ to prototypes m_i. The differences are called *residuals*. Suppose that the sum of the residuals has a Gaussian distribution with variance α:

$$P(D \mid M) = \left(\frac{1}{2\pi\alpha}\right)^{\frac{N}{2}} e^{-\frac{1}{2\alpha}\sum_{i=1}^{N}(x_i - m_i)^2}$$

* Consider now that the model is a neural network parameterized by the weights w_i $(1 \le i \le W)$. This gives a distribution of all neural networks, supposed to be Gaussian with variance β:

$$P(M) = \left(\frac{1}{2\pi\beta}\right)^{\frac{W}{2}} e^{-\frac{1}{2\beta}\sum_{i=1}^{W} w_i^2}$$

RESIDUALS (2)

* The application of the MDL principle gives:

$$\arg\min_{M}[-\log P(D \mid M) - \log P(M)] = \frac{1}{2\alpha}\sum_{i=1}^{N}(x_i - m_i)^2 + \frac{1}{2\beta}\sum_{i=1}^{W} w_i^2 + const.$$

* This explains why neural network training aims at minimizing the squared sum between actual and desired outputs for the training data (the error).
* Note that there is a trade-off between minimizing the error and the cost of the model.

IMAGE CODING EXAMPLE (1)

* One decides to encode an $n \times n$ image with pixels I_{ij} $(1 \le i,j \le n)$ with m neurons and reconstruct it as follows:

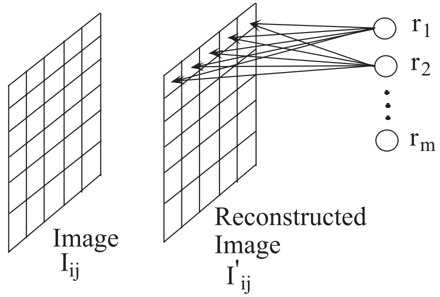


IMAGE CODING EXAMPLE (2)

* The pixels in the reconstructed image: $I'_{ij} = \sum_{k=1}^{m} w_{ijk} r_k$.
* According to the MDL principle, the w_{ijk} and r_k should be chosen such as to minimize:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(I_{ij} - I'_{ij})^2 + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} w_{ijk}^2 + \sum_{k=1}^{m} r_k^2$$