**Probabilistic and statistical methods for recognition:**

**Statistical approach:** In statistical pattern recognition, a pattern is represented by a set of d features, or attributes, viewed as a d-dimensional feature vector. Well-known concepts from statistical decision theory are utilized to establish decision boundaries between pattern classes. The recognition system is operated in two modes: training (learning) and classification (testing) (see Fig. 1). The role of the preprocessing module is to segment the pattern of interest from the background, remove noise, normalize the pattern, and any other operation which will contribute in defining a compact representation of the pattern. In the training mode, the feature extraction/selection module finds the appropriate features for representing the input patterns and the classifier is trained to partition the feature space. The feedback path allows a designer to optimize the preprocessing and feature extraction/selection strategies. In the classification mode, the trained classifier assigns the input pattern to one of the pattern classes under consideration based on the measured features.



The decision making process in statistical pattern recognition can be summarized as follows: A given pattern is to be assigned to one of c categories ω1, ω 2 … ω c based on a vector of d feature values x = (x1, x2 … xd). The features are assumed to have a probability density or mass (depending on whether the features are continuous or discrete) functions conditioned on the pattern class. Thus, a pattern vector x belonging to class ωi is viewed as an observation drawn randomly from the class-conditional probability function p(x|ωi).

**Probabilistic approach:** Define a joint probability distribution on X×Y denoted PX, Y. Let (X, Y ) denote a pair of random variables distributed according to PX,Y. We will also have use for marginal and conditional distributions. Let PX denote the marginal distribution on X, and let PY|X denote the conditional distribution of Y given X. For any distribution P, let p denote its density function with respect to the corresponding dominating measure; e.g., Lebesgue measure for continuous random variables or counting measure for discrete random variables. Define the expectation operator:



Suppose that (X, Y) are distributed according to PX,Y((X, Y) PX,Y for short). Our goal is to find a map so that f(X) Y with high probability. Ideally, we would chose f to minimize the risk R(f) = E[***l***(f(X), Y )]. However, in order to compute the risk (and hence optimize it) we need to know the joint distribution PX,Y. In many problems of practical interest, the joint distribution is unknown, and directly minimizing the risk is not possible. Suppose that we have some “training examples”, that is, samples from the distribution PX,Y. Specifically, consider n samples  distributed independently and identically (i.i.d.) according to the otherwise unknown PX,Y . These are called training data, and denote the collection by . Let’s also define a collection of candidate mappings F. We will use the training data Dn to pick a mapping  that we hope will be a good predictor. This is sometimes called the Model Selection problem. Note that the selected model fn is a function of the training data:

****

**Bayes approach:**

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known.

Let *ω*1*, ..., ωc* be the finite set of *c* states of nature (“categories”) and *α*1*, ..., αa* be the finite set of *a* possible actions. The loss function *λ*(*αi|ωj*) describes the loss incurred for taking action *αi* when the state of nature is *ωj* . Let the feature vector **x** be a *d*-component vector-valued random variable, and let *p*(**x***|ωj*) be the state conditional probability density function for **x** — the probability density function for **x** conditioned on *ωj* being the true state of nature. As before, *P*(*ωj*) describes the prior probability that nature is in state *ωj* . Then the posterior probability *P*(*ωj |***x**) can be computed from *p*(**x***|ωj*) by Bayes’ formula:



where the evidence is now



Suppose that we observe a particular **x** and that we contemplate taking action *αi*. If the true state of nature is *ωj* , by definition we will incur the loss *λ*(*αi|ωj* ).Since *P*(*ωj |***x**) is the probability that the true state of nature is *ωj* , the expected loss associated with taking action *αi* is merely:



In decision-theoretic terminology, an expected loss is called a *risk*, and *R*(*αi|***x**) is called the *conditional risk*. Stated formally, our problem is to find a decision rule against *P*(*ωj*) that minimizes the overall risk. A general *decision rule* is a function *α*(**x**) that tells us which rule action to take for every possible observation. To be more specific, for every **x** the *decision function α*(**x**) assumes one of the *a* values *α*1*, ..., αa*. The overall risk *R* is the expected loss associated with a given decision rule. Since *R*(*αi|***x**) is the conditional risk associated with action *αi*, and since the decision rule specifies the action, the overall risk is given by:



where *d***x** is our notation for a *d*-space volume element, and where the integral extends over the entire feature space. Clearly, if *α*(**x**) is chosen so that *R*(*αi*(**x**)) is as small as possible for every **x**, then the overall risk will be minimized. This justifies the following statement of the *Bayes decision rule*: To minimize the overall risk, compute the conditional risk:



for *i* = 1,...,*a* and select the action *αi* for which *R*(*αi|***x**) is minimum.

**Gaussian distribution and Mehalonobis distance:**

**Gaussian distribution** is a continuous probability distribution that describes data that cluster around a mean or average. The graph of the associated probability density function is bell-shaped, with a peak at the mean, and is known as the **Gaussian function** or **bell curve**.

The normal distribution can be used to describe, at least approximately, any variable that tends to cluster around the mean.

The probability density function for a normal distribution is given by the formula:



For mean 0 and standard deviation 1 the distribution is called standard normal distribution:



The continuous probability density function of the **normal distribution** is the Gaussian function



where *σ* > 0 is the standard deviation, the real parameter *μ* is the expected value, and



The probability density function has notable properties including:

* symmetry about its mean *μ*
* the mode and median both equal the mean *μ*

the inflection points of the curve occur one standard deviation away from the mean (i.e., at *μ* − *σ* and *μ* + *σ*).

Some properties of the normal distribution:

1. If \scriptstyle X\,\sim\,N(\mu, \sigma^2)and \scriptstyle aand \scriptstyle bare real numbers, then \scriptstyle a X + b\,\sim\,N(a \mu + b, (a \sigma)^2)(see expected value and variance).
2. If \scriptstyle X\,\sim\,N(\mu_X, \sigma^2_X)and \scriptstyle Y\,\sim\,N(\mu_Y, \sigma^2_Y)are independent normal random variables, then:
   1. Their sum is normally distributed with \scriptstyle U = X + Y\,\sim\,N(\mu_X + \mu_Y, \sigma^2_X + \sigma^2_Y)(proof). Thus the normal distribution is infinitely divisible. Interestingly, the converse holds: if two independent random variables have a normally-distributed sum, then they must be normal themselves; this is known as Cramér's theorem.
   2. Their difference is normally distributed with \scriptstyle V = X - Y\,\sim\,N(\mu_X - \mu_Y, \sigma^2_X + \sigma^2_Y).
3. If the variances of *X* and *Y* are equal, then *U* and *V* are independent of each other (and equal variance); in the bivariate distribution, this corresponds to a change of coordinates in the plane from \scriptstyle (x,y)to \scriptstyle (x+y,x-y), or equivalently a change of axes (rotation and dilation).
4. The Kullback-Leibler divergence is given by: \scriptstyle D_{\rm KL}( X \| Y ) =
   { 1 \over 2 } \left[2 \log \left( { \sigma_Y \over \sigma_X } \right) + \frac{\sigma^2_X}{\sigma^2_Y} +
   \frac{\left(\mu_Y - \mu_X\right)^2}{\sigma^2_Y} - 1\right].
5. If \scriptstyle X\,\sim\,N(0, \sigma^2_X)and \scriptstyle Y\,\sim\,N(0, \sigma^2_Y)are independent normal random variables, then:
   1. Their product \scriptstyle X Yfollows a distribution with density \scriptstyle pgiven by
   2. \scriptstyle p(z) = \frac{1}{\pi\,\sigma_X\,\sigma_Y} \; K_0\left(\frac{|z|}{\sigma_X\,\sigma_Y}\right),where \scriptstyle K_0is a modified Bessel function of the second kind.
   3. Their ratio follows a Cauchy distribution with \scriptstyle \frac{X}{Y}\,\sim\,\mathrm{Cauchy}(0, \frac{\sigma_X}{\sigma_Y}). Thus the Cauchy distribution is a special kind of ratio distribution.
6. If \scriptstyle X_1, \dots, X_nare independent standard normal variables, then \scriptstyle X_1^2 + \cdots + X_n^2has a [chi-square distribution](http://en.wikipedia.org/wiki/Chi-square_distribution) with *n* degrees of freedom.
7. If \scriptstyle X_1,\dots,X_nare independent standard normal variables, then the [sample mean](http://en.wikipedia.org/wiki/Sample_mean) \scriptstyle\bar{X}=\frac{1}{n}(X_1+\cdots+X_n)and sample variance \scriptstyle S^2\,=\,\frac{1}{n-1}[(X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2]are [independent](http://en.wikipedia.org/wiki/Statistical_independence). This can be proven using [Basu's theorem](http://en.wikipedia.org/wiki/Basu%27s_theorem) or [Cochran's theorem](http://en.wikipedia.org/wiki/Cochran%27s_theorem).

**Mahalanobis distance**

It is based on [correlations](http://en.wikipedia.org/wiki/Correlation) between variables by which different patterns can be identified and analyzed. It is a useful way of determining *similarity* of an unknown [sample set](http://en.wikipedia.org/wiki/Sample_set) to a known one. It differs from [Euclidean distance](http://en.wikipedia.org/wiki/Euclidean_distance) in that it takes into account the correlations of the [data set](http://en.wikipedia.org/wiki/Data_set) and is [scale-invariant](http://en.wikipedia.org/wiki/Scale_invariance), i.e. not dependent on the scale of measurements.

Formally, the Mahalanobis distance of a multivariate vector x = ( x_1, x_2, x_3, \dots, x_N )^Tfrom a group of values with mean \mu = ( \mu_1, \mu_2, \mu_3, \dots , \mu_N )^Tand [covariance matrix](http://en.wikipedia.org/wiki/Covariance_matrix) *S* is defined as:



Mahalanobis distance (or "generalized squared interpoint distance" for its squared value[[3]](http://en.wikipedia.org/wiki/Mahalanobis_distance#cite_note-2)) can also be defined as a dissimilarity measure between two [random vectors](http://en.wikipedia.org/wiki/Random_vector)  \vec{x}and  \vec{y}of the same [distribution](http://en.wikipedia.org/wiki/Probability_distribution) with the [covariance matrix](http://en.wikipedia.org/wiki/Covariance_matrix) *S* :



If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the [Euclidean distance](http://en.wikipedia.org/wiki/Euclidean_distance). If the covariance matrix is diagonal, then the resulting distance measure is called the *normalized Euclidean distance*:



where σ*i* is the [standard deviation](http://en.wikipedia.org/wiki/Standard_deviation) of the *xi* over the sample set.

**Recognition methods by distance:**