Anàlisi d'Imatges i Reconeixement de Formes

Image Analysis and Pattern Recognition:

1. Statistical Approach

Francesc J. Ferri

Dept. d'Informàtica. Universitat de València

Gener 2007

Problem definition

Let X be a vector space and $W = \{w_1, w_2\}$ (two-class case).

A measure $x \in X$ corresponds to a stochastic process conditioned to the fact that x belongs either to class w_1 or class w_2.

A priori probability

the probability of any x belongs to a class, $P(w_i)$, also called *prior*.

Exemples:

OCR systems, vending machine, men-women, etc.

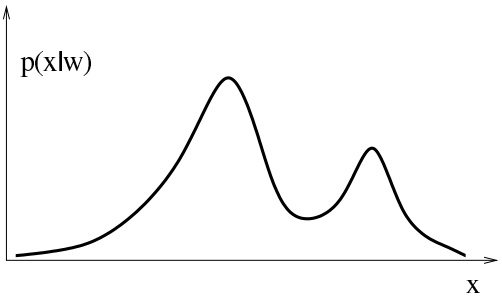
Priors can be fixed by the problem or estimated from frequencies in (representative) training sets.

Problem definition

Once we know (or are sure) that a measure x belongs to a particular class, w_i, how problable is to obtain a particular vector? Or how the variability of measures in this class is modelled?

class-conditional probability

is the *probability density function* (pdf) that describes the probability of obtaining a particular value x given that it belongs to class w_i. It is written as $p(x|w_i)$.

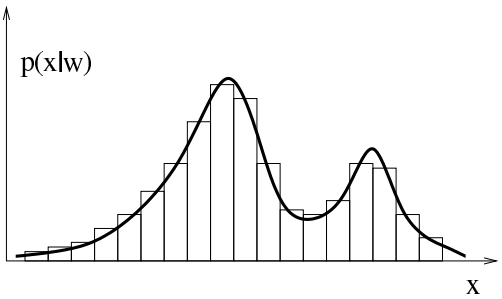


Problem definition

It holds that

$$\sum_i P(w_i) = 1 \quad \text{and} \quad \int_X p(x|w_i)dx = 1$$

If a representative sample (of points from w_i) is given, $p(x|w_i)$ can be approximated by a histogram.



Problem definition

The problem in pattern recognition in most cases reduces to the fact of assigning a class to a particular measurement, x.

In other words, we are interested in the probability of an object belongs to class w_i given that its measure is the vector x. .

A posteriori probability or likelihood

is the probability of a class given the measure, $p(w_i|x)$.

Problem definition

Unconditional probability

The probability of a particular vector x is obtained regardless of its class, $p(x)$. It is a pdf and its called sometimes *evidence factor*.

The unconditional probability is obviously given by

$$p(x) = \sum_i P(w_i)p(x|w_i)$$

Bayes theorem

All defined concepts are conveniently related by the Bayes theorem that states that

$$p(x, w_i) = p(x|w_i)P(w_i) = P(w_i|x)p(x)$$

that can be rewritten as

$$P(w_i|x) = \frac{p(x|w_i)P(w_i)}{p(x)}$$

which states that the **posterior probability** can be computed once **priors** and **conditional** probabilities (which should be fixed for a particular problem) are known.

Bayes rule

If all previous assumptions are met, it is possible to the define the **optimal** decision rule in the sense of minimal probability of error.

Minimum error Bayes classification rule

$$\mathcal{F}(x) = \begin{cases} w_1 & \text{if} \quad P(w_1|x) > P(w_2|x) \\ w_2 & \text{if} \quad P(w_2|x) > P(w_1|x) \\ \textit{undefined} & \text{otherwise} \end{cases}$$

The optimality of the Bayes rule

The probability of error for each x

$$P(\text{error}|x) = \begin{cases} P(w_2|x) \text{ if } P(w_1|x) > P(w_2|x) \\ P(w_1|x) \text{ if } P(w_2|x) > P(w_1|x) \end{cases}$$

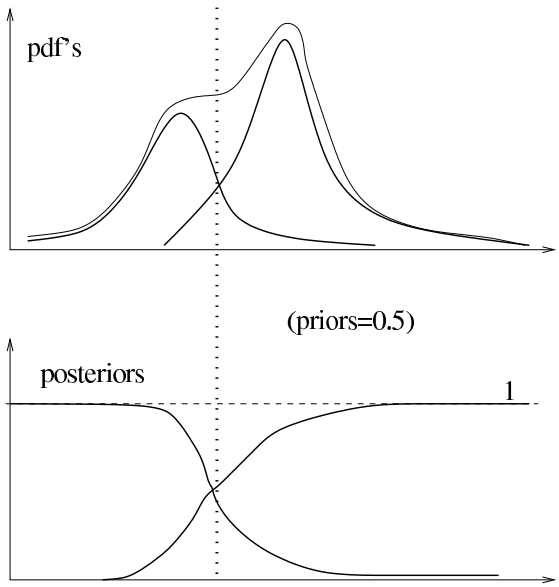
and averaging...

$$P(error) = \int_X P(\text{error}, x)dx = \int_X P(\text{error}|x)p(x)dx$$

and then (as $P(\text{error}|x) = \text{mín}[P(w_1|x), P(w_2|x)] \forall x$)

the value of the averaged $P(error)$ is obviously the minimum.

The Bayes rule (graphically)



The minimum-risk Bayes rule

The Bayes rule is optimal in the sense that it minimizes the number of expected errors.

But can be extended to the (interesting) case in which not all kind of errors are **equally important**. (vending machine, defect detection, etc.)

loss function/matrix

is a measure about how important is deciding class w_i when the *true* class is w_j. It is noted as $\lambda(w_i|w_j)$.

Risk

class-conditional risk

is the expected loss associated to each decision $w \in W$ given x. That is,

$$R(w|\mathbf{x}) = \sum_{j=1}^{c} \lambda(w|w_j) P(w_j|\mathbf{x})$$

Once a classification rule,$\alpha : R^d \longrightarrow W$, is given, one can define the

Risk (or loss) of the classifier

as

$$R = \int_{R^d} R(\alpha(\mathbf{x})|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

Minimum-risk Bayes rule

The optimal rule is then

$$\alpha(x) = \arg\min_{w \in \Omega} R(w|\mathbf{x})$$

as it minimizes the expected risk over all X.

The two-class case

The loss matrix is $\lambda_{ij} = (i \neq j)$, $i, j \in \{1, 2\}$.

and α decides w_1 or w_2 if $R(\alpha_1|x) < R(\alpha_2|x)$ or not. Then

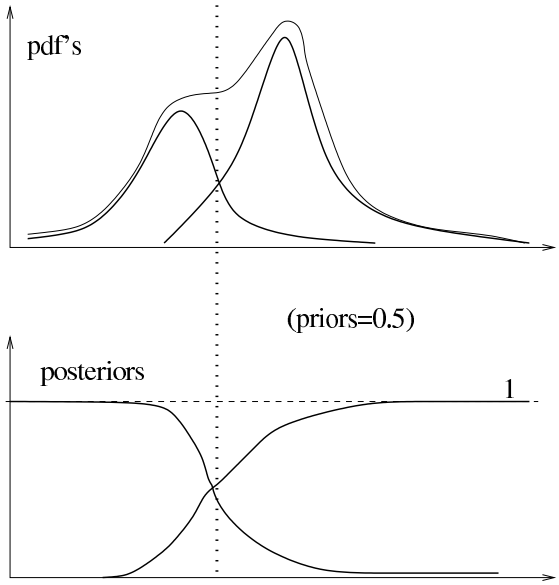
$$(\lambda_{21} - \lambda_{11})P(w_1|x) > (\lambda_{12} - \lambda_{22})P(w_2|x)$$

and applying the Bayes formula

$$\frac{p(x|w_1)}{p(x|w_2)} = \underbrace{\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(w_2)}{P(w_1)}}_{\text{indep. from } \mathbf{x}} = \theta$$

The minimum-risk Bayes rule

Observe that when $\lambda_{ij} = (i \neq j)$ we have the original rule.



Classifying with a reject option

In practice, sometimes is convenient **not** to assign a class to certain objects if there is not enough evidence (relative probability) about its true class.

Reject can be considered as a new class, w_0 but obviously cannot conveniently "modelled" as the other classes.

A convenient way of implementing the reject option is through reject thresholds

Bayes rule with reject

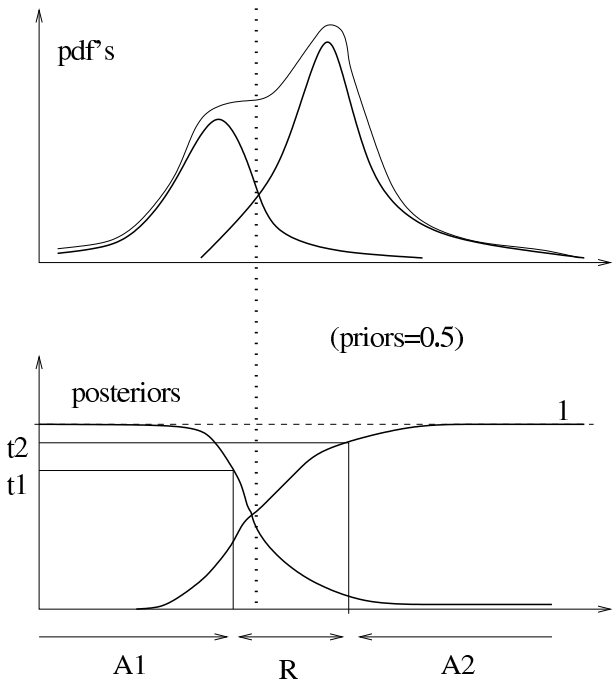
The classification rule is modified as

$$\mathcal{F}(x) = \begin{cases} w_i = \arg\max_{w \in \{w_1,\ldots,w_c\}} P(w|x) & \text{if } P(w_i|x) > t \\ w_0 & \text{otherwise} \end{cases}$$

The parameter t is the **reject threshold** and obviously $\frac{c-1}{c} \leq t \leq 1$.

This rule can be extended by using class-dependent reject thresholds, t_i.

Bayes rule with reject (graphically)



Minimum-risk Bayes rule with reject

The most general version of the Bayes rule can be defined by assigning loss values to the different reject options.

This can be accomplished by extending the loss matrix with a "zero" row

$$\lambda_{01}, \ldots, \lambda_{0c}$$

and using the same expressions.

Discriminant functions

A convenient and general way of describing and representing classifiers is using a function associated to each class, $g_i : R^d \longrightarrow R$

in such a way that the classifier is written as

$$\alpha(x) = \arg \max_{i=1,\ldots,c} g_i(x)$$

For example, $g_i = -R(\alpha_i|x)$ or $g_i(x) = P(w_i|x)$ for the Bayes classifiers.

As far as classification is concerned, any monotonically increasing function applied over these functions will not modify the resulting classifier

It is quite common to use logarithms of the above expressions for the Bayes classifiers in the normal case.

Discriminant functions

Acceptance regions

$$A_i = \{x \mid g_i(x) > g_j(x), j \neq i\}$$

Decision boundaries

$$g_i(x) = g_j(x)$$

When $c = 2$ usually only one discriminant function is considered

$$g(x) = g_1(x) - g_2(x)$$

The Bayes rule in the normal case

Univariate Gaussian pdf

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Multivariate Gaussian pdf

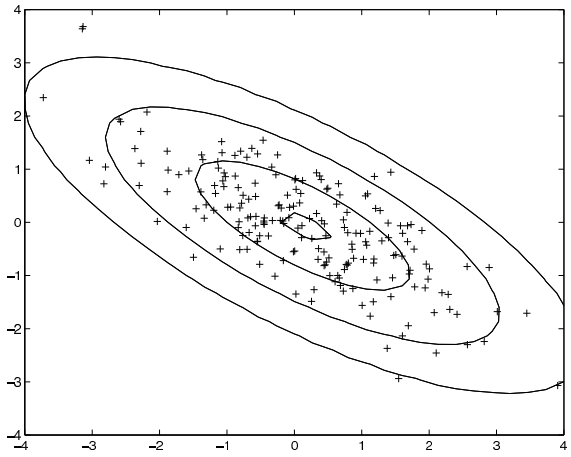
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma_i^{-1} (\mathbf{x}-\mu)}$$

Mahalanobis distance (between x and μ)

$$(\mathbf{x} - \mu)^T \Sigma_i^{-1} (\mathbf{x} - \mu)$$

The Mahalanobis distance

to measure distances between samples with regard to its mean in the normal case.



The Gaussian Bayes rule

Let us assume $c = 2$ and the minimum error case.

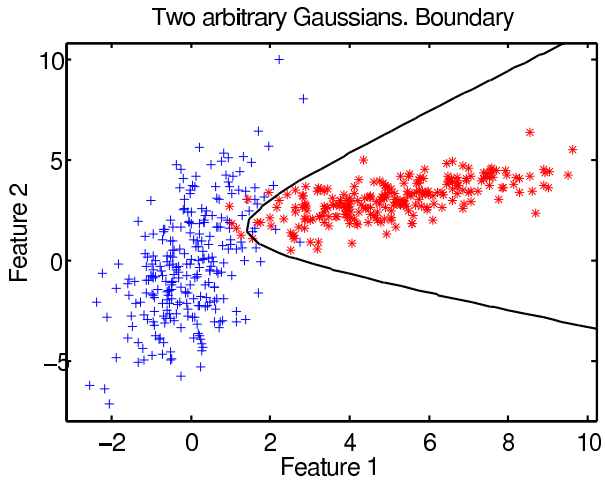
$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|w_i) + \ln P(w_i)$$

Then,

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\underbrace{-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)}_{\text{constant}}$$

This is the equation of a hyperquadrics

Bayes boundaries. General case



Particular case: hyperspheric classes

That is, all covariance matrices are as $\Sigma_i = \sigma^2 I$ Then, $|\Sigma_i| = \sigma^{2d}$ i $\Sigma_i^{-1} = \frac{1}{\sigma^2} I$

(at every step, terms independent of i are neglected)

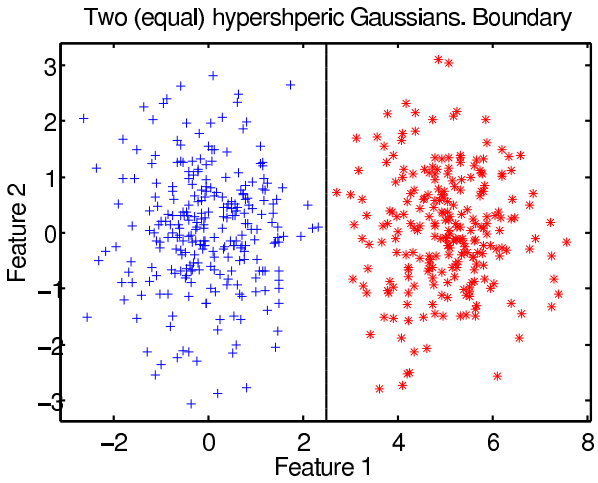
$$g_i(\mathbf{x}) = -\frac{||x - \mu_i||^2}{2\sigma^2} + \ln P(w_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(x^T x - 2\mu_i x + \mu_i^T \mu_i) + \ln P(w_i)$$

$$g_i(\mathbf{x}) = \frac{\mu_i}{\sigma^2} \cdot \mathbf{x} + \left[\frac{-\mu_i^T \mu_i}{2\sigma^2} + \ln P(w_i)\right]$$

This is a linear discriminant function.

Bayes boundaries. Hyperspherical case



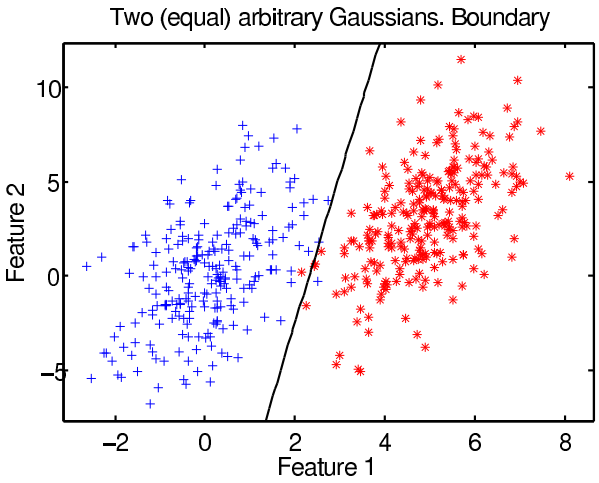
Particular case: equal covariance case

Each classe is supposed to corresond to a Gaussian with the same covariance matrix, Σ. Now,

$$g_i(\mathbf{x}) = \Sigma^{-1}\mu_i \cdot \mathbf{x} + \left[-\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln P(w_i)\right]$$

This is also a linear discriminant function.

Bayes boundaries. Equal covariance case



Learning Classifiers

Given the training data and once a particular (classification) model has been chosen, the problem consists of obatining the particular instance of the classifier that best adapts to the given data.

Parametric Pattern Recognition

In the particular case in which class-conditional densities can be characterized by a few parameters, the computation of the Bayes rule may be feasible. In this case the problem reduces to **parameter estimation**.

Parameter Estimation

Let's assume we have a dataset $D = \{x_1, \ldots, x_N\}$ which is drawn independently and identically distributed (i.i.d.) from a probability distribution $p(x|\theta)$ where θ is a parameter vector.

Likelihood

The function that measures how likely is to obtain precisely this dataset, given the parameters is the so-called **likelihood** of θ with regard to D.

$$p(D|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$$

One is interested in finding θ such that $p(D|\theta)$ is maximum

Maximum Likelihood Estimation

Once D is given and a functional form is assumed for $p(x|\theta)$, the parameters θ can be found by maximizing any monotonic function (usually the logarithm) of the likelihood

$$\theta = \arg\max \prod_{n=1}^{N} p(x_n|\theta) = \arg\max \sum_{n=1}^{N} \log p(x_n|\theta)$$

This is usually done by differencing and equating to zero

$$\frac{\partial}{\partial\theta} [\log p(D|\theta)] = 0$$

Information Theoretic Measures

It is possible to attach a measure on the amount of information carried by a random variable, x through its **differential entropy**

$$H[x] = -\int p(x) \log p(x) dx$$

measured in **bits** or **nats**

Entropy of a univariate Gaussian with μ and σ^2

$$H[x] = \frac{1}{2}(1 + \log(2\pi\sigma^2))$$

Relative Entropy

Given an (unknown) $p(x)$ and a (parametric) model $q(x|\theta)$

Relative entropy or Kullback-Leibler divergence

$$KL(p||q) = -\int p(x) \log\{\frac{q(x|\theta)}{p(x)}\}$$

When approximately evaluated using an i.i.d. sample $x_1, \ldots, x_N$

$$KL(p||q) \approx \sum_{n=1}^{N} -\log q(x_n|\theta) + \log p(x_n)$$

This implies that minimizing relative entropy is equivalent to maximize the log-likelihood

Maximum Likelihood. Normal Case

The corresponding log-likelihood is (1D case)

$$\ln(D|\mu,\sigma^2) = \frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

from which

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$

This generalizes as

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})(x_n-\mu_{ML})^T$$

Bayesian Estimation

Consider that the parameters θ themselves are random variables with corresponding pdf's, prior and posteriors.

What we are interested in is the probability of particular values of θ given the observed data, D. That is $p(\theta|D)$

According to Bayes theorem

$$p(\theta|D) \propto p(\theta)p(D|\theta)$$

$$\text{posterior} \propto \text{prior} \cdot \text{likelihood}$$

According to the likelihood, we can fix a convenient prior in order to be able to compute the posterior.

The posterior (given the observed samples) is interpreted as the current knowledge (and corresponding uncertainty) about the value of θ.