Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2009/2010

Lesson 16 9 april 2010

Bayesian Discriminant Functions

Sources Bibliographiques :

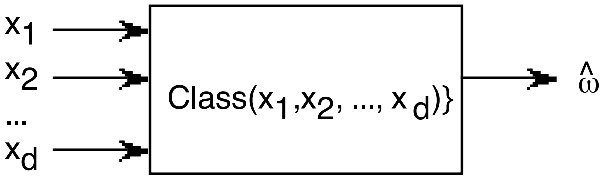
"Neural Networks for Pattern Recognition", C. M. Bishop, Oxford Univ. Press, 1995.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notation

x	a variable
X	a random variable (unpredictable value)
N	The number of possible values for X (Can be infinite).
$\vec{x}$	A vector of D variables.
$\vec{X}$	A vector of D random variables.
D	The number of dimensions for the vector $\vec{x}$ or $\vec{X}$
E	An observation. An event.
T_k	The class (tribe) k
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that $E \in T_k$
$p(\omega_k) = p(E \in T_k)$	Probability that the observation E is a member of the class k. Note that $p(\omega_k)$ is lower case.
M_k	Number of examples for the class k. (think M = Mass)
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{ X_m^k \}$	A set of M_k examples for the class k. $\{X_m\} = \bigcup_{k=1,K} \{X_m^k\}$
P(X)	Probability density function for X
$P(\vec{X})$	Probability density function for $\vec{X}$
$P(\vec{X} / \omega_k)$	Probability density for $\vec{X}$ the class k. $\omega_k = E \in T_k$.
h(n)	A histogram of random values for the feature n.
$h_k(n)$	A histogram of random values for the feature n for the class k. $h(x) = \sum_{k=1}^{K} h_k(x)$
Q	Number of cells in $h(n)$. $Q = N^D$

Bayesian Classification (Reminder)

Our problem is to build a box that maps a set of features $\vec{X}$ from an Observation, E into a class T_k from a set of K possible Classes.



Let ω_k be the proposition that the event belongs to class k: $\omega_k = E \in T_k$

ω_k Proposition that event E $\in$ the class k

In order to minimize the number of mistakes, we will maximize the probability that $\omega_k \equiv E \in T_k$

$$\hat{\omega}_k = \arg-\max_k \left\{ \Pr(\omega_k \mid \vec{X}) \right\}$$

We will call on two tools for this:

1) Baye's Rule :

$$p(\omega_k \mid \vec{X}) = \frac{P(\vec{X} \mid \omega_k) p(\omega_k)}{P(\vec{X})}$$

2) Normal Density Functions

$$P(\vec{X} \mid \omega_k) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(C_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1} (\vec{X}-\vec{\mu}_k)}$$

Last week we looked at Baye's rule. Today we concentrate on Normal Density Functions.

Linear Transforms of the Normal Multivariate Density

The Normal (Gaussian) function is a defined only by its moments.
It is thus invariant to transformations of its moments, that is affine transformations.
The affine transformations include rotation, translation, scale changes and other linear transformations.
For example consider a rotation vector of cosine angles about each component of X :
Rotation is projection onto a vector $\vec{R}$

$$\vec{R} = \begin{pmatrix} \cos(\alpha_1) \\ \cos(\alpha_2) \\ \dots \\ \cos(\alpha_D) \end{pmatrix}$$

Note that $\| \vec{R} \| = 1$.
A vector $\vec{X}$ may be rotated by projection onto $\vec{R}$

$$\vec{Y} = \vec{R}^T \vec{X}$$

Projection of a Gaussian is the Gaussian of the projection.
For a projection R : $\vec{Y} = \vec{R}^T \vec{X}$

$$\mu_y = \vec{R}^T \vec{\mu}_x, \qquad \sigma_y^2 = \vec{R}^T \mathbf{C}_X \vec{R}$$

For the Covariance:
Projection of a covariance requires pre- and post- multiplication by $\vec{R}$.

For $C_x = E\{\vec{V}\vec{V}^T\}$ where $\vec{V}_m = \vec{X}_m - E\{\vec{X}_m\} = \vec{X}_m - \vec{\mu}_m$

$$C_Y = E\{(\vec{R}^T\vec{V})(\vec{R}^T\vec{V})^T\} = E\{(\vec{R}^T\vec{V})(\vec{V}^T\vec{R})\} = E\{(\vec{R}^T(\vec{V}\vec{V}^T)\vec{R})\} = E\{(\vec{R}^T C_X \vec{R})\}$$

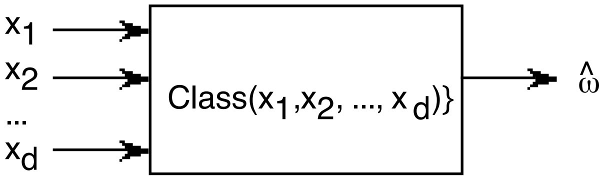
$$P(Y) = \mathcal{N}(y;\ \vec{R}^T \vec{\mu}_x,\ \vec{R}^T \mathbf{C}_X \vec{R}^T) = \mathcal{N}(y;\ \mu_y, \sigma_y^2)$$

This can be computed by projecting the moments or by projecting the data and recomputing the moments.

$$\mu_y = E\{P(Y)\} = \vec{R}^T \vec{\mu}_x \qquad \sigma_y^2 = E\{(P(Y)-\mu_y)(P(Y)-\mu_y)\} = \vec{R}^T \mathbf{C}_X \vec{R}$$

Quadratic Discrimination

Classification is a process of estimating the membership of an observation in a class based on the features of the observation, $\vec{X}$.



$$\hat{\omega}_k = \text{Class}(E) = \text{Decide}(E \in A_k)$$

$\hat{\omega}_k$ is the proposition that $(E \in \omega_k)$.

The classification fonction can be decomposed into two parts: d() and $g_k()$:

$$\hat{\omega}_k = d(g(\vec{X})).$$

$g(\vec{X})$: A discriminant function : $R^D \rightarrow R^K$
d() : a decision function $R^K \rightarrow \{\omega_K\}$

Discrimination

$g(\vec{X})$: Is a discrimination function that maps from $R^D \rightarrow R^K$

$$\vec{g}(\vec{X}) = \begin{pmatrix} g_1(\vec{X}) \\ g_2(\vec{X}) \\ \ldots \\ g_D(\vec{X}) \end{pmatrix}$$

Quadratic discrimination functions can be derived directly from $p(\omega_k \mid X)$

$$p(\omega_k \mid \vec{X}) = \frac{P(\vec{X} \mid \omega_k) p(\omega_k)}{P(\vec{X})}$$

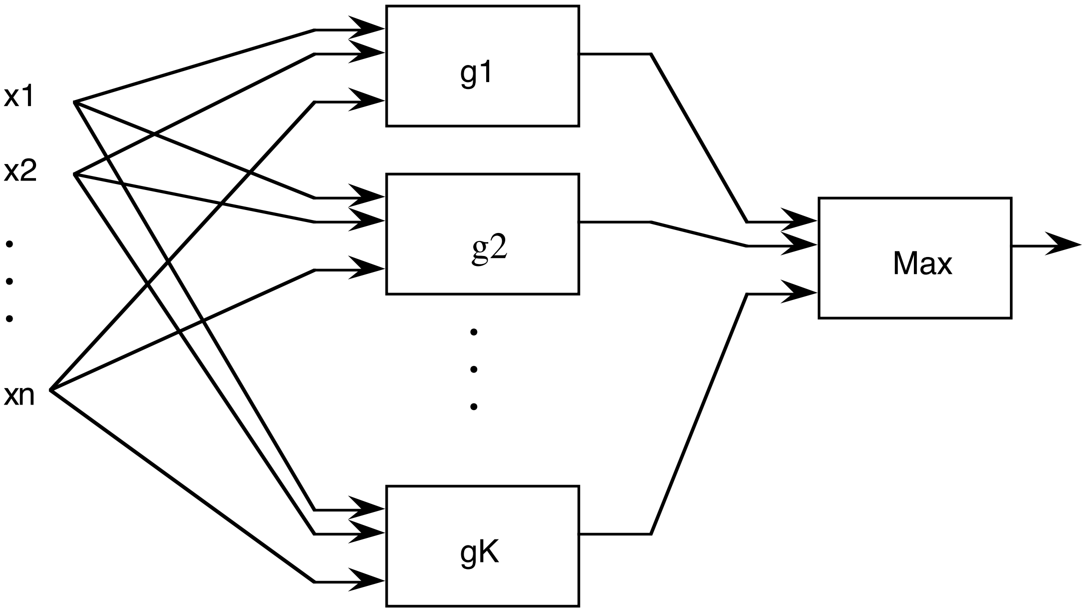
To minimize the number of errors, we will choose k such that

$$k = \underset{k}{\text{arg-max}}\ \{g_k(X)\} = \underset{k}{\text{arg-max}}\ \{p(\omega_k \mid \vec{X})\} = \underset{k}{\text{arg-max}}\{\frac{P(\vec{X} \mid \omega_k)p(\omega_k)}{P(\vec{X})}\}$$

but because P(X) is constant for all k, it is common to use:

$$k = \underset{k}{\text{arg-max}}\{P(\vec{X} \mid \omega_k)p(\omega_k)\}$$

Thus the classifier is decomposed to a selection among a set of parallel discriminant functions.



This is easily applied to the multivariate norm:

$$P(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \mathbf{C}_k)$$

or with a sum of normals (Gaussian Mixture Model).

$$P(\vec{X} \mid \omega_k) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(\vec{X}; \vec{\mu}_{kn}, \mathbf{C}_{kn})$$

Discrimination using Log Likelihood

Let D=1, with

$$p(X=x \mid \omega_k) = \mathcal{N}(x; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

The discrimination function takes the form:

$$g_k(X) = p(\omega_k) \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

Note that $k = \underset{k}{\text{arg-max}}\{g_k(X)\} = \underset{\omega_k}{\text{arg-max}}\{\text{Log}\{g_k(X)\}\}$

because Log{} is a monotonic function.

$$k = \underset{k}{\text{arg-max}} \{\text{Log}\{ \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \} + \text{Log}\{p(\omega_k)\} \}$$

$$k = \underset{k}{\text{arg-max}} \{\text{Log}\{ \frac{1}{\sqrt{2\pi}\sigma_k} \} + \text{Log}\{e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \} + \text{Log}\{p(\omega_k)\} \}$$

$$k = \underset{k}{\text{arg-max}} \{-\text{Log}\{\sqrt{2\pi} \ \sigma_k\} - \frac{(x-\mu_k)^2}{2\sigma_k^2} + \text{Log}\{p(\omega_k)\} \}$$

$$k = \underset{k}{\text{arg-max}} \{-\text{Log}\{\sigma_k\} - \frac{(x-\mu_k)^2}{2\sigma_k^2} + \text{Log}\{p(\omega_k)\} \}$$

Example for K > 2 and D > 1

In the general case, there are D characteristics.

$$\boxed{g_k(\vec{X}) = p(\omega_k \mid \vec{X})\, p(\omega_k)}$$

The decision rule is

$$\boxed{\hat{\omega}_i \ : \ \text{si } \forall\ j \neq i \quad g_i(\vec{X}) > g_j(\vec{X})}$$

Thus the classifier is a machine that calculates K functions $g_k(\vec{X})$
Followed by a maximum selection.

The discrimination function is $g_k(\vec{X}) = p(\vec{X} \mid \omega_k)\, p(\omega_k)$

On sélection la classe ω_k pour laquelle $\underset{k}{\text{arg-max}} \{g_k(\vec{X})\}$

par règle de Bayes :

$$\underset{k}{\text{arg-max}} \{p(\omega_k \mid \vec{X})\} = k = \underset{k}{\text{arg-max}} \{ p(\vec{X} \mid \omega_k)\, p(\omega_k) \}$$

$$= \underset{k}{\text{arg-max}} \{\text{Log}\{p(\vec{X} \mid \omega_k)\} + \text{Log}\{p(\omega_k)\}$$

Si les caractéristiques suivent une densité Normale :

$$p(\vec{X} \mid w_k) = \mathcal{N}(\vec{X}, \vec{\mu}_k, C_k)$$

$$Log(P(\vec{X} \mid \omega_k)\} = Log\{\frac{1}{(2\pi)^{\frac{D}{2}} \det(C_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1} (\vec{X}-\vec{\mu}_k)}\}$$

$$Log(P(\vec{X} \mid \omega_k)\} = -\frac{D}{2} Log(2\pi) - \frac{1}{2} Log\{Det(C_x)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1} (\vec{X}-\vec{\mu}_k)$$

We observe that $-\frac{D}{2}$ Log{2π} can be ignored because it is constant for all k.
The discrimination function becomes:

$$g_k(\vec{X}) = -\frac{1}{2}Log\{\det(C_k)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1}(\vec{X}-\vec{\mu}_k) + Log\{p(\omega_k)\}$$

$$\boxed{g_k(\vec{X}) = -\frac{1}{2}Log\{\det(C_k)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1}(\vec{X}-\vec{\mu}_k) + Log\{p(\omega_k)\}}$$

Different families of Bayesian classifiers can be defined by variations of this formula. This becomes more evident if we reduce the equation to a quadratic polynomial.

Canonical Form for the discrimination function

The quadratic discriminant can be reduced to a standard (canonical) form.

$$g_k(\vec{X}) = -\frac{1}{2}Log\{\det(C_k)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1}(\vec{X}-\vec{\mu}_k) + Log\{p(\omega_k)\}$$

Let us start with the term $(\vec{X}-\vec{\mu}_k)^T C_k^{-1}(\vec{X}-\vec{\mu}_k)$.

This can be rewritten as :

$$(\vec{X}-\vec{\mu}_k)^T C_k^{-1}(\vec{X}-\vec{\mu}_k) = \vec{X}^T C_k^{-1}\vec{X} - \vec{X}^T C_k^{-1}\vec{\mu}_k - \vec{\mu}_k{}^T C_k^{-1}\vec{X} + \vec{\mu}_k{}^T C_k^{-1}\vec{\mu}_k$$

We note that $\vec{X}^T C_k^{-1}\vec{\mu}_k = \vec{\mu}_k{}^T C_k^{-1}\vec{X}$
and thus : $-\vec{X}^T C_k^{-1}\vec{\mu}_k - \vec{\mu}_k{}^T C_k^{-1}\vec{X} = -(2C_k^{-1}\vec{\mu}_k)^T\vec{X}$

we define: $\vec{W}_k = -2C_k^{-1}\vec{\mu}_k$
to obtain $-\vec{X}^T C_k^{-1}\vec{\mu}_k - \vec{\mu}_k{}^T C_k^{-1}\vec{X} = \vec{W}_k{}^T\vec{X}$

Let us also define $D_k = -\frac{1}{2}C_k^{-1}$

The remaining terms are constant. Let us defined the constant

$$b_k = -\frac{1}{2}\vec{\mu}_k{}^T C_k^{-1}\vec{\mu}_k - Log\{\det(C_k)\} + Log\{p(\omega_k)\}$$

which gives a quadratic polynomial

$$\boxed{g_k(\vec{X}) = \vec{X}^T D_k\vec{X} + \vec{W}_k^T\vec{X} + b_k}$$

where: $D_k = -\frac{1}{2}C_k^{-1}$

$\vec{W}_k = -2C_k^{-1}\vec{\mu}_k$

and $b_k = -\frac{1}{2}\vec{\mu}_k{}^T C_k^{-1}\vec{\mu}_k - Log\{\det(C_k)\} + Log\{p(\omega_k)\}$

A set of K discrimination functions $g_k(\vec{X})$ partitions the space $\vec{X}$ into a disjoint set of regions with qudratic boundaries. The boundaries are points for which

$$g_i(\vec{X}) = g_j(\vec{X}) \geq \ g_k(\vec{X}) \ \forall k \neq i, j$$

The boundaries are the functions $g_i(\vec{X}) - g_j(\vec{X}) = 0$

Noise and Discrimination

Under certain conditions, the quadratic discrimination function can be simplified by eliminating either the quadratic or the linear term.

If we could perfectly model the universe, then sensor reading would be a predictable value, $\vec{x}$. The normal density attempts to represent this with the "average" feature $\vec{\mu}_k$

In reality, the features of a class are generally dispersed by un-modeled phenomena. These may be effects that are beyond the abilities of the available sensors, or they may be effects that we choose to ignore because they are "unimportant".

Although the true variation my not be additive, we will model it as an additive random term N_k. The term is random because we are unable to predict it.

Thus the observed feature is random: $\vec{X} = \vec{x} + N_k$

For example, the color of your eyes could be predicted from your genetic code, but in the absence of a genetic decoder, this becomes random.

In addition, every observation system (or sensor) is subject to some form of sensor noise. This sensor Noise is modeled as an additive random term N_s. Sensor noise is generally independent of the class k.

Thus the sensor returns a random feature $\vec{X} = \vec{x} + \vec{N}_k + \vec{N}_s$

The Normal density function represents these two forms of "noise" as a second moment of the class, C_k.

Thus $C_k = E\{(E\{(N_k + N_s)(N_k + N_s)^T\}$

Depending on the nature of $\vec{N}_k \text{ and } \vec{N}_s$ different simplifications are possible.

For example if $\vec{N}_s >> \vec{N}_k$ then the term C_k will be nearly constant for all k.
In this case, the discrimination function can be reduced to a linear equation.

$$g_k(\vec{X}) = \vec{W}_k^T \vec{X} + b_k$$

This is very useful because there are simple powerful techniques to calculate the terms of such an equation.

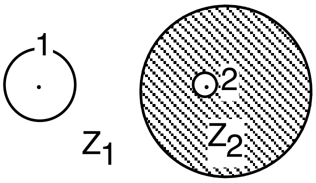
Decision Surfaces for different Noise assumptions

In the more general case we can not make any assumptions on $\vec{N}_k$ and $\vec{N}_s$
Depending on the nature $\vec{N}_k$ we may find a variety of different second order decision surfaces :

For eaxample (K=2, D=2)

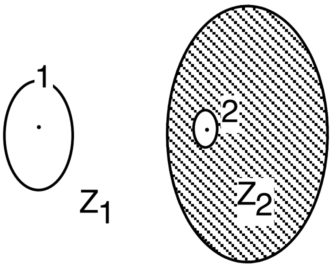
Hyper-sphere :

Let $C_k = \sigma_k^2 I$
and $\det\{C_1\} > \det\{C_2\}$



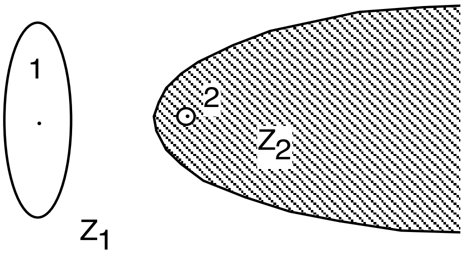
Hyper-ellipsoid :

For $\sigma_{x1}^2 > \sigma_{x2}^2$
and $\det\{C_1\} > \det\{C_2\}$

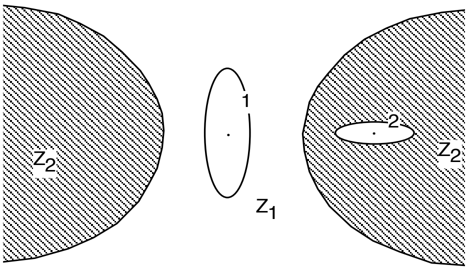


Hyper-paraboloid :

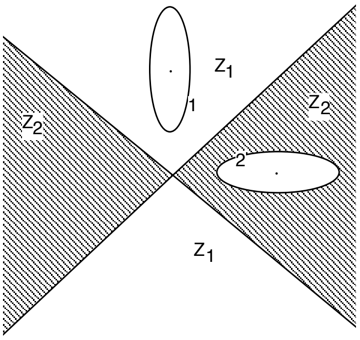
for $\sigma^2_{x1k=1} >> \sigma^2_{x1k=2}$
et $\sigma^2_{x2k=1} > \sigma^2_{x2k=2}$



Hyper-hyperboloids :

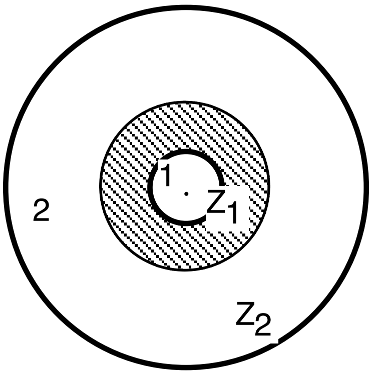


Hyperplanes.



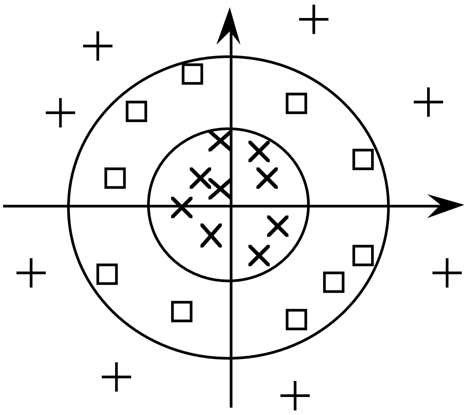
$\vec{\mu}_1 = \vec{\mu}_2$ et $C_1 << C_2$

with $\sigma_{11} = \sigma_{22}$ et $\sigma_{12} = \sigma_{21} = 0$.



a hypershere.

Two classes with equal means



Suppose tht we have 2 classes i, j such that

$$\vec{\mu}_i = \vec{\mu}_j \text{ and } \det(C_i) > \det(C_j).$$

Is it possible to assign an observation to one of the classes?

$$g_i(\vec{X}) - g_j(\vec{X}) = 0$$

takes the form of a sphere with observations assigned to T_i outside the sphere and T_j on the inside.

$$g_k(\vec{X}) = \vec{X}^T D_k \vec{X} + \vec{W}_k^T \vec{X} + b_k$$