Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2013/2014

Lesson 17 18 april 2014

Discriminant Functions

Notation 2

Bayesian Classification...3

Quadratic Discrimination...4

Discrimination using Log Likelihood 6

Example for K > 2 and D > 1... 7

Canonical Form for the discrimination function 8

Noise and Discrimination 9

Decision Surfaces for different Noise assumptions 11

Two classes with equal means 12

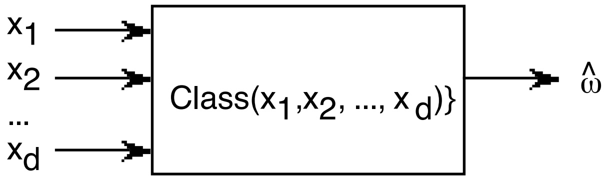
Source:
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notation

$\vec{x}$	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.
C_k	The class k.
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that E $\in C_k$
$P(\omega_k) = P(E \in C_k)$	Probability that the observation E is a member of the class k. Note that $p(\omega_k)$ is lower case.
$p(X)$	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. ω_k = E $\in T_k$.

Bayesian Classification

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



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

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

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

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

We will call on two tools for this:

1) Baye's Rule :

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

2) Normal Density Functions

$$p(\vec{X}) = \mathcal{N}(\vec{X}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1} (\vec{X}-\vec{\mu})}$$

and

$$p(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1} (\vec{X}-\vec{\mu}_k)}$$

Quadratic Discrimination

The classification function can be decomposed into two parts: d() and g_K():

$$\hat{\omega}_k = d\left(g_k\left(\vec{X}\right)\right)$$

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

The discriminant is a vector of functions:

$$\vec{g}(\vec{X}) = \begin{pmatrix} g_1(\vec{X}) \\ g_2(\vec{X}) \\ \vdots \\ g_K(\vec{X}) \end{pmatrix}$$

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

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

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

$$\hat{\omega}_k = \underset{\omega_k}{\arg-\max}\{\frac{P(\vec{X} \mid \omega_k)p(\omega_k)}{P(\vec{X})}\}$$

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

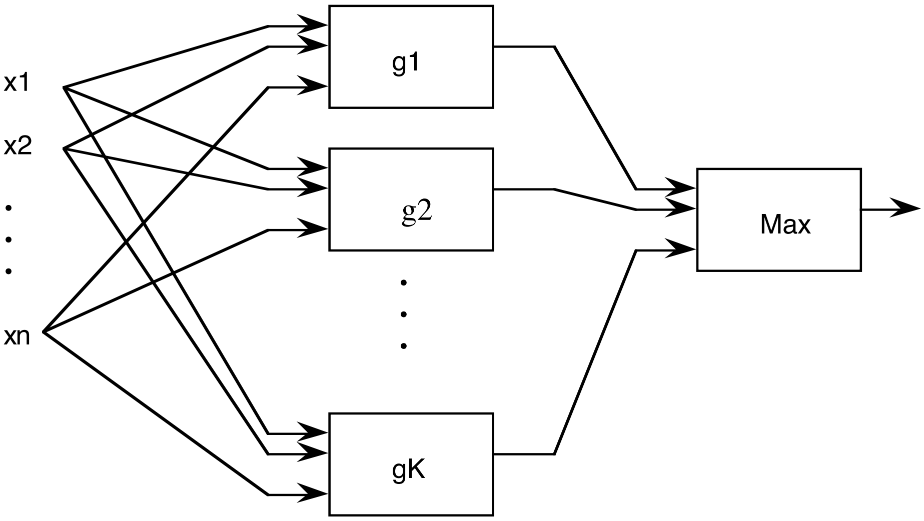
$$\hat{\omega}_k = \underset{\omega_k}{\arg-\max}\{P(\vec{X} \mid \omega_k)p(\omega_k)\}$$

This is called a "Maximum Likelihood" classifier.

Warning: Maximum likelihood can sometimes give a highly improbable answer. Remember that confidence in the choice $\hat{\omega}_k$ is provided by the full probability:

$$CF_{\hat{\omega}_k} = p(\hat{\omega}_k \mid \vec{X}) = \frac{P(\vec{X} \mid \hat{\omega}_k)p(\hat{\omega}_k)}{P(\vec{X})}$$

The maximum likelihood classifier can be organized as a set of parallel discriminant functions. This gives a sort of parallel machine that resembles:



The functions $g_k()$ are commonly constructed by from the Log of the likelihood:

$$g_k(X) = Log\{P(\vec{X} \mid \omega_k)P(\omega_k)\}$$

This gives a simple function in the case where the pdf is a multivariate norm:

$$p(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k)$$

or even to a Gaussian Mixture Model

$$p(\vec{X} \mid \omega_k) = \sum_{n=1}^{M} \alpha_n \mathcal{N}(\vec{X}; \vec{\mu}_n, \Sigma_n)$$

Discrimination using Log Likelihood

As a simple example, let D=1

$$p(X) = \mathcal{N}(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

The Likelihood function takes the form:

$$L_k(X) = p(X \mid \omega_k) \cdot P(\omega_k) = P(\omega_k) \cdot \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(X-\mu_k)^2}{2\sigma_k^2}}$$

We can simplify the math by working with the Logarithm of the likelihood.
We note that:

$$\hat{k} = \arg\text{-}\max_k\left\{L_k(X)\right\} = \arg\text{-}\max_k\left\{Log\{L_k(X)\}\right\}$$

because Log{} is a monotonic function. In this case we define g_k() to be the Log-likelihood.

$$g_k(X) = Log\{\mathcal{N}(X;\mu_k,\sigma_k) \cdot P(\omega_k)\}$$

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

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

$$g_k(X) = -Log\{\sqrt{2\pi}\} - Log\{\sigma_k\} - \frac{(X-\mu_k)^2}{2\sigma_k^2} + Log\{P(\omega_k)\}$$

$$g_k(X) = -Log\{\sigma_k\} - \frac{(X-\mu_k)^2}{2\sigma_k^2} + Log\{P(\omega_k)\}$$

and

$$\hat{k} = \arg\text{-}\max_k\left\{g_k(X)\right\}$$

Example for K > 2 and D > 1

In general, it is more effective to work with D>1 features.
In this case:

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

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

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

For a Gaussian (Normal) density function

$$p(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X}; \vec{\mu}_k, \Sigma_k)$$

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

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

We can observe that $-\frac{D}{2}Log(2\pi)$ can be ignored because it is constant for all k.

The discrimination function becomes:

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

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.

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

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

This can be rewritten as :

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

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

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

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

The remaining terms are constant. Let us defined the constant

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

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

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

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

A set of K discrimination functions $g_k(\vec{X})$ partitions the space $\vec{X}$ into a disjoint set of regions with quadratic boundaries. The boundaries are the functions $g_i(\vec{X}) - g_j(\vec{X}) = 0$

The boundaries are defined by points for which

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

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

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

For example if $\vec{N}_s$ >> $\vec{N}_k$ then the term Σ_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 $\vec{N}_s$

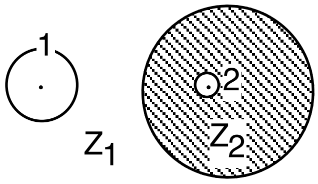
Depending on the nature $\vec{N}_k$ we may find a variety of different second order decision surfaces :

For eaxample (K=2, D=2)

Hyper-sphere :

Let $\Sigma_k = \sigma_k^2 I$

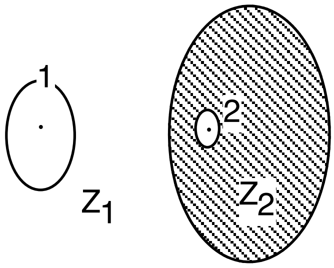
and $\det\{\Sigma_1\} > \det\{\Sigma_2\}$



Hyper-ellipsoid :

For $\sigma_{x1}^2 > \sigma_{x2}^2$

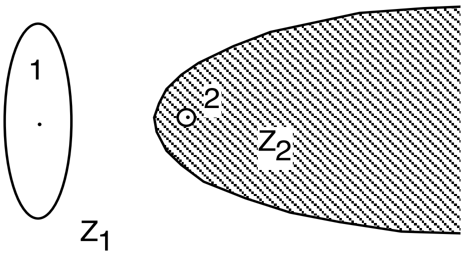
and $\det\{\Sigma_1\} > \det\{\Sigma_2\}$



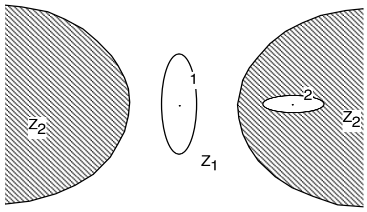
Hyper-paraboloid :

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

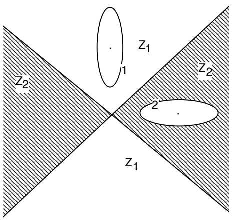
et $\sigma^2_{x2k=1} > \sigma^2_{x2k=2}$



Hyper-hyperboloids :



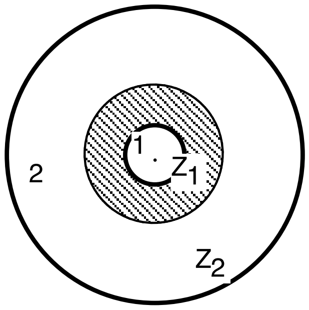
Hyperplanes.



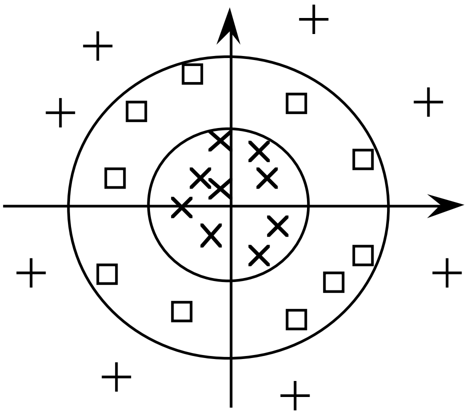
$\vec{\mu}_1 = \vec{\mu}_2$ and $\det\{\Sigma_1\} > \det\{\Sigma_2\}$

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

$\vec{\mu}_i = \vec{\mu}_j$ and det{Σ_1} > det{Σ_2}.

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 C_i outside the sphere and C_j on the inside.

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