Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2011/2012

Lesson 19 25 April 2012

Linear Classification Methods

Contents

Notation 2

Bayesian Discrimination Functions (Rappel) 3

Linear Classification 4

Pattern detectors as linear classifiers...4

ROC Curve 6

Vector between center of gravities. 7

Fisher Linear Discriminant. 8

Two Class solution 8

Fisher's Discriminant for Multiple Classes....................................11

Sources Bibliographiques :
"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notation

x	a variable
X	a random variable (unpredictable value)
$\vec{x}$	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 (tribe) k
k	Class index
K	Total number of classes
ω_k	The fact that E $\in C_k$
$\hat{\omega}_k$	The decision (estimation) 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.
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. ω_k = E $\in T_k$.

Bayesian Discrimination Functions (Rappel)

In lesson 17 we saw that the classification function in a Bayesian Classifier can be decomposed into two parts: a decision function – d() and a discrimination function – g_k():

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

Quadratic discrimination functions can be derived directly from maximizing the probability of $p(\omega_K \mid X)$

$$\vec{g}(\vec{X}) = \begin{pmatrix} g_1(\vec{X}) \\ g_2(\vec{X}) \\ \ldots \\ g_K(\vec{X}) \end{pmatrix}$$ A set of discriminant functions : $R^D \rightarrow R^K$

d() : a decision function $R^K \rightarrow \{\omega_K\}$

We derived the canonical form for the discriminant function.

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

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

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

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. At the boundaries between classes:

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

Linear Classification

In lesson 17 we saw that in many cases the quadratic term can be ignored and the partitions take on the form of hyper-surfaces. 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 coefficients for linear functions from training data.

Pattern detectors as linear classifiers.

Linear classifiers are widely used to define pattern "detection" systems. Such systems can be provide discrimination between a target class and everything else. (K=2) This is widely used in computer vision, for example, to detect faces or publicity logos, or other patterns of interest.

Class k=1: The target pattern.
Class k=2: Everything else.

In the following examples, we will assume that our training data is composed of M sample observations $\{\vec{X}_m\}$ where each sample is labeled with an indicator y_m

y_m = +1 for examples of the target pattern (class 1)
y_m = –1 for all other examples.

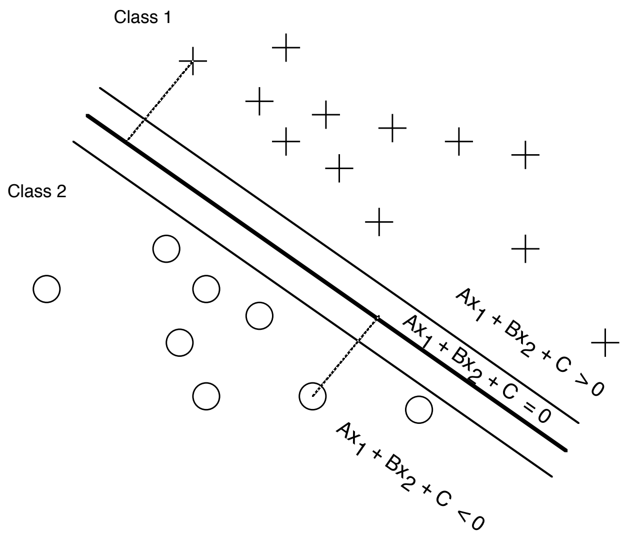
Our goal is to build a hyper-plane that provides a best separation of class 1 from class 2.

$$\vec{W}^T \vec{X} + B$$

B is an adjustable gain that sets the sensitivity of the detector.

In this case, the decision rule can be expressed using the Sgn function:

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{if x } > 0 \\ 0 & \text{if } \text{x} \leq 0 \end{cases}$$



A hyperplane is a set of points such that $\vec{W}^T \vec{X} + B = 0$

$$w_1x_1 + w_2x_2 + ... + w_Dx_D + b = 0$$

Where $\vec{W} = \begin{pmatrix} w_1 \\ w_2 \\ ... \\ w_D \end{pmatrix}$ is the normal to the hyperplane.

When $\vec{W}$ is normalized to unit length, $\|\vec{W}\| = 1$, then

$B = -\vec{W}^T \vec{X}$ is the perpendicular distance to the origin.

if $\|\vec{W}\| \neq 1$ then normalize as $\vec{W}' = \frac{\vec{W}}{\|\vec{W}\|}$ and $B' = \frac{B}{\|\vec{W}\|}$

A variety of techniques exist to calculate the plane. The best choice can depend on the nature of the pattern class as well as the nature of the non-class data.

For example:

1) Vector between center of gravities.
2) Fisher linear discriminant analysis,
3) Regression
4) Perceptrons

ROC Curve

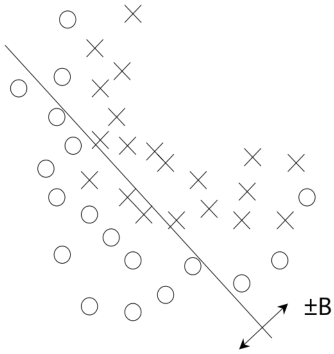
Two-class linear discrimination functions are practical for many problems. Among other uses, they provide the optimal solution to many signal detection problems in communications theories. In the case of radio communications, the noise is typically additive, Gaussian and independent of the signal, and the Bayesian Classifier reduces to a linear classifier.

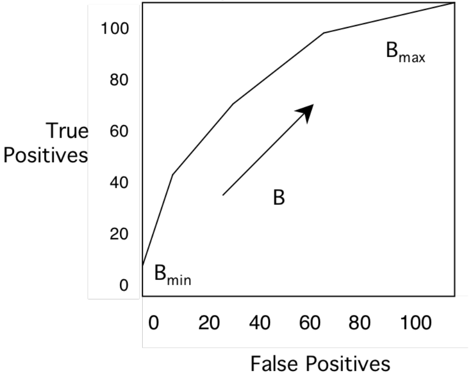
Historically two class linear classifiers have been used to demonstrate optimality for some signal detection methods. The quality metric that is used is the Receiver Operating Characteristic curve. This curve should be used to describe or compare any method for signal or pattern detection.

We can bias the classifier to one or the other class by adding and Bias term B.

$$y(\vec{X}) = \vec{W}^T \cdot \vec{X} + B$$

B is a free variable that can be swept through a range of values.
Changing B changes the ratio of true positive detection to false detections.
This is illustrated by a curve called the Receiver Operating Characteristics (ROC) curve. The ROC plots False Positives against False Negatives as a function of B.





The more the curve approaches the upper left corner the better the detector.
The ROC is a powerful descriptor for the “goodness” of a linear classifier.
For a target class C_1 a Positive (P) detection is the decision that $E \in C_1$
a Negative (N) detection is the decision that $E \in C_2$

		True Class	
Estimated Decision		ω_1	ω_2
	$\hat{\omega}_1$	True Positive (TP)	False Positive (FP)
	$\hat{\omega}_2$	False Negative (FN)	True Negative (TN)

Vector between center of gravities.

Suppose that we have two classes, defined with training data sets $\{X_m^1\}$ and $\{X_m^2\}$, with mean and covariance $(\vec{\mu}_1, \Sigma_1)$, and $(\vec{\mu}_2, \Sigma_2)$. These can be used to define two linear discriminant functions:

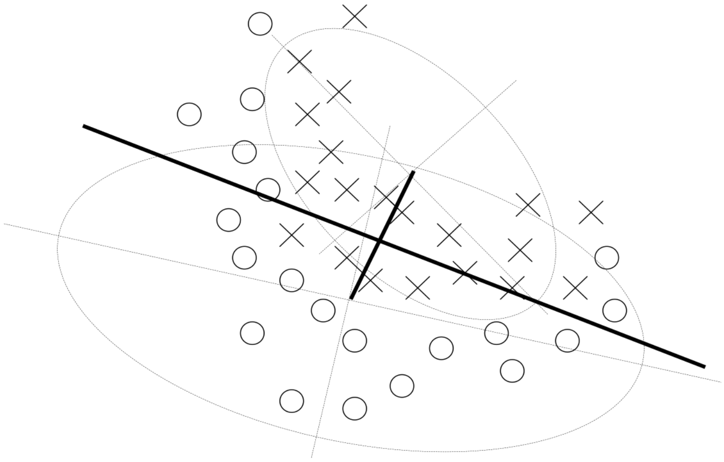
Let $g_1(\vec{X}) = \vec{W}_1^T \vec{X} + b_1$ and $g_2(\vec{X}) = \vec{W}_2^T \vec{X} + b_2$

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

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

The decision boundary is

$$g_1(\vec{X}) - g_2(\vec{X}) = 0$$
$$(\vec{W}_1^T - \vec{W}_2^T)\vec{X} + b_1 - b_2 = 0$$
$$(C_1^{-1}\vec{\mu}_1 - C_2^{-1}\vec{\mu}_2) + b_1 - b_2 = 0$$



The direction is determined by the vector between the center of gravities of the two classes, weighted by the inverse of the covariance matrices.

This approach is based on the assumption that the two classes are well modeled by Normal density functions. This assumption is not reasonable in many cases.
If one of the classes is not well modeled as a normal, the results can be unreliable.

Fisher Linear Discriminant.

The Discrimination problem can be viewed as a problem of projecting the D dimensional feature space onto a lower dimensional K dimensional space.

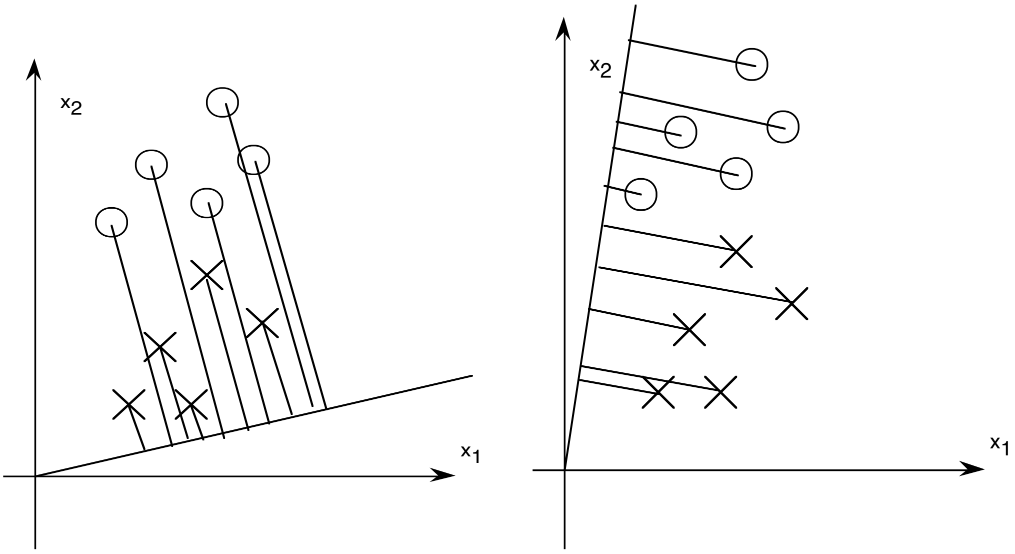
The tool for such projection is the Fisher discriminant.

Two Class solution

The principle of the Fisher linear discriminant is to project the vector X with D_x onto a variable z (D=1) by a linear projection F such that the classes are most separated.

$$z = \vec{F}^T \cdot \vec{X}$$

A Fisher metric, J(F) is used to choose F such that the two classes are most separated.



The error rates of the classification (FP, FN) depends on the direction of F.

Note that F is commonly normalized so that $\|\vec{F}\| = 1$

Assume a set of M_k training samples for each class, $\{\vec{X}_m^k\}$

The average for each class is:

$$\vec{\mu}^k = E\{\vec{X}^k\} = \frac{1}{M_k}\sum_{m=1}^{M_k}\vec{X}_m^k$$

Moments are invariant under projections. Thus the projection of the average is the average of the projection.

$$\mu_z^k = E\{F^T \cdot \vec{X}_m^k\} = F^T \cdot E\{\vec{X}_m^k\} = F^T \cdot \vec{\mu}_k$$

The inter-class distance between between classes 1 and 2 is

$$d_{12} = \mu_z^1 - \mu_z^2 = \vec{F}(\vec{\mu}_1 - \vec{\mu}_2)$$

The Fisher metric is designed to make the inter-class distance, d_{12}, as large as possible. The key concept is the "scatter" of the samples. Scatter can be seen as unormalised covariance.

The “scatter” for the M_k samples $\{\vec{X}_m^k\}$ of the set k is a matrix : **S**$_k$.
This is the same as an "unnormalised" covariance.

$$S_k = M_k \Sigma_k = \sum_{m=1}^{M_k} (\vec{X}_m^k - \vec{\mu}^k)(\vec{X}_m^k - \vec{\mu}^k)^T$$

The transformation F projects the vector $\vec{X}$ onto a scalar z.

$$z = \vec{F}^T \cdot \vec{X}$$

The scatter of the class after projection is

$$S_z^k = \sum_{m=1}^{M_k} (z_m^k - \mu_z^k)^2$$

The fisher criteria tries to maximize the ratio of the separation of the classes compared to their scatter by maximizing the ratio of within and between class scatter.

$$J(F) = \frac{\left(\mu_z^1 - \mu_z^2\right)^2}{s_z^1 + s_z^2}$$

Let us define the between class scatter as $S_B = (\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T$

then $\left(\mu_z^1 - \mu_z^2\right)^2 = F^T\left((\vec{\mu}_1 - \vec{\mu}_2)(\vec{\mu}_1 - \vec{\mu}_2)^T\right)F = F^T S_B F$

And let us define within class scatter as

$$S_W = S_1 + S_2 = \sum_{m=1}^{M_1} (\vec{X}_m^1 - \vec{\mu}_1)(\vec{X}_m^1 - \vec{\mu}_1)^T + \sum_{m=1}^{M2} (\vec{X}_m^2 - \vec{\mu}_2)(\vec{X}_m^2 - \vec{\mu}_2)^T$$

Then

$$s_z^1 + s_z^2 = F^T(S_1 + S_2)F = F^T S_W F$$

Then

$$J(F) = \frac{\left(\mu_z^1 - \mu_z^2\right)^2}{s_z^1 + s_z^2} = \frac{F^T S_B F}{F^T S_W F}$$

Taking the derivative with respect to F, we find that J(F) is maximized when

$$\left(F^T S_B F\right) S_W F = \left(F^T S_W F\right) S_B F$$

Because $S_B F$ is always in the direction $\vec{\mu}_1 - \vec{\mu}_2$

Dropping the scale factors $\left(F^T S_B F\right)$ and $\left(F^T S_W F\right)$ we obtain

$$S_W F = \vec{\mu}_1 - \vec{\mu}_2$$

and thus $F = S_W^{-1}\left(\vec{\mu}_1 - \vec{\mu}_2\right)$

Fisher's Discriminant for Multiple Classes.

Fisher's method can be extended to the derivation of K > 2 linar discriminants.
Let us assume that the number of features is greater than the number of classes, D > K.

We will look for functions that project the D features on D' < D features to form a new feature vector, $\vec{Y} = \vec{w}^T \vec{X}$ (note that there is no constant term).

as before, we define the class Mean, $\vec{\mu}_k$, class Scatter S_k and within-class scatter S_W

Class Mean: $\vec{\mu}_k = \frac{1}{M_k} \sum_{m=1}^{M_k} \vec{X}_m^k$

Class Scatter: $S_k = \sum_{m=1}^{M_k} (\vec{X}_m^k - \vec{\mu}_k)(\vec{X}_m^k - \vec{\mu}_k)^T$

Within Class Scattter $\vec{\mu}_k = \frac{1}{M_k} \sum_{m=1}^{M_k} \vec{X}_m^k$

We need to generalization of the between class covariant.
The total mean is:

$$\vec{\mu} = \frac{1}{M} \sum_{k=1}^{K} \sum_{m=1}^{M_k} \vec{X}_m^k = \frac{1}{M} \sum_{k=1}^{K} M_k \vec{\mu}_k$$

The between class scatter is:

$$S_B = \sum_{k=1}^{K} M_k (\vec{\mu}_k - \vec{\mu})(\vec{\mu}_k - \vec{\mu})^T$$

Which gives the total scatter as

$$S_T = S_W + S_B$$

We can define similar scatters in the target space:

$$\vec{\mu}_k = \frac{1}{M_k} \sum_{m=1}^{M_k} \vec{Y}_m^k \qquad \vec{\mu} = \frac{1}{M} \sum_{k=1}^{K} \sum_{m=1}^{M_k} \vec{Y}_m^k = \frac{1}{M} \sum_{k=1}^{K} M_k \vec{\mu}_k$$

$$S'_W = \sum_{k=1}^{K} \sum_{m=1}^{M_k} (\vec{Y}_m^k - \vec{\mu}_k)(\vec{Y}_m^k - \vec{\mu}_k)^T$$

$$S'_B = \sum_{k=1}^{K} M_k(\vec{\mu}_k - \vec{\mu})(\vec{\mu}_k - \vec{\mu})^T$$

We want to construct a set of projections that maximizes the between class scatter

$$J(W)=Tr\{W \cdot S_W \cdot W^T)^{-1}(WS_BW^T)$$

The W valuies are deteremined by the D' eigenvectors of $S_W^{-1}S_B$ that correspond to the D' largest Eigen values.