Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2009/2010

Lesson 19 5 May 2010

Linear Classification Methods

Contents

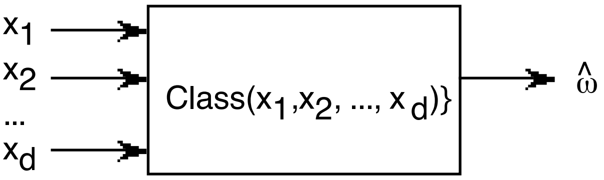
Sources Bibliographiques :
"Neural Networks for Pattern Recognition", C. M. Bishop, Oxford Univ. Press, 1995.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notation

x	a variable
X	a random variable (unpredictable value)
$\vec{x}$	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. $\omega_k = E \in T_k$.

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 lesson 16 we saw that the classification function can be decomposed into two parts: d() and $g_k()$:

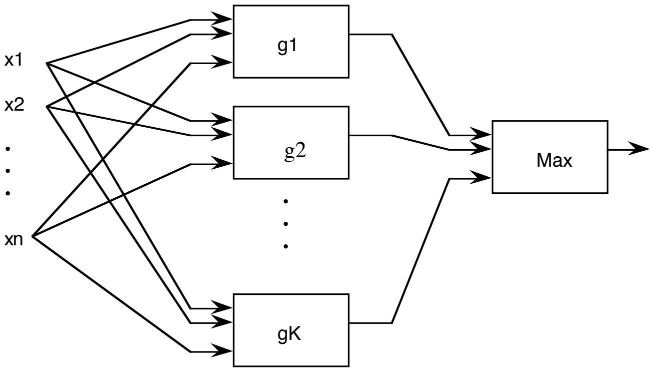
$$\hat{\omega}_k = d(\vec{g}(\vec{X}))$$

where :

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

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

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



Quadratic discrimination functions can be derived directly from p(ωk | X)

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

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

but because P(X) is constant for all k, Log{} is a monotonic function.

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

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

And when the features are modeled by a Normal density.

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

Which can be reduced to a standard (canonical) form.

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

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

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

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

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

In many cases the quadratic term can be ignored and the partitions take on the form of hyper-surfaces.

Linear Classification.

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 often independent of the class k.

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

If $\vec{N}_s$ >> $\vec{N}_k$ the term D_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 coefficients for linear functions from training data.

In communications theory, the noise is generally independent from the class. Thus is becomes possible to simplify the signal detector to:

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

where $\vec{W}_k^T$ is a “prototype” of the signal obtained as an average observation, and B is a bias or tunable gain factor. This is called a “correlation” detector.

Linear classifiers are widely used to define pattern “detection” systems, Such systems can be seen as two class discrimination. This is widely used in computer vision, for example, to detect faces or publicity logos, or other patterns of interest.

Pattern detectors as linear classifiers.

In the case of pattern detectors, generally there are two classes (K=2)

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 $\{X_m\}$ where each sample is labeled with an indicator Y_m

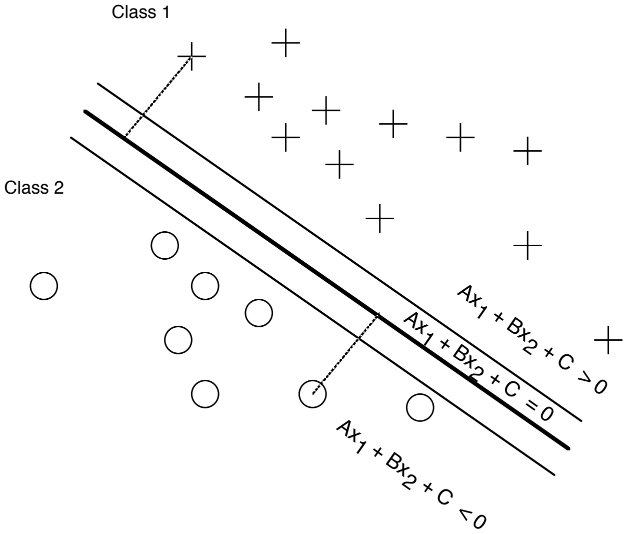
$Y_m = +1$ for examples of the target pattern (class 1)
$Y_m = -1$ for all other examples.

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

$$\vec{W}^T \vec{X} + B = 0$$

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

In this case, the decision rule reduces to a sgn function: d( )= sgn().



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.

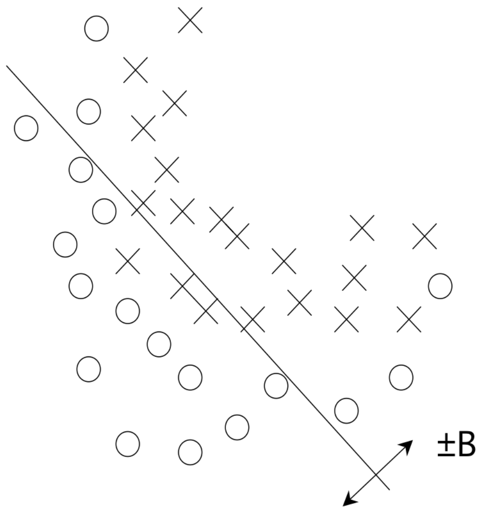
When $\vec{W}$ is normalized to unit length, $\|\vec{W}\| = 1$, then

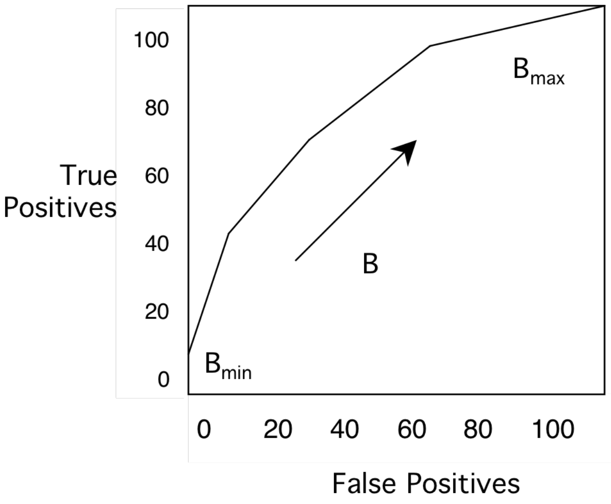
$$B = -\vec{W}^T\vec{X} \quad \text{is the perpendicular distance to the origin. .}$$

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

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 Reciever Operating Characteristics (ROC) curve.

The ROC is a powerful descriptor for the "goodness" of a linear classifier.





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

Vector between center of gravities.

Let $g_1(\vec{X}) = \vec{W}_1^T \vec{X} + B_1.$ et $g_2(\vec{X}) = \vec{W}_2 \vec{X} + B_2.$

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

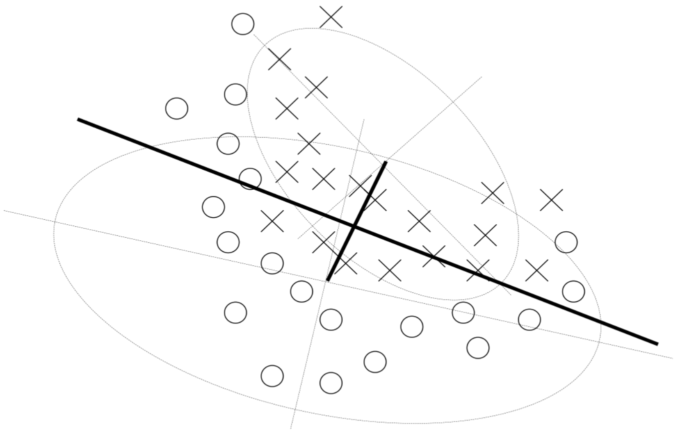
and $B_k = -\frac{1}{2}(\mu_k^T C_k^{-1} \mu_k) - \frac{1}{2} Log\{\det(C_k)\} + Log\{p(\omega_k)\}$

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.

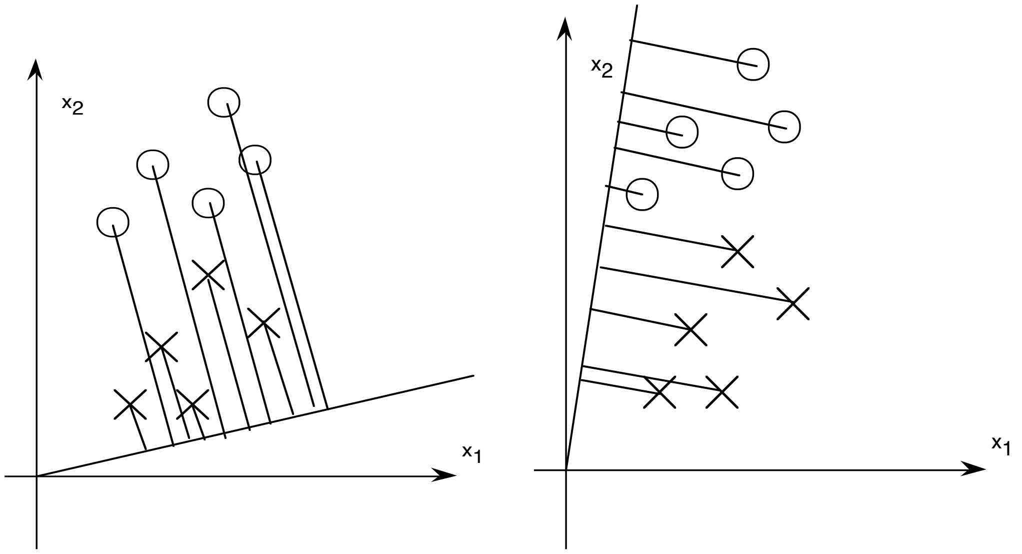
This is a reasonable choice, when the two classes are relatively compact.

Fisher Linear Discriminant (LDA).

The principle of the Fisher linear discriminant is to project the vector X with D_x is projected onto a space Z with D_z dimensions ($D_Z << D_X$) by a linear projection F.

$$\vec{z} = F^T \vec{x}$$

F is chosen such that the two classes are most separated.



The power of descrimination depends on the direction of $\vec{F}$

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

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.

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

For two classes, the inter-class distance is $d_{12} = \| \tilde{\mu}_1 - \tilde{\mu}_2 \| = \| \vec{F}^T (\vec{\mu}_1 - \vec{\mu}_2) \|$

The Fisher metric is designed to make the inter-class distance, d_{12}, as large as possible.

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

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

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

$$Z = F^T \vec{X}$$

The scatter of the class after projection is

$$\tilde{S}_k = \sum_{m=1}^{M_k} (Z_m^k - \tilde{\mu}_k)^2$$

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

$$J(F) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)}{\tilde{s}_1 + \tilde{s}_2} = \frac{\| F(\vec{\mu}_1 - \vec{\mu}_2) \|^2}{\tilde{s}_1 + \tilde{s}_2}$$

$$F = \arg-\max_F \{ \frac{\| F(\vec{\mu}_1 - \vec{\mu}_2) \|^2}{\tilde{s}_1 + \tilde{s}_2} \}$$

For K=2, $M = M_1 + M_2$

The average for each class is $\vec{\mu}_k = E\{\vec{X}^{(k)}\} = \frac{1}{M_k} \sum_{m=1}^{M_k} \vec{X}_m^{(k)}$

The complete set of $M = M_1 + M_2$ data samples is: $\{\vec{X}_m\} = \bigcup_k \{\vec{X}_m^k\}$

The average for ALL the data is:

$$\vec{\mu} = \frac{1}{M}\sum_{m=1}^{M}\vec{X}_m = \frac{1}{M}(M_1\vec{m}_1 + M_2\vec{m}_2)$$

The inter-class dispersion matrix, S_B, (B for between) is the scatter of the average of the classes.

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

The intra-class dispersion, S_W (W for within) is

$$S_W = \sum_{k=1}^{K} S_k = \sum_{k=1}^{K} M_k C_k$$

for 2 classes this is :

$$S_W = S_1 + S_2$$

Fisher showed that the best F is

$$F = \underset{F}{\operatorname{argmax}} \left\{ \frac{\| F^T S_B F \|}{\| F^T S_W F \|} \right\}$$

For K=2, $D_Z = 1$ ($F^T S_B$)

We have :

$$J(F) = \frac{F^T S_B F}{F^T S_W F}$$

It can be shown that $S_B = \lambda S_W F$.

Thus

$$S_W^{-1} S_B = \lambda F.$$

The scale factor is not important because it can be determined from the data. Thus.

$$F = S_W^{-1} S_B = S_W^{-1}(\vec{\mu}_1 - \vec{\mu}_2)$$

This is the Fisher LDA for 2 classes.
The decision surface takes the form:

$$F^T \vec{X} + b_o = 0 \qquad \text{where } F = C^{-1} (\vec{\mu}_1 - \vec{\mu}_2)$$

And b_o is an adjustable gain.

Recall that the Bayesian approach gave:

$$\vec{W}_{12}^T X + B_{12} = 0$$

where $\vec{W}_{12}^T = (C_1^{-1}\vec{\mu}_1 - C_2^{-1}\vec{\mu}_2)$
and $B_{12} = B_1 - B_2$