Pattern Recognition and Machine Learning

James L. Crowley

ENSIMAG 3 MMIS First Semester 2010/2011

Lesson 7 1 December 2010

Linear Methods for Discriminative Recognition

Contents

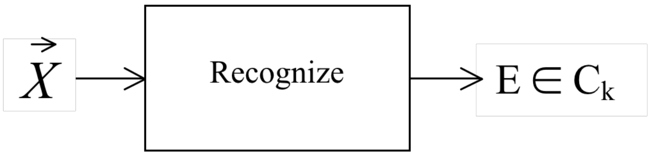
Source:
"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.

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.
k	Class index
K	Total number of classes
C_k	The kth class.
ω_k	The statement (assertion) that $E \in C_k$
$\left\{\vec{X}_m\right\}$	A set of M Training samples.
$\left\{t_m\right\}$	A set of class labels (indicators) for the samples.
	For a 2 Class problem, t_m is -1 or +1
$\left\{\vec{t}_m\right\}$	For a K class problem, $\vec{t}_m$, is a vector of 0's with a 1 in the position k of the class: $\vec{t}_m = (0,0,...,1,....,0)^T$
M	Total number of training samples. (think M = Mass)

Discriminant Recognition

As before, our problem is to build a box that assigns an observation (E, K) to one of K classes $\{C_k\}$ labeled k=1 to K. We assume that each observation, the sensor provides a vector of D features, $\vec{X}$



The decision process is decomposed into two component functions *d()* and $y(\vec{X})$:

$$\hat{\omega}_k \leftarrow d(y(\vec{X}))$$

Where $y(\vec{X})$ is a discriminant function that maps $R^D \rightarrow R^K$
d() is a decision function d(): $R^K \rightarrow \{\hat{\omega}_k\}$

Today we are going to begin looking at discriminative methods for the function $y(\vec{X})$. We can look at the function $y(\vec{X})$as a partition function, that divides the space R^D into disjoint "decision regions." The boundaries of these are the "decision surfaces".

Today we will examine methods for learning linear decision surfaces.

We will assume a training data set composed of M sample observations $\{\vec{X}_m\}$ labeled with a binary "indicator" vector $\{\vec{t}_m\}$ that gives the class k for each observation.

$$\vec{t}_m = (0,0,...,1,....,0)^T$$

Each vector $\{\vec{t}_m\}$ is composed of a vector of zeros with a single 1 in the kth position.

The decision surfaces correspond to linear functions of $\vec{X}$, followed by a non-linear function $f(\cdot)$. These are referred to as generalized linear models.

Homogeneous Coordinate Notation

Note it will often be convenient to use "homogeneous coordinates" to represent $\vec{X}$.

That is, we will add an extra "dummy" dimension to $\vec{X}$ to represent y(X) as vector product. In this case, $\vec{X}$ becomes a D+1 vector, with 1 as the last coefficient.

We also add w_o as the D+1 coefficient of $\vec{w}$

$$\vec{X} = (1, x_1, x_2, \ldots, x_D)$$
$$\vec{w} = (w_o, \vec{w})$$

The linear decision surface becomes

$$y(\vec{X}) = \vec{w}^T \cdot \vec{X} = \sum_{d=0}^{D} w_d x_d$$

The function $y(\vec{X}) = \vec{w}^T \cdot \vec{X}$ is sometimes known as a linear regression on $\vec{X}$.

There are a variety of techniques for estimating the decision surfaces

1) Least Squares Regression
2) Fisher Linear Disciminant Analysis
3) The Perceptron algorithm

Least Squares Estimation

Two Class problem. (K=2)

We first illustrate least squares with a two class problem. Our problem is to estimatea weight vector, $\vec{w}$ and a constant w_o that separates two classes. The constant is referred to as the "bias" (not to be confused with "bias" in estimating a variance).
The decision rule is

$$\text{if } y(X) = (\vec{w}^T \cdot \vec{X} + w_o) \geq 0 \text{ then } C_1 \text{ else } C_2$$

The decision surface is the hyperplane where $y(X) = \vec{w}^T \cdot \vec{X} + w_o = 0$

We can "bias" the decision surface towards class 1 or class 2 by adding a constant, b, to w_o.

If we normalize the vector $\vec{w}$ to unit norm, then

$\vec{N} = \frac{\vec{w}}{\|\vec{w}\|}$ is the normal to this hyperplane. and

$d = \frac{w_o}{\|\vec{w}\|}$ is the (signed) perpendicular distance from the plane to the origin

For least square regression, assume that each of the M training samples $\{\vec{X}_m\}$ are labeled with an indicator variable, t_m such that t_m=1 for Class 1 and t_m= –1 for class 2.

A least-squares estimate for the function $y(\vec{X}) = \vec{w}^T\vec{X} + w_o$ can be obtained in closed form.

Define a "Loss" function: $L(\hat{W}) = \sum_{m=1}^{M}(t_m - \vec{w}^T\vec{X}_m)^2$

We will use the M training samples to compose a matrix **X** and a vector **T**.

$$X = \begin{pmatrix}\vec{X}_1 & \vec{X}_2 & \cdots & \vec{X}_M\end{pmatrix} = \begin{pmatrix}1 & 1 & \cdots & 1 \\ \vec{X}_1 & \vec{X}_2 & \cdots & \vec{X}_M\end{pmatrix} \text{ (D+1 rows by M columns)}$$

T = $(t_1, t_2, ..., t_M)^T$ (M rows).

We seek the D+1 Coefficient weight vector $\vec{w} = \begin{pmatrix} w_o \\ \vec{w} \end{pmatrix}$

We write $L(\vec{w}) = (T - X^T \vec{w})^T (T - X^T \vec{w})$

To minimize the loss function, we calculate the derivative and solve for $\vec{W}$ when the derivative is 0.

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = -2X^T T + 2X^T X \vec{w} = 0$$

Thus $X^T T = X^T X \vec{w}$ and $\vec{w} = (X^T X)^{-1} X^T T$

Our decision surface is : $y(\vec{X}) = \vec{w}^T \cdot \vec{X} = 0$

The term $X^+ = (X^T X)^{-1} X^T$ is the Moore Penrose pseudo inverse.

ROC Curve

Two class linear discrimination functions are not only relatively simple – they also can be useful. 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. In such a case, a Bayesian parametric 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 Reciever Operating Characteristic curve. This curve should be used to describe or compare any method for signal or pattern detection.

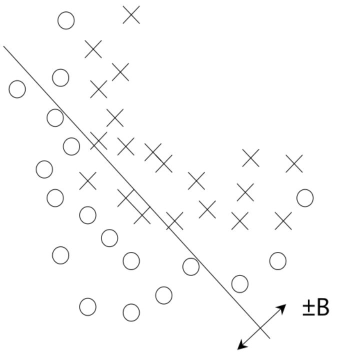
As we saw above, expressed in homogenous coordinates, the decision surface for a two class problem is a hyperplane: $y(\vec{X}) = \vec{w}^T \cdot \vec{X} = 0$

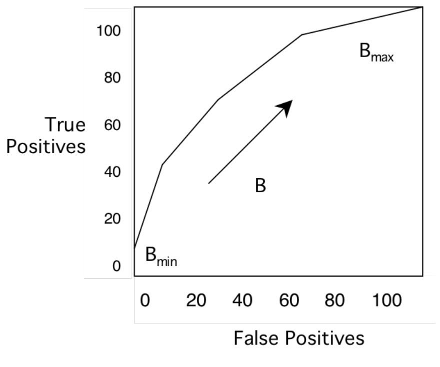
W can bias the classifier to one or the other class by adding and Bias term B.

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

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 Reciever Operating Characteristics (ROC) curve.

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





The more the curve approaches the upper left corner the better the detector.

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	
Decision		$E \in C_1$ (P)	$E \in C_2$ (N)
	$E \in C_1$ (P)	TP	FP
	$E \in C_2$ (N)	FN	TN

The ROC plots FP against FN as a function of B.

Least Squares for Multiple Class Discrimination

The Multi-class problem is a bit more complex than the two-class problem.
The set of discriminant functions must be learned together.
We will assume a training data set composed of M sample observations $\{\vec{X}_m\}$ labeled with a binary "indicator" vector $\{\vec{t}_m\}$ that gives the class k for each observation.

$$\vec{t}_m = (0,0,...,1,....,0)^T$$

Each vector $\{\vec{t}_m\}$ is composed of a vector of zeros with a single 1 in the kth position.
We need to learn a single K class discriminant function comprising K linear functions of the form

$$y_k(\vec{X}) = \vec{w}_k^T \cdot \vec{X} + w_{ko}$$

With this function, the decision surface between class i and j is:

$$y_i(\vec{X}) - y_j(\vec{X}) = (\vec{w}_i - \vec{w}_j)^T \cdot \vec{X} + (w_{io} - w_{jo})$$

The decision regions for such functions are always "singly connected" and convex. That is, any two points $\vec{X}_a$and $\vec{X}_b$within the region for class C$_k$ can be joined by a straight line segment that lies entirely within the region for class C$_k$.

Least Squares Estimation for Multi-Class Discrimination.

Least squares provides a closed form solution for the K discrimination functions $y_k(\vec{X})$. Using homogenous coordinates, we can group the K equations into a single matrix

$$\vec{Y}(\vec{X}) = \vec{w}^T \cdot \vec{X}$$

where $\vec{Y}(\vec{X})$ is a vector of K coefficients, and W^T is a matrix of K rows and D+1 columns.

Let us organize the training data into a matrix X composed of M rows and D+1 Columns, where the mth column contains the mth training sample, $\vec{X}_m$ augmented the dummy value 1.

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{D1} & x_{D2} & \cdots & x_{DM} \end{pmatrix}$$

Let us define the Truth matrix, T as an K x M matrix, where each Mth col is the binary indicator vector for the Mth data sample:

$$T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1M} \\ t_{21} & t_{22} & \cdots & t_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ t_{K1} & t_{K2} & \cdots & t_{KM} \end{pmatrix}$$

Our goal it to estimate a D+1 column by K row matrix W where each row is the coefficients for the Kth linear function $\vec{W}_k^T$.

$$W = \begin{pmatrix} w_{10} & w_{11} & w_{12} & \cdots & w_{1D} \\ w_{2o} & w_{21} & w_{22} & \cdots & w_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{Ko} & w_{K1} & w_{K2} & \cdots & w_{KD} \end{pmatrix}$$

We seek to estimate W such that we minimize a sum of square function of the form

$$L(W) = \frac{1}{2} Tr\{(WX - T)^T (WX - T)\}$$

Computing the derivative gives:

$$\frac{\partial L(W)}{\partial W} = (X^T X)W + X^T T = 0$$

Setting the derivative to 0 and rearranging we can obtain

$$W = (X^T X)^{-1} X^T T = X^+ T$$

As before, the term $X^+ = (X^T X)^{-1} X^T$ is the Moore Penrose pseudo inverse.

This gives a simple formula $\vec{Y}(\vec{X}) = W^T \vec{X}$ to estimate the K indicator variables T as $y_k(\vec{X}) = \vec{w}_k^{\,T} \cdot \vec{X}$

An interesting property of the solution $\vec{Y}(\vec{X}) = W^T \vec{X}$ is that the values are not binary. We can interpret the K values of the vector $\vec{Y}(\vec{X})$ as confidence factors for the K classes.

While least squares gives a simple closed form solution for the discriminant function, the solution is unduly influenced by outliers. A variety of alternative solutions are available.