Supervised Learning

Simplest case is learning to distinguish between two classes of objects, class +1 and class -1. Each object is encoded as a feature vector, an N-dimensional vector of real numbers, each representing a different measured feature of the object. In this way each object is represented as a point in an N-dimensional space, labeled either +1 or -1. e.g.

A decision rule (or “classification rule” or “discriminant function”) is a division of the N-dimensional feature space into two regions, one in which the prediction for the points will be +1 and one where the prediction will be -1. The simplest kind of decision rule is a hyperplane, or in 2 dimensions, a line

Even if the hyperplane does not perfectly separate the training data, it may still be an optimal decision rule, in the sense of representing the maximum likelihood, or the maximum a posteriori estimate of the correct class.

Example 1: It turns out that the TATA box classifier we designed is a linear discriminant function, if you define the feature space appropriately.
For each position $i$, $1 \leq i \leq 6$, and each base $a$ in {A, C, G, T}

$$f_{i,a} = \begin{cases} 1 & \text{if } x_i = a \\ 0 & \text{else} \end{cases}$$

where $x = x_1, \ldots, x_6$ is a 6 base window of DNA to be classified

A MAP classifier for $x$ can be described by the decision rule:

$$\begin{align*}
\text{if } \log \frac{P(\theta^{(\text{motif})} | x)}{P(\theta^{(bg)} | x)} > 0 & \quad \text{decide class = +1} \\
\text{else decide class = -1} & 
\end{align*}$$

However,

$$\log \frac{P(\theta^{(\text{motif})} | x)}{P(\theta^{(bg)} | x)} = \log \frac{P(x_i | \theta^{(\text{motif})})}{P(x_i | \theta^{(bg)})} + \log \frac{P(\theta^{(\text{motif})})}{P(\theta^{(bg)})} = s(x) \text{ (score)} + l \text{ (log prior odds)}$$

where $s(x) = \sum_{i=1}^{6} \log \frac{P(x_i | \theta^{(\text{motif})})}{P(x_i | \theta^{(bg)})}$.

For $= 1, 2, \ldots, 6$ and $a = A, C, G, T$ let $w_{i,a} = \log \frac{P(x_i = a | \theta^{(\text{motif})})}{P(x_i = a | \theta^{(bg)})}$. 
Then \( s(x) \) can also be written as
\[
s(x) = \sum_{i=1,2,\ldots,6} \sum_{a=A, C, G, T} w_{i,a} f_{i,a}
\]

So the rule is a linear rule in the feature vector.

\[
\begin{bmatrix}
f_{1,\lambda} \\
f_{1,c} \\
f_{1,g} \\
f_{1,t} \\
f_{2,\lambda} \\
\vdots \\
f_{6,\lambda}
\end{bmatrix}
\]

i.e. if \( \sum_{i,a} w_{i,a} f_{i,a} + l > 0 \) decide class = +1
else decide class = −1

If we re-label the input feature vector as \( \vec{x} = \begin{bmatrix} x^1 \\ \vdots \\ x^{24} \end{bmatrix} \)
where now the \( x_i \)'s are the binary \( f_{i,a} \) features, and we relabel the “weight vector” \( \{w_{i,a}\} \) as \( \vec{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_{24} \end{bmatrix} \)
and the log prior odds \( l \) as the “bias” \( \beta \).

Then the rule becomes
if \( \vec{w} \cdot \vec{x} + \beta > 0 \) decide class = +1
else decide class = −1.

Here \( \vec{w} \cdot \vec{x} \) is the dot product, \( \vec{w} \cdot \vec{x} = \sum_i w_i x_i \).

This class prediction is even more succinctly written as
\[
\text{sign} (\vec{w} \cdot \vec{x} + \beta) \quad \text{where} \quad \text{sign} (y) = \begin{bmatrix} +1 \text{ if } y > 0 \\ -1 \text{ else} \end{bmatrix}
\]

\( \vec{w} \cdot \vec{x} + \beta \) is the canonical form for a linear function in the space defined by vectors \( \vec{x} \) (the feature space).
The geometry of linear discriminate functions:

Let us scale the decision rule by dividing through by the length of the vector \( \mathbf{w} \), i.e. \( ||\mathbf{w}|| = \sqrt{\sum w_i^2} \).

\[
\text{sign} (\mathbf{w} \cdot \mathbf{x} + \beta) \text{ becomes } \text{sign} \left( \frac{\mathbf{w}}{||\mathbf{w}||} \cdot \mathbf{x} + \frac{\beta}{||\mathbf{w}||} \right)
\]

which is an equivalent function, with a new weight vector \( \mathbf{w}' = \frac{\mathbf{w}}{||\mathbf{w}||} \)

and a new bias \( \beta' = \frac{\beta}{||\mathbf{w}||} \), \( ||\mathbf{w}'|| = 1 \).

So, since this function is the same, without loss of generality, we can assume that the length of \( \mathbf{w} \) is 1. In this case we can easily and intuitively draw the discriminant function:

Let \( T = -\beta \). \( T \) is called the threshold.

**Paramecric Models**

The simplest parametric models for the classification problem are Gaussian models. Consider the 1-dimensional feature space. The input is a scalar \( x \), the decision function is just a single cut off value

\[
\text{Predict } -1 \quad \text{Predict } +1
\]

Suppose the class \(-1\) data is Gaussian with mean 0 and variance \( \sigma^2 \) (if not, shift the axis to make it mean 0).

Suppose the class \(+1\) data is Gaussian with mean \( \mu \) and variance \( \sigma^2 \). Then the MAP Classifier is

\[
\text{sign} \left( \log \frac{P(x|\theta^{(1)})}{P(x|\theta^{(-1)})} + l \right)
\]
Where \( \theta^{(+1)} = (\mu, \sigma) \) and \( \theta^{(-1)} = (0, \sigma) \) and \( l = \log \frac{P(x|\theta^{(+1)})}{P(x|\theta^{(-1)})} \)

\[
\log \frac{P(x|\theta^{(+1)})}{P(x|\theta^{(-1)})} = \log \left[ \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2} \right] = \log \left( \frac{1}{e} \right) \left[ \left( \frac{x}{\sigma} \right)^2 - \left( \frac{x - \mu}{\sigma} \right)^2 \right]
\]

\[= \frac{1}{2\sigma^2} [x^2 - x^2 + 2x \mu - \mu^2] = \frac{\mu}{\sigma^2} x - \frac{\mu}{2\sigma} \]

So the rule is

\[
sign \left( \frac{\mu}{\sigma^2} x - \frac{\mu}{2\sigma} + l \right) = sign \left( x - \frac{\mu}{2} + \frac{\sigma^2}{\mu} \right) = sign (x + \beta), \text{ where } \beta = \frac{\sigma^2}{\mu} - \frac{\mu}{2}.
\]

Note that if the prior odds are even, so \( l = 0 \). Then the decision rule is \( sign (x - \frac{\mu}{2}) \) i.e. you have a cut-off halfway between the mean of the \(-1\) class (0) and the mean of the \(+1\) classes (\(\mu\)). Otherwise the prior odds bias you toward one class or the other.

**General N-dimensional Case**

In this case the Gaussian is defined by a mean vector \( \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} \), \( \mu_i = E(x_i) \)

and a covariance matrix \( \Sigma \), where \( \Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] \).

\[
P(x|\theta) = P(x|\hat{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2} (\vec{x} - \hat{\mu})^T \Sigma^{-1} (\vec{x} - \hat{\mu})}
\]

where \( \vec{v} \) is the transpose \([v_1 \ldots v_N]\) of the vector \( \vec{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} \) and \( |\Sigma| \) is the determinant of the matrix \( \Sigma \).

Assume the mean of the \(-1\) class is 0, and the covariance matrix is the same for both classes, equal to \( \Sigma \). Let the mean of the \(+1\) class be \( \hat{\mu} \).

\( \theta^{(-1)} = (0, \Sigma) \) and \( \theta^{(+1)} = (\mu, \Sigma) \)

\[
\log \frac{P(x|\theta^{(+1)})}{P(x|\theta^{(-1)})} = \log \left( \frac{1}{e} \right) \frac{1}{2} \left[ x^T \Sigma^{-1} x + (\vec{x} - \hat{\mu})^T \Sigma^{-1} (\vec{x} - \hat{\mu}) \right]
\]

\[= \frac{1}{2} \left[ x^T \Sigma^{-1} x + \vec{x}^T \Sigma^{-1} \vec{x} + \hat{\mu}^T \Sigma^{-1} \hat{\mu} + \vec{x}^T \Sigma^{-1} \vec{x} - \vec{x}^T \Sigma^{-1} \hat{\mu} - \hat{\mu}^T \Sigma^{-1} \vec{x} + \hat{\mu}^T \Sigma^{-1} \hat{\mu} \right]
\]

\[= \frac{1}{2} \left[ x^T \Sigma^{-1} x + \vec{x}^T \Sigma^{-1} \vec{x} - \vec{x}^T \Sigma^{-1} \hat{\mu} - \hat{\mu}^T \Sigma^{-1} \vec{x} + \hat{\mu}^T \Sigma^{-1} \hat{\mu} \right]
\]

which is a linear function on \( x_1 \ldots x_N \).
In general, if the covariance matrix is different for the two classes, the discriminant function is quadratic.

Examples: