Linear models for classification

\[ \tilde{x} \rightarrow C_k, \quad k = 1, \ldots, K \]

\text{Input} \downarrow \text{Discrete classes}

\text{Input space divided into decision regions by decision surfaces.}

For \( \tilde{t} = (0, 1, 0, 0, 0) \)

\text{ex., target variables} \quad K=5 \text{ classes, target variable indicates class 2}

\text{Classification approaches:}

1. \text{Discriminant function: directly assigns} \tilde{x} \text{ to a class, e.g. for 2 classes}\n\[ y(\tilde{x}) > 0 \iff C_1, \quad y(\tilde{x}) < 0 \iff C_2. \]

2. \text{Probabilistic approach:} \quad \begin{align*}
\text{use} \quad p(C_k | \tilde{x}) &= \frac{p(\tilde{x} | C_k) p(C_k)}{p(\tilde{x})} \\
\text{Bayesian framework}
\end{align*}

Previously, we focused on \( \text{explicit bias term} \)
\[ y(\tilde{x}) = \tilde{w}^T \tilde{\phi}(\tilde{x}) + b_0. \]

Now we will consider \( y(\tilde{x}) = f(\tilde{w}^T \tilde{\phi}(\tilde{x}) + b_0) \)

or, more generally, \( y(\tilde{x}) = f(\tilde{w}^T \tilde{\phi}(\tilde{x}) + b_0) \) \text{ non-linear activation function}
Decision surfaces are given by
\[ \mathbf{w}^T \mathbf{x} + w_0 = \text{const} \quad \text{linear f's of } \mathbf{x} \]

\[ \text{Discriminant functions} \]

1. **Two classes**

Consider \[ y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 \quad \text{[linear discriminant]} \]
\[
\begin{align*}
\{ \quad & y(\mathbf{x}) > 0 \Rightarrow C_1 \\
& y(\mathbf{x}) < 0 \Rightarrow C_2
\end{align*}
\]

\[ y(\mathbf{x}) = 0 \equiv \text{decision boundary (DB)} \]

Consider \( \mathbf{x}_A, \mathbf{x}_B \in \text{DB} \)

\[ y(\mathbf{x}_A) = y(\mathbf{x}_B) = 0 \Rightarrow \mathbf{w}^T (\mathbf{x}_A - \mathbf{x}_B) = 0 \]

- lies on \((D - 1)\) dim's DB

So, \( \mathbf{w} \perp \text{DB} \Rightarrow \mathbf{n} = \frac{\mathbf{w}}{||\mathbf{w}||} \) is a unit \text{vector in DB}

Similarly, if \( \mathbf{x} \in \text{DB}, \)

\[ y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 = 0 \quad \text{or} \]
\[ \frac{\mathbf{w}^T \mathbf{x}}{||\mathbf{w}||} = -\frac{w_0}{||\mathbf{w}||} \]

\[ \Rightarrow \quad \mathbf{n} \cdot \mathbf{x} = \text{normal distance} \]
from the origin to DB

So, \( w_0 \) determines the location of DB
Moreover, for any $x$,

$$x = x_{\parallel \text{DB}} + \frac{y}{\|y\|} \quad \text{in DB}$$

Then

$$y(x) = \frac{y(x)}{\|y\|} \|y\|$$

Here, we used

$$\|y\| x_{\parallel \text{DB}} + u_0 = 0$$

Indeed,
Then
\[-\frac{1}{n} \frac{w_0}{\|w\|} + \bar{x},_2 = \bar{x}\]

\[= \bar{x},_2, \text{LLDB} + \bar{x},_2, \text{LLDB} \]

\[= \bar{w}^T w_0 + \bar{w}^T \bar{x},_2, \text{LLDB} + r \frac{\bar{w}^T w_0}{\|w\|} \]

\[= r \bar{w}^T w_0 \sqrt{\|w\|} \]

\[
\bar{w}^T \bar{x}_i + w_0 = r \|w\| \quad , \text{as before} \]

\[y(x) \]

2. **Multiple classes**

Consider \(K \geq 2\) classes. Difficulties in generalizing from the \(K=2\) case:

![Diagram of multiple classes](image)

Rather, consider a single \(K\)-class discriminant:

\[y_k(x) = \bar{w}_k^T \bar{x} + w_0 \]

Assign a point to \(C_k\) if \(y_k(x) > y_j(x), \forall j \neq k\).
D\v{s}s are then given by $y_k(x) = y_j(x)$, s.t.

$$(\omega_k - \omega_j)^T x + (\omega_{k,0} - \omega_{j,0}) = 0$$

Now, consider $\bar{x}_A, \bar{x}_B \in C_k$

The line connecting $\bar{x}_A$ and $\bar{x}_B$

$$\bar{x} = \lambda \bar{x}_A + (1 - \lambda) \bar{x}_B, \quad 0 \leq \lambda \leq 1$$

By linearity of discriminant $f$'s

$$y_k(\bar{x}) = \lambda y_k(\bar{x}_A) + (1 - \lambda) y_k(\bar{x}_B)$$

$$> y_j(\bar{x}_A), \quad \forall j \neq k$$

$$> y_j(\bar{x}_B), \quad \forall j \neq k$$

$$y_k(\bar{x}) > y_j(\bar{x}), \quad \forall j \neq k$$

So, $\bar{x} \in C_k$ as well.

Since $\bar{x}_A, \bar{x}_B$ are arbitrary, $C_k$ is

arbitrary, $C_k$ is

singly connected and

convex, $\forall k$. Convex

concave

-5-
Least squares for classification

Consider a problem with \( K \) classes, s.t. \( i \) are \( K \)-dim unit vectors.

\[
y_k(x) = \mathbf{w}_k^T \mathbf{x} + \mathbf{b}_{k0} \Rightarrow \hat{y}(x) = \tilde{\mathbf{w}}^T \tilde{x}
\]

\[
\tilde{\mathbf{w}} = \begin{bmatrix}
\mathbf{w}_{k0} \\
\mathbf{w}_{k1} \\
\vdots \\
\mathbf{w}_{kD}
\end{bmatrix}
\]

\( D+1 \) (prms) or \# entries in \( \tilde{x} \)

\( k \)th column

\[
\tilde{x} = (1, \mathbf{x})^T
\]

Training set: \( \{\tilde{x}_n, t_n\} \quad n=1, \ldots, N \)

Define

\[
\mathbf{T} = \begin{bmatrix}
t_{n,0} & t_{n,1} & \ldots & t_{n,K-1}
\end{bmatrix}^T
\]

\( t_n \) is the \( n \)th row of \( \mathbf{T} \)

\[
\tilde{\mathbf{x}} = \begin{bmatrix}
x_{n,0} \\
\vdots \\
x_{n,D}
\end{bmatrix}
\]

\( D+1 \)

\( \tilde{x}_n \) is the \( n \)th row of \( \tilde{x} \)

\( \tilde{x} \tilde{\mathbf{w}} \) is an \( N \times K \) matrix like \( \mathbf{T} \)

-6-
Then \[ E(\tilde{W}) = \frac{1}{2} \operatorname{Tr} \{ (\tilde{X}\tilde{W} - T)^T(\tilde{X}\tilde{W} - T) \} \].

Indeed,
\[
E(\tilde{W}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} \left( (\tilde{X}_n)_{nk} - t_{nk} \right)^2 \quad \otimes \\
\sum_{j=0}^{d} \tilde{X}_{nj} \tilde{W}_{jk} = \sum_{j=0}^{d} \tilde{X}_{nj} \tilde{w}_{k,j} = \tilde{w}_{k,0} + \tilde{w}_{k}^T \cdot \tilde{X} \\
\otimes \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} \left( \sum_{j=0}^{d} \tilde{w}_{k,j} \tilde{X}_{nj} - t_{nk} \right)^2.
\]

Then
\[
\frac{\partial E}{\partial w_{k,j}} = \frac{1}{2} 2 \sum_{n} \left[ \sum_{j} \tilde{w}_{k,j} \tilde{X}_{nj} - t_{nk} \right] \tilde{X}_{nj} = 0, \quad \forall \ k,j
\]

\[
(\tilde{X}^T \tilde{X})\tilde{W} = \tilde{X}^T \tilde{T}
\]

Finally, \[
\tilde{W} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{T}
\] "\(\tilde{X}^+\) pseudo-inverse of \(\tilde{X}\)

\[
\tilde{Y}(\tilde{X}) = \tilde{W}^T \tilde{X} = T^T (\tilde{X}^+)^T \tilde{X}.
\]

This is a closed-form solution which however is sensitive to outliers.
And may even fail completely:

Indeed, least-squares assumes Gaussian distribution of $E$'s, and binary target vectors often have non-Gaussian distributions.