Recall discriminant function $g_j(x)\), j=1,...,K.

Choose $C_i$ of $g_i(x) = \max_{j=i} g_j(x)$

In previous chapters, the $g_j(x)$ were obtained from Bayesian decision theory by estimating probabilities $p(x|C_i), p(C_i)$ from the data.

Now discriminant-based classification where we assume a model for the discriminant

Model $g_i(x|\phi_i)$ where set of parameter $\phi_i$ now need to learn the $\phi_i$ from data.

In this lecture (Chp 10), we assume model $g_i(x|w_i, w_0) = w_i^T x + w_0$

linear discriminant.

I recall linear discriminant arises from Bayes decision of the distribution $p(x|C_i)$ are Gaussian with identical covariance.

Simple generalizations

$g_j(x|\nu_i, \omega_i, w_0) = \nu_i^T x + \omega_i + w_0$

for other generalizations - see kernel trick.
Geometry of Linear Discriminant

Two classes:

\[ g(x) = g_1(x) - g_2(x) = (\omega_1 - \omega_2)^T x + (\omega_{10} - \omega_{20}) \]

Choose \( C_1 \) if \( g(x) > 0 \)

\( C_2 \) otherwise

\( g(x) = 0 \) defines a hyperplane

it divides the space into

two regions \( g(x) > 0 \) \( C_1 \)

\& \( g(x) < 0 \) \( C_2 \).

The normal to the hyperplane is \( \omega \)

closest distance to the origin is \( \frac{w_{10}}{||w||} \)

If \( K > 2 \) classes:

\( K \) discriminant functions

Adjust the parameters \( \omega_i, \omega_{10} \), so that

\[ g_i(x) = w_i^T x + \omega_{i0} \geq 0, \quad \text{if } x \in C_i \]

\[ g_i(x) < 0, \quad \text{otherwise} \]

In general, it will not be possible to linearly separate the data

(see support vector machines)
\(X = \{x^t, r^t\} \quad r^t = 1, \text{ if } x^t \in C_1, \quad r^t = -1, \text{ if } x^t \in C_2.\)

Want to find \(w \perp w_0\) s.t.

\[
\begin{align*}
&\forall x^t: w^T x^t + w_0 \geq 1 & \text{if } r^t = +1 \\
&\forall x^t: w^T x^t + w_0 \leq -1 & \text{if } r^t = -1.
\end{align*}
\]

re-express as \(r^t (w^T x^t + w_0) \geq 1\) why "11" require the data to be past the margin.

\[
\begin{align*}
&\min_{w \in \mathbb{R}^d} \frac{1}{2} ||w||^2 \\
&\text{subject to } r^t (w^T x^t + w_0) \geq 1, \forall x^t
\end{align*}
\]

make the margin \(\frac{1}{||w||}\) as big as possible (why? more likely to generalize)

Distance of \(x^t\) to the plane \(w^T x^t + w_0 = 0\)

\[x^t \triangleq x^t - \hat{x} w\]

this hits the plane when

\[
\begin{align*}
&\hat{x}^T x^t - \hat{x}^T \hat{x} w + w_0 = 0 \\
&\Rightarrow \hat{x} = \frac{w^T x^t + w_0}{||w||^2}
\end{align*}
\]

distance to plane is

\[
\frac{1}{||w||} = \frac{1}{w^T x^t + w_0}.
\]

sign-distance is

\[
\frac{1}{||w||}.
\]
Formalize as constrained optimization problem

\[ L_p = \frac{1}{2} \|w\|^2 - \sum_{t=1}^{N} x_t \mathbf{r}_t (\mathbf{w}^T \mathbf{x}_t + \omega_0) - 1 \]

minimized w.r.t. \( w \)

maximized w.r.t. \( \alpha^t \)

note if \( r^t (\mathbf{w}^T \mathbf{x}_t + \omega_0) - 1 > 0 \)

then we maximize \( \alpha^t \) by setting \( \alpha^t = 0 \)

convert optimization problem:

\[ \frac{\partial L_p}{\partial w} = 0 \Rightarrow \mathbf{w} = \frac{1}{N} \sum_{t} \alpha^t r^t \mathbf{x}_t \]

\[ \frac{\partial L_p}{\partial \alpha^t} = 0 \Rightarrow \mathbf{x}_t^T \mathbf{r}_t = 0. \]

Substituting back into \( L_p \) gives the dual

\[ L_d = -\frac{1}{2} \sum_{t} \sum_{s} \alpha^t \alpha^s \mathbf{r}_t \mathbf{r}_s (\mathbf{x}_s^T \mathbf{x}_t + 2 \omega_0 \mathbf{x}_s \cdot \mathbf{x}_t) \]

\[ \text{goal: maximize w.r.t. } \{ \alpha^t \} \text{ subject to } \sum_{t} \alpha^t \mathbf{r}_t = 0 \text{ and } \alpha^t > 0 \text{ for } t \]

\[ \text{can be solved by quadratic optimization methods, time complexity } O(N^3) \]

\[ \text{space complexity } O(N^2) \]
Support Vectors:

the solution is of form \( w = \sum_{t} \lambda^t r^t x^t \).

But \( \lambda^t = 0 \) unless \( r^t (w_0^T x^t + \omega_0 - 1) = 0 \)
which happens only if \( x^t \) is on the margin
- i.e. distance of \( x^t \) to the hyperplane is \( \frac{1}{|w|} \).

So only the data points on the margin determine the
separating hyperplane.

To determine \( \omega_0 \), we only need to find one support vector \( x^s \) - support
vectors.

- any data vector with \( z^s > 0 \),

then \( \omega_0 = 1 - \frac{\omega_0^T x^s}{z^s} \)

where \( \omega_0 = \frac{1}{z^s} \sum_{t} \lambda^t r^t x^t \)

but for stability, average over all the support

vector \( \omega_0 = 1 - \frac{1}{|s|} \sum_{s \in S} \frac{\omega_0^T x^s}{z^s} \) where \( S = \{ t : z^t > 0 \} \).
(6) \textbf{Nonseparable Case}

Define slack variables \( \xi_t \geq 0 \) to allow a data point to move.

\[
p_t (\omega^T x^t + \omega_0) \geq 1 - \xi_t
\]

rewrite as

\[
1 - \omega_0 \leq p_t (\omega^T x^t + \omega_0) \leq \frac{1}{\omega_0^2}
\]

Think of this as moving the data point to put it on the correct side of the margin.

pentalogue slackness by \( \frac{2}{\xi_t} \)

minimize

\[
\min_{\xi_t} \quad \frac{1}{2} \| \omega \|^2 + C \sum_t \xi_t - \frac{1}{2} \sum_t (\omega^T x^t + \omega_0)^2 \frac{1}{1 - \xi_t}
\]

maximize w.r.t. \( (\omega^t, \xi_t, \mu_t) \)

subject to \( \mu_t \geq 0 \), set \( \mu_t = 0 \)

To obtain the dual

\[
\frac{\partial L}{\partial \omega} = 0, \quad \frac{\partial L}{\partial \xi_t} = 0
\]

\[
L_d = \frac{1}{2} \sum_t x_t^T x_t - \frac{1}{2} \sum_t \lambda_t \xi_t - \sum_t \omega^T x_t
\]

subject to \( \frac{1}{2} \lambda_t = 0 \), \( 0 \leq \lambda_t \leq \infty \), \( t \neq 0 \)

Solve the dual - quadratic optimization - to obtain the \( \hat{\xi}_t \), solve \( \hat{\omega} = \sum \hat{x}_t^T \hat{x}_t^t \) to get \( \hat{\omega} \), estimate \( \omega_0 \) from support vectors as before.
The Kernel Trick.

Example: Go from $x$-space to $z$-space

$z_1 = x_1$, $z_2 = x_2$, $z_3 = x_1^2$, $z_4 = x_2^2$, $z_5 = x_1 x_2$

take $z = (z_1 \ldots z_5)$ as input.

linear function in $z$-space is non-linear in $x$-space

$z = \Phi(x)$, $z_j = \phi_j(x)$, $j = 1 \ldots k$

$g(z) = w^T \Phi(x)$

The solution of (VM) is

$w = \frac{1}{t} \sum x^t \Phi(x)^T \Phi(x)$

The discriminant is

$g(x) = w^T \Phi(x) = \frac{1}{t} \sum x^t \Phi(x)^T \Phi(x)$

kernel trick - all that matters is

$K(x^t, x) = \Phi(x)^T \Phi(x)$

we don't care what $\Phi(x)$ is.

$g(x) = \frac{1}{t} \sum x^t K(x^t, x)$

Kernels - three main types:

- Polys $K(x^t, x) = (x^T x^t + 1)^d$
- Radial basis function $K(x^t, x) = \exp(-\frac{||x^t - x||^2}{\sigma^2})$
- Sigmoid function $K(x^t, x) = \tanh(2x^T x^t + 1)$