(1)
Alpaydin.
Chp 10.

Linear Discrimination

4/27/2008

Recall discriminant functions $g_j(\underline{x})$, $j = 1, \ldots K$.

Choose C_i if $g_i(\underline{x}) = \max_{j=1}^{K} g_j(\underline{x})$

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

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

Model $g_i(\underline{x} | \Phi_i)$ Φ_i set of parameters
$\rightarrow$ now need to learn the Φ_i from data!

In this lecture (chp 10), we assume model
$g_i(\underline{x} | \underline{w}_i, w_{i0}) = \underline{w}_i^T \underline{x} + w_{i0}$
linear discriminant.

(recall linear discriminant arises from Bayes decision if the distribution $p(\underline{x} | C_i)$ are Gaussian with identical covariance).

simple generalizations
$g_i(\underline{x} | \underline{\underline{W}}_i, \underline{w}_i, w_{i0}) = \underline{x}^T \underline{\underline{W}}_i \underline{x} + \underline{w}_i^T \underline{x} + w_{i0}$

for other generalizations – see kernel trick.

(2)

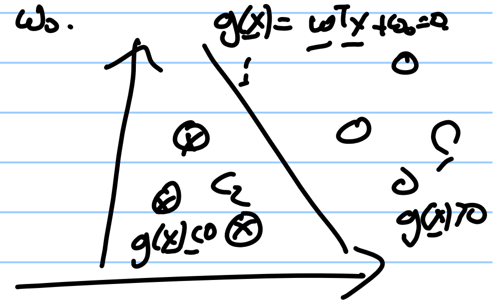
Geometry of Linear Discriminant

Two Classes: $g(\underline{x}) = g_1(\underline{x}) - g_2(\underline{x})$

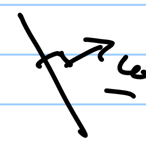
$$= (\underline{\omega}_1 - \underline{\omega}_2)^T \underline{x} + (\omega_{10} - \omega_{20})$$

$$= \underline{\omega}^T \underline{x} + \omega_0.$$

Choose C_1 if $g(\underline{x}) > 0$
C_2 otherwise



$g(\underline{x}) = 0$ defines a hyperplane
it divides the space into
two regions $g(\underline{x}) > 0$ C_1
& $g(\underline{x}) < 0$ C_2.

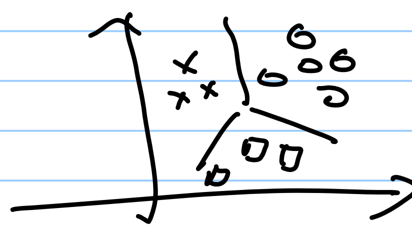


The normal to the hyperplane is $\underline{\omega}$.
closest distance to the origin is $|\omega_0| / \|\underline{\omega}\|$

If $K > 2$ classes:
K discriminant functions
Adjust the parameters $\underline{\omega}_i, \omega_{i0}$, so that

$$g_i(\underline{x} \mid \underline{\omega}_i, \omega_{i0}) \begin{cases} > 0, & \text{if } \underline{x} \in C_i \\ \leq 0, & \text{otherwise.} \end{cases}$$

In general, it will
not be possible to linearly
separate the data
(see support vector machines)



(3)

Support Vector Machines Two class

$\mathcal{X} = \{\underline{x}^t, r^t\}$ $\quad r^t = +1$, if $\underline{x}^t \in C_1$, $r^t = -1$, if $\underline{x}^t \in C_2$.

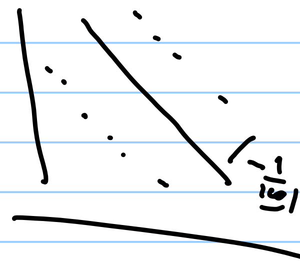
Want to find $\underline{\omega}$ & ω_0 s.t.

$$\underline{\omega}^T \underline{x}^t + \omega_0 \geq 1 \quad \text{if } r^t = +1$$
$$\underline{\omega}^T \underline{x}^t + \omega_0 \leq -1 \quad \text{if } r^t = -1.$$

re-express as $\quad r^t(\underline{\omega}^T \underline{x}^t + \omega_0) \geq 1$

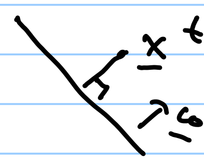
why "1" require the data to be past the margin.

min $\frac{1}{2}|\underline{\omega}|^2$ subject to
$$r^t(\underline{\omega}^T \underline{x}^t + \omega_0) \geq 1, \ \forall t$$



Make the margin $\frac{1}{|\underline{\omega}|}$ as big as possible. (why? more likely to generalize)

Distance of $\underline{x}^t$ to the plane $\underline{\omega}^T \underline{x} + \omega_0 = 0$



line $\quad \underline{x}(\lambda) = \underline{x}^t - \lambda \underline{\omega}$

this hits the plane when.

$$\underline{\omega}^T\{\underline{x}^t - \hat{\lambda}\underline{\omega}\} + \omega_0 = 0. \quad \Rightarrow \quad \hat{\lambda} = \frac{\underline{\omega}^T \underline{x}^t + \omega_0}{|\underline{\omega}|^2}$$

distance to plane is $|\hat{\lambda}\underline{\omega}| = \frac{|\underline{\omega}^T \underline{x}^t + \omega_0|}{|\underline{\omega}|}$.

sign-distance is $\frac{\underline{\omega}^T \underline{x}^t + \omega_0}{|\underline{\omega}|}$

4)

Formalize as constrained optimization problem
→ find separating plane with biggest margin.

$$L_p = \frac{1}{2}\|\underline{w}\|^2 - \sum_{t=1}^{N} \alpha^t [r^t(\underline{w}^T \underline{x}^t + w_0) - 1]$$

$\{\alpha^t\}$ Lagrange parameters
constrained so that $\alpha^t \geq 0$

minimized w.r.t. $\underline{w}$
maximized w.r.t. α^t.
note if $r^t(\underline{w}^T \underline{x}^t + w_0) - 1 > 0$
then we maximize α^t by setting $\alpha^t = 0$

convex optimization problem.

$$\frac{\partial L_p}{\partial \underline{w}} = 0 \Rightarrow \underline{w} = \sum_t \alpha^t r^t \underline{x}^t$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_t \alpha^t r^t = 0.$$

Substituting back into L_p gives the dual

$$L_d = -\frac{1}{2} \sum_t \sum_s \alpha^t \alpha^s r^t r^s (\underline{x}^t)^T \underline{x}^s + \sum_t \alpha^t.$$

goal maximize w.r.t. $\{\alpha^t\}$ subject to $\sum_t \alpha^t \underline{r}^t = 0$ & $\alpha^t \geq 0$, $\forall t$

can be solved by quadratic optimization methods. time complexity $O(N^3)$
space complexity $O(N^2)$.

(5) **Support Vectors:**

the solution is of form $\underline{\omega} = \sum_t \alpha^t r^t \underline{x}^t$.

But $\alpha^t = 0$ unless $r^t(\underline{\omega}^T \underline{x}^t + \omega_0 - 1) = 0$
which happens only if $\underline{x}^t$ is on the margin
– ie. distance of $\underline{x}^t$ to the hyperplane is $\frac{1}{|\underline{\omega}|}$

So only the datapoints on the margin determine the seperating hyperplane.

support vectors.

To determine ω_0, we only need to find one support vector $\underline{x}^s$ – any datavector with $\alpha^s > 0$ –.

then $\omega_0 = 1 - \hat{\underline{\omega}}^T \underline{x}^s$

where $\hat{\underline{\omega}} = \sum_t \hat{\alpha}^t r^t \underline{x}^t$

But, for stability, average over all the support vectors

$$\omega_0 = 1 - \frac{1}{|S|} \sum_{s \in S} \hat{\underline{\omega}}^T \underline{x}^s \quad \text{where } S = \{t : \alpha_t > 0\}$$

(6) Nonseparable Case

Define slack variables $\xi^t \geq 0$ to allow a data point to move.

$$r^t(\underline{w}^T \underline{x}^t + w_0) \geq 1 - \xi^t$$

rewrite as $r^t\left(\underline{w}^T\left(\underline{x}^t + \frac{r^t \underline{w}\, \xi^t}{|\underline{w}|^2}\right) + w_0\right) \geq 1.$

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

penalize slackness by $\sum_t \xi^t$

minimize w.r.t. $\underline{w}, \{\xi^t\}$, maximize w.r.t. $\{\alpha^t\}, \{\mu^t\}$

$$L_p = \frac{1}{2}|\underline{w}|^2 + C\sum_t \xi^t - \sum_t \alpha^t\{r^t(\underline{w}^T\underline{x}^t + w_0) - 1 - \xi^t\} - \sum_t \mu^t \xi^t$$

If $\xi^t > 0$, set $\mu^t = 0$.

To obtain the dual $\frac{\partial L_p}{\partial \underline{w}} = 0$ $\frac{\partial L_p}{\partial \xi^t} = 0$

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t\sum_s \alpha^t \alpha^s r^t r^s (\underline{x}^t)^T \underline{x}^s$$

subject to $\sum_t \alpha^t r^t = 0$ $\quad 0 \leq \alpha^t \leq C, \ \forall t$

Solve the dual — quadratic optimization — to obtain the $\hat{\alpha}^t$, solve $\hat{\underline{w}} = \sum_t \hat{\alpha}^t r^t \underline{x}^t$ to get $\hat{\underline{w}}$, estimate w_0 from support vectors as before.

(7)

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 $z = (z_1 \dots z_5)$ as input.

linear function in z-space is non-linear in X-space

$$\underline{z} = \underline{\Phi}(\underline{x}) \quad , z_j = \phi_j(\underline{x}),\ j = 1 \dots k.$$

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

The solution of SVM is $\quad \underline{w} = \sum_t \alpha^t r^t \underline{z}^t = \sum_t \alpha^t r^t \underline{\Phi}(\underline{x}^t)$

The discriminant is $\quad g(\underline{x}) = \underline{w}^T \underline{\Phi}(\underline{x}) = \sum_t \alpha^t r^t \underline{\Phi}(\underline{x}^t)^T \underline{\Phi}(\underline{x})$

kernel trick – all that matters is

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

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

$$g(\underline{x}) = \sum_t \alpha^t r^t K(\underline{x}^t, \underline{x})$$

Kernels $\rightarrow$ three main types.

- polys $\quad K(\underline{x}^t, \underline{x}) = (\underline{x}^T \underline{x}^t + 1)^d$
- radial basis function $\quad K(\underline{x}^t, \underline{x}) = \exp\left\{ - \frac{\|\underline{x}^t - \underline{x}\|^2}{\sigma^2} \right\}$
- sigmoid function $\quad K(\underline{x}^t, \underline{x}) = \tanh(2\underline{x}^T \underline{x}^t + 1)$