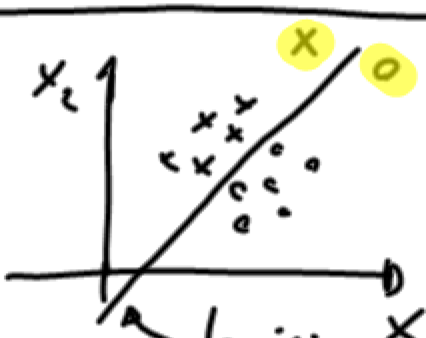
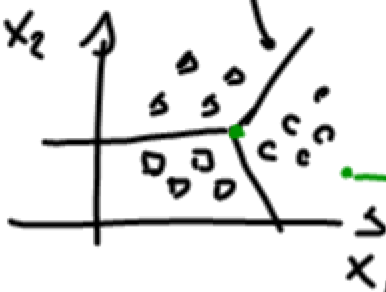
Linear models for classification



- we saw that we can divide the input space in regions, and assign a label to each of them

decision boundary ← when linear

linear methods for classification



$\hat{f}_o(x) = \hat{f}_\Delta(x) = \hat{f}_\square(x)$

$\hat{f}_\square(x) > \hat{f}_\Delta(x) \wedge \hat{f}_o(x) > \hat{f}_\Delta(x)$

until now: classification based on linear regression

→ it is a member of a class of methods, namely methods based on discriminant functions

$\forall k \quad \hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$

decision boundary between two classes k and ℓ

$x: \hat{f}_k(x) = \hat{f}_\ell(x)$

discriminant functions $\delta_k(x)$

$$\left.\begin{aligned} f_A(x) &= \Pr[G=A \mid X=x] \\ f_B(x) &= \Pr[G=B \mid X=x] \\ f_C(x) &= \Pr[G=C \mid X=x] \end{aligned}\right\} \text{posterior probabilities}$$

If $\delta_k(x)$ or $\Pr[G=k \mid X=x]$ are linear in x → linear decision boundaries

Actually, we only need monotone transformation of $\delta_k(x)$ or $Pr[G=k|X=x]$ to be linear

Examples

(i) $\hat{\delta}_k(x) = \hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_{k1} x_1 + \hat{\beta}_{k2} x_2 + \hat{\beta}_{k3} x_1^2 + \hat{\beta}_{k4} x_2^2$

the relationship is linear in the augmented input space, but the decision boundaries are quadratic in the original space

(ii) when there are two classes

$$\left.\begin{aligned} Pr[G=1|X=x] &= \frac{\exp(\beta_0 + \beta^T x)}{1+\exp(\beta_0+\beta^T x)} \\ Pr[G=2|X=x] &= \frac{1}{1+\exp(\beta_0+\beta^T x)} \end{aligned}\right\} \text{logit transformation } \log\frac{p}{1-p}$$

$$\log \frac{Pr[G=1|X=x]}{Pr[G=2|X=x]} = \beta_0 + \beta^T x$$

Linear regression of an Indicator Matrix

- codify each of the classes $1, \dots, K$ with an indicator variable

ex. $K=3 \Rightarrow$

$$\begin{array}{c} \text{class } 1 \quad 2 \quad 3 \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array} = \underset{N \times K}{Y}$$

1^{st} obs. is of class 1

2^{nd} and 3^{rd} " 2

4^{th} " 5^{th} " 3

$\rightarrow$ linear regression: $\underset{N\times K}{\hat{Y}} = \underset{N\times p}{X} \underbrace{(\underset{p\times N}{X^T} \underset{N\times p}{X})^{-1} \underset{p\times N}{X^T} \underset{N\times K}{Y}}_{\hat{B}}$

new observation x_{new} $\quad \hat{y} = \hat{f}(x_{new}) = \underset{1\times p}{x_0} \underset{p\times K}{\hat{B}}$

$$\left(\hat{f}_1(x_{new}) \quad \hat{f}_2(x_{new}) \quad \hat{f}_3(x_{new}) \right)$$

$$\hat{G}(x_{new}) = \arg\max_k \hat{f}_k(x)$$

Why does it work

$$E[Y_k | X=x] = Pr[G=k | X=x]$$

Are $\hat{f}_k(x)$ reasonable estimates of $Pr[G=k|X=x]$?

Yes and no

→ if intercept is included $\sum_{k=1}^{K} \hat{f}_k(x) = 1$

→ $\hat{f}_k(x)$ can be <0 or >1 → problems happen when the new observation is outside the traning hull, due to the rigidity of linear regression

many times it works despite this issue

A bigger problem is the so called <u>masking effect</u>

- only if $K \geq 3$

As we saw if Fig 4.3, when $K=3$, it is sufficient to have a quadratic rule.

More generally, if we have K classes, we need a curve of degree $K-1$

LINEAR DISCRIMINANT ANALYSIS

- from the decision theory (Section 2.4) we know that for optimal classification we need to know the class posterior $\Pr(G=k \mid X=x)$

Suppose:

- $f_k(x)$ is the density of x conditional to class $G=k$
  $\Pr[X=x \mid G=k]$
- $\pi_k(x)$ is the prior probability to be in class k, $\Pr[G=k]$

Then

$$\Pr[G=k \mid X=x] = \frac{\Pr[X=x \mid G=k]\,\Pr[G=k]}{\Pr[X=x]} \qquad = \sum \Pr[X=x \mid G=k]\,\Pr[G=k]$$

$$= \frac{f_k(x)\,\pi_k(x)}{\sum_{\ell=1}^{K} f_\ell(x)\,\pi_\ell(x)}$$

We can choose $f_k(x)$ and $\pi_k(x)$ as we prefer...

- when $f_k(x)$ is from a multivariate Gaussian distribution, then Linear Discriminant Analysis (LDA)
  Quadratic Discriminant Analysis (QDA)

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right\}$$

In particular, for LDA we suppose $\Sigma_k = \Sigma \quad \forall k$

We can then compare two classes by the log-ratio

$$\log \frac{\Pr[G=k \mid X=x]}{\Pr[G=\ell \mid X=x]} = \log \frac{\cancel{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right\} \hat{\pi}_k / \cancel{D}}{\cancel{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}} \exp\left\{-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell)\right\} \hat{\pi}_\ell / \cancel{D}}$$

$$= \log \frac{\hat{\pi}_k}{\hat{\pi}_\ell} - \frac{1}{2}\Big(\cancel{x^T\Sigma^{-1}x} - \underbrace{2x^T\Sigma^{-1}\mu_k}_{\substack{\text{scalar}\\ =\mu_k^T\Sigma^{-1}x}} + \mu_k^T\Sigma^{-1}\mu_k - \cancel{x^T\Sigma^{-1}x} + 2x^T\Sigma^{-1}\mu_\ell - \mu_\ell^T\Sigma^{-1}\mu_\ell$$

$$= \log \frac{\hat{\pi}_k}{\hat{\pi}_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^T\Sigma^{-1}(\mu_k-\mu_\ell) + x^T\Sigma^{-1}(\mu_k-\mu_\ell)$$

Note

- the decision boundaries are NOT the perpendicular bisectors of the segments joining the centroids (happens only if $\Sigma = \sigma^2 I$)
- the linear discriminant function $\delta_k(x)$ is

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \hat{\pi}_k$$

$$\rightarrow \hat{G}(x) = \underset{k}{\operatorname{argmax}}\ \hat{\delta}_k(x)$$

Since we do not know the values of the parameters $\hat{\pi}_k, \mu_k, \Sigma$, we need to estimate them

- $\hat{\pi}_k = N_k / N = \dfrac{\#\text{ observations in class } k}{\text{total } \#\text{ of observations}}$
- $\hat{\mu}_k = \sum_{g_i=k} x_i / N_k$
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i-\mu_k)(x_i-\mu_k)^T / (N-K)$

With two classes, there is a simple correspondence between LDA and classification by linear regression ($\rightarrow$ Ex. 4.2)

With $K > 2$, there are substantial differences

LDA does NOT suffer from the masking effect.