Linear Classification Methods

$\mathcal{G}$ = discrete target space $\{1, \ldots, K\}$ say.

$G(x)$ = predictor (i.e. $h_\theta(x)$)

For many methods the decision boundaries are linear in input variables - these are "linear methods".

If a fit linear model on K^{th} class,

$$\hat{f}_k(x) \overset{def}{=} \theta_{k0} + \vec{\theta}_k^T \vec{x}$$

then $\{x \mid \hat{f}_k = \hat{f}_\ell\} = \{x \mid (\hat{\theta}_{k0} - \hat{\theta}_{\ell 0}) + (\vec{\theta}_k - \hat{\theta}_\ell)^T x = 0\}$

which is an affine space

i.e. the predictors are divided into regions separated by hyperplanes (a translated hyperplane).

One approach is to do linear regression on indicator matrix, namely,

$$Y_k = \begin{cases} 1 & \text{if } G = k \\ 0 & \text{else} \end{cases}$$

$\vec{Y} \in \mathbb{R}^K$ and $\underset{=}{Y} = \begin{bmatrix} - & Y^{(1)T} & - \\ - & Y^{(2)T} & - \\ & \vdots & \\ - & Y^{(m)T} & - \end{bmatrix}$

indicator response

is design matrix of zeros & ones.

Perform simple regression $\underset{=}{\hat{Y}} = \underset{=}{X}(\underset{=}{X}^T\underset{=}{X})^{-1}\underset{=}{X}^T\underset{=}{Y}$

$\hat{Y} = (p+1) \times K$

each column of $\underset{=}{Y}$ is a response vector

and each column of $\underset{=}{\hat{Y}}$ is a predicted response vector

$$\underset{=}{X} = \begin{bmatrix} 1 & - x^{(1)T} - \\ \vdots & - x^{(2)T} - \\ 1 & \vdots \\ & - x^{(m)T} - \end{bmatrix}$$

Set $\hat{\underset{=}{B}} = (\underset{=}{X}^T \underset{=}{X})^{-1} \underset{=}{X}^T \underset{=}{Y}$

Then Given a new $\vec{x}$, classify according to

(1) Compute $\hat{f}(\vec{x})^T = (1, x^T)\hat{B} \in \mathbb{R}^K$

(2) i.e. project $\vec{x}$ using $\hat{B}$

(2) Output largest component

$$\hat{G}(\vec{x}) = \underset{k \in G}{\operatorname{argmax}} \hat{f}_k(\vec{x})$$

Why? $\mathbb{E}[Y_k | X = x] = P(G = k | X = x)$

and, we saw earlier, that linear regression matches this kind of model framework.

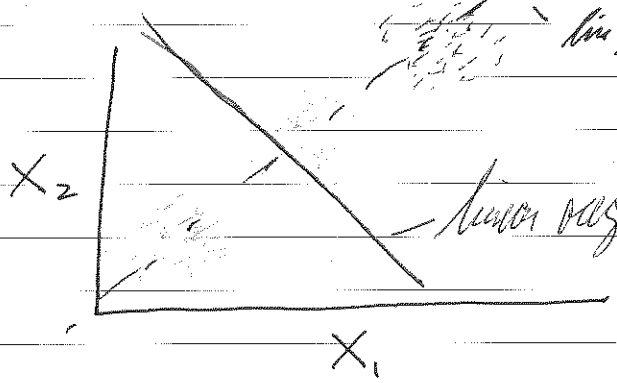
Of course $|\hat{f}_k(x)| > 1$ could happen, but that's not necessarily problematic.

Equivalently, can fit $\min_{\underset{=}{B}} \sum_i \| y_i - (1, x^{(i)T})B \|^2$

and then $\hat{G}(x) = \underset{k}{\operatorname{argmin}} \| \hat{f}(x) - \vec{e}_k \|^2$

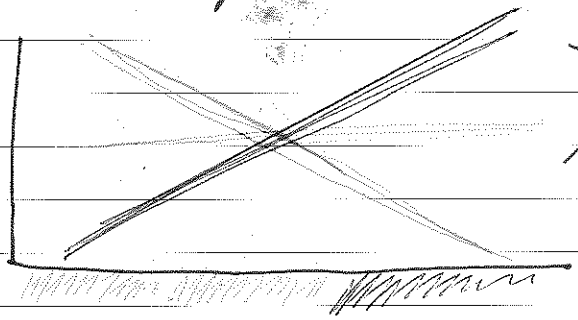
When K = # classes in logistic a subtle problem can occur.

usually classes can be "masked" by other classes.



linear regression failed to identify the middle class even though data is linearly separable.

Notice in projected space, projecting onto line through the 3 centroids of data we have:



Linear Discriminant Analysis LDA

Suppose want to know $P(G|X)$ where $f_k(x)$ is class-conditional ($f_k(x) = P(X|G)$) & π_k = class prior, $\sum \pi_k = 1$

Then $P(G=k|X=x) = \dfrac{f_k(x)\,\pi_k}{\sum_{j=1}^{K} f_j(x)\,\pi_j}$

ie. knowing $f_k(x)$ is, basically, solving the problem.

Suppose each class conditional is multivariate Gaussian

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma_k^{-1}(x-\mu)}$$

Assume a common Covariance Matrix $\Sigma_k = \Sigma \ \forall k$

then can look at the log-odds function

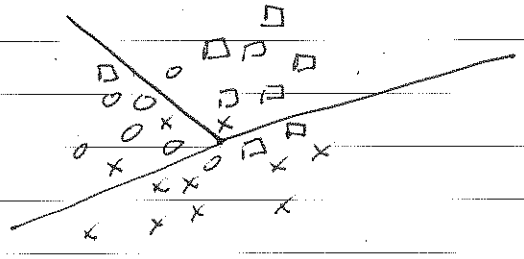
(in general $\frac{p}{1-p}$ is "odds" so $\log \frac{p}{1-p}$ is "log-odds")

$$\log \frac{P(G=k|X=x)}{P(G=\ell|X=x)} = \log \frac{\pi_k}{\pi_\ell} + \log \frac{f_k(x)}{f_\ell(x)}$$

$$= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k+\mu_\ell)^T \Sigma^{-1}(\mu_k-\mu_\ell) + x^T \Sigma^{-1}(\mu_k-\mu_\ell)$$

But then we get the decision boundary is linear

i.e. hyperplane decision boundaries.



- Decision boundaries are NOT the $\perp$ bisectors of the centroids.

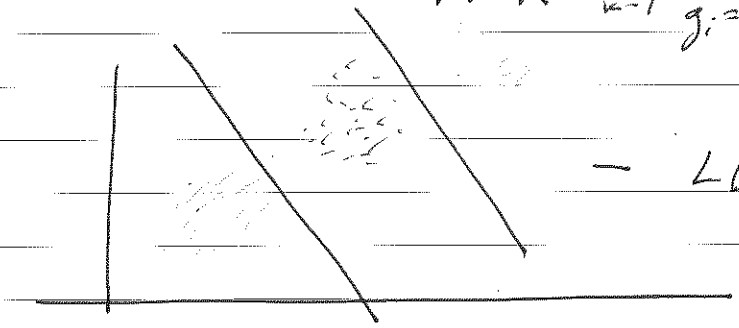
Leads to linear discriminant functions

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

So that $$G(x) = \underset{k}{\operatorname{argmax}}\ \delta_k(x)$$

Of course, need to estimate the parameters

- $\hat{\pi}_k = \frac{M_k}{M}$ = % of observations in k^{th} class.
- $\hat{\mu}_k = \frac{1}{M_k} \sum_{g_i = k} x_i$
- $\hat{\Sigma} = \frac{1}{M-K} \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$



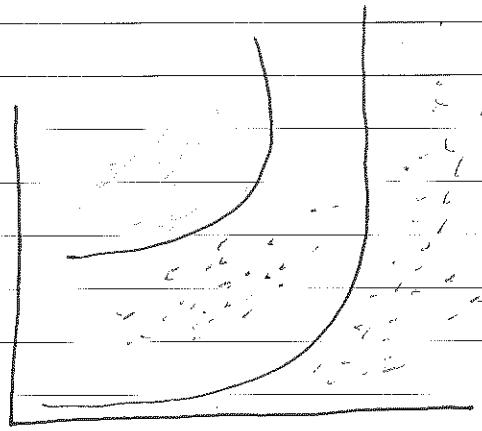
- LDA boundaries avoid masking

If not all Σ_k's are assumed same, then we don't get the cancellations from before & instead we get

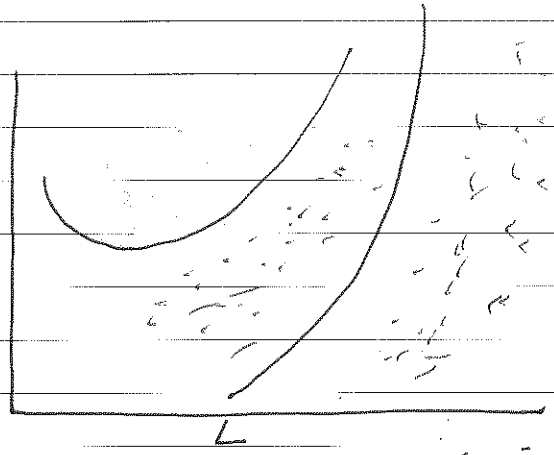
$$\delta_k(x) = \log \pi_k - \frac{1}{2} \log \det \Sigma_k - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)$$

so that $\{x \mid \delta_k(x) = \delta_\ell(x)\}$ is quadratic

is QDA quadratic discriminant analysis



QDA decision boundaries



LDA boundaries in 5-dimensional $(X_1, X_2, X_1^2, X_2^2, X_1X_2)$ space.

Very similar decision boundaries, but QDA is advised in general.

Why use LDA/QDA?

Data isn't Gaussian & it still works well.
Trade bias-variance. Accept Gaussian bias to reduce variance in more complicated models.

Can Regularize LDA/QDA via

$$\hat{\Sigma}_k^{-1}(\alpha,\gamma) = \alpha\hat{\Sigma}_k^{-1} + (1-\alpha)\hat{\Sigma}^{-1}(\gamma)$$

$$\hat{\Sigma}^{-1}(\gamma) = \gamma\hat{\Sigma}^{-1} + (1-\gamma)\hat{\sigma}^2\underline{\underline{I}} \qquad \gamma\in[0,1]$$

This gives a nice tradeoff between LDA & QDA