Lecture 11: LDA, QDA, logistic regression

For $P>1$ LDA model

$$\underset{\sim}{X} \mid Y=c \sim N(\overset{P}{\mu_c}, \overset{P\times P}{\Sigma})$$

$$P(Y=c)=\pi_c \;;\; \pi_c \geq 0,\; \sum_{c=1}^{K} \pi_c = 1$$

$$\delta_c(\underset{\sim}{x}) \equiv \mathbb{P}(\underset{\sim}{X}=\underset{\sim}{x} \mid Y=c)\,\mathbb{P}(Y=c)$$

$$= (2\pi)^{-P/2} \det(\Sigma)^{-1/2} \exp\left(-\tfrac{1}{2}(\underset{\sim}{x}-\mu_c)^T \Sigma^{-1}(\underset{\sim}{x}-\mu_c)\right)\pi_c$$

apply log

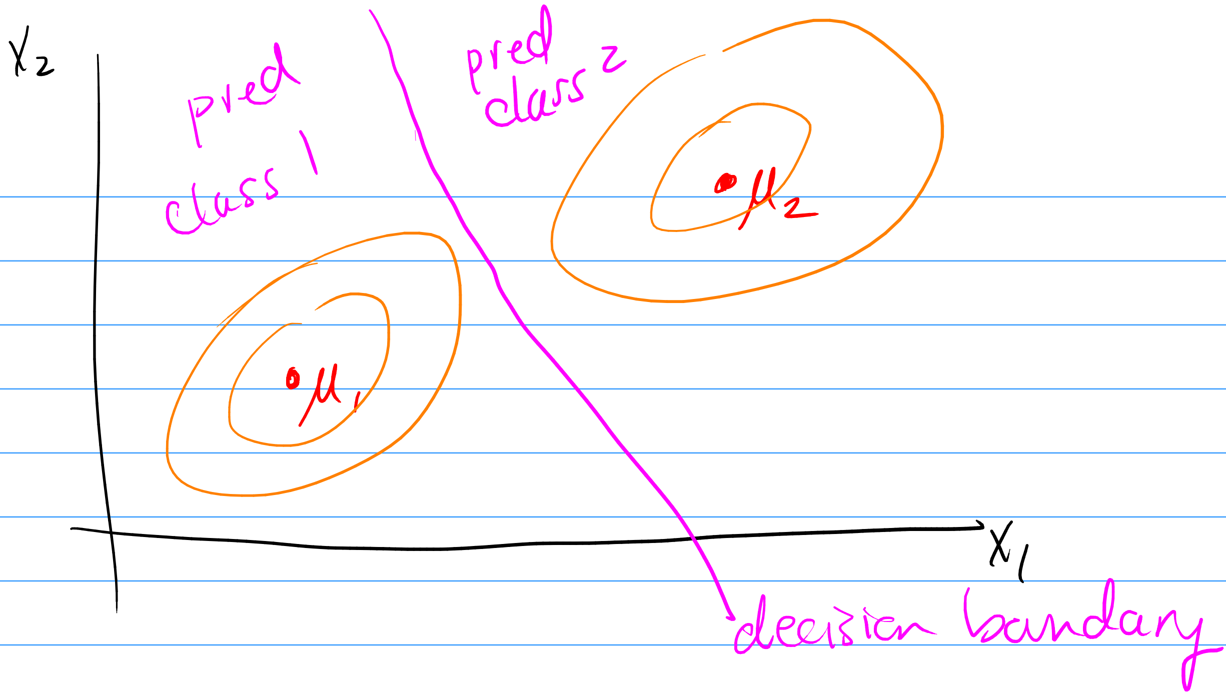
$$\equiv \cancel{-\tfrac{P}{2}\log 2\pi} \cancel{-\tfrac{1}{2}\log\det\Sigma} - \tfrac{1}{2}(\underset{\sim}{x}-\mu_c)^T \Sigma^{-1}(\underset{\sim}{x}-\mu_c) + \log \pi_c$$

$$\equiv -\tfrac{1}{2}(\underset{\sim}{x}-\mu_c)^T \Sigma^{-1}(\underset{\sim}{x}-\mu_c) + \log \pi_c$$

$$= -\tfrac{1}{2}\left(\cancel{\underset{\sim}{x}^T\Sigma^{-1}\underset{\sim}{x}} + \mu_c^T\Sigma^{-1}\mu_c - 2\mu_c^T\Sigma^{-1}\underset{\sim}{x}\right) + \log \pi_c$$

$$= \underbrace{\log \pi_c - \tfrac{1}{2}\mu_c^T\Sigma^{-1}\mu_c}_{\beta_{0,c}} + \overset{\beta_c}{\underbrace{\mu_c^T\Sigma^{-1}\underset{\sim}{x}}_{1\times P \;\; P\times P \;\; P\times 1}}$$

$$\delta_c(\underset{\sim}{x}) = \beta_{0,c} + \beta_c \underset{\sim}{x} \longleftarrow \underline{\text{linear!}}$$



To estimate μ_c, Σ, π_c for $c = 1, \ldots, K$

MLE $\begin{cases} \hat{\mu}_c = \text{mean vector of } X\text{s in class } c \\ \hat{\pi}_c = \text{\% of data in class } c \\ \hat{\Sigma} = \text{pooled cov. mtx.} \end{cases}$

Just like in regression, I can do feature engineering

$$\underset{\sim}{X} = \begin{bmatrix} | & | & | & | \\ x_1 & x_2 & x_1 x_2 & \log x_1 \\ | & | & | & | \end{bmatrix}$$

Quadratic Discriminant Analysis (QDA)

LDA: assumes that variances are equal (Σ)

QDA: relax this, assume different Σ_c for each class.

$$\delta_c(\underset{\sim}{x}) = \underbrace{\mathbb{P}(\underset{\sim}{X} = \underset{\sim}{x} \mid Y = c)}_{N(\mu_c, \Sigma_c)} \underbrace{\mathbb{P}(Y = c)}_{\pi_c}$$

Similar algebra ...

$\boxed{P=1}$

$$\delta_c(x) = -\log \sigma_c - \frac{\overset{\text{quadratic}}{(x-\mu_c)^2}}{2\sigma_c^2} + \log(\pi_c)$$

$\boxed{P>1}$

$$\delta_c(\underset{\sim}{x}) = -\log \det \Sigma_c \underbrace{- \frac{1}{2}(\underset{\sim}{x}-\mu_c)^T \Sigma_c^{-1} (\underset{\sim}{x}-\mu_c)}_{\text{quadratic}} + \log \pi_c$$

Count up number of parameters

	LDA	QDA
# params	$(K-1)(P+1)$ $\approx$ KP	$(K-1)\left(\frac{P(P+3)}{2}+1\right)$ $\approx$ KP^2

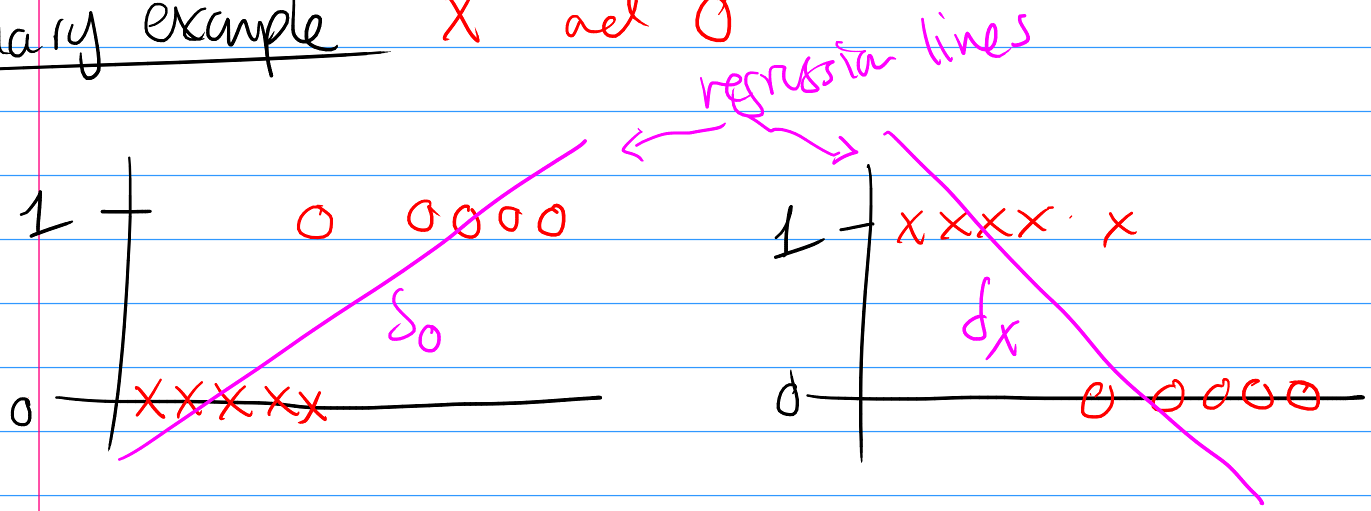
LDA is a linear classifier:

$$\delta_c(\underset{\sim}{x}) = \hat{\beta}_{0,c} + \hat{\beta}_c^T \underset{\sim}{x}$$

uses normality assumptions to calc $\hat{\beta}_{0,c}$ and $\hat{\beta}_c$ s.

Why not just use linear regression to calc $\hat{\beta}$s?

Binary example X and O



Punchline: - reasonable when K is small

(K=2 exactly gives LDA)*

- when K is large we have a potential masking problem