Lecture 4 - 17/09/24

Strategies for building a classifier

- So far our models have approximated posterior class probs:

  $P(C_k|x)$ with $C_1, C_2, \ldots, C_K$ & input x

  - This is because we have used discriminative models to classify data.
    - They model posterior class probabilities directly from data.

So far the discriminative models:

- Logistic regression

$$P(Y = C_1 | x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- K-nearest neighbours

$$P(Y = C_k | x) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = C_k).$$

Generative models for Classification

- Another approach would be to take a step back.
  - Model joint distribution of features
    - Gives also posterior probability.

$$P(y|x) = \frac{P(x,y)}{P(x)} \longrightarrow P(y=k)\,p(x|y=k)$$

$C_k = k$

- $P(y=k)$: Class prior describes prob of obtaining data point from class k → often π_k
- $p(x|y=k)$: Class conditional describes $\forall k$ the distribution of features within k

- This generative model will then give us the posterior class distribution that we use for classification in discriminative models
- Generative model also gives estimates of joint dist. → $p(x,y)$ can be used for simulating data

Prob of obtain a point in this class

Distribution of the class

- There are then the strategies:

① Model the full joint distribution (Generative models)
   - Gives posterior probabilities

② Model class posterior probabilities (Discriminative model)
   - Posterior probs used as discriminant functions

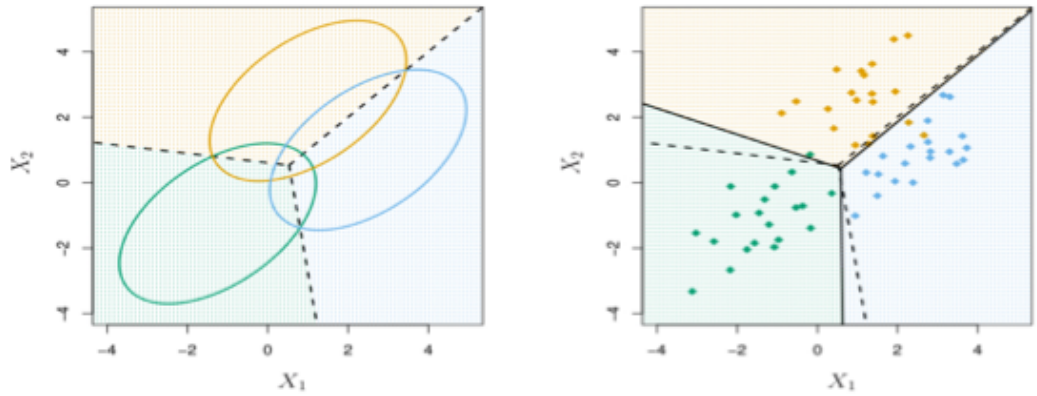
③ Model directly a discriminant function
   - Remember $d(x)$ → maps input into prediction

Hierarchy!!!

Number ② refers to ⟶ Logistic Regression, K-nearest neighbors.

Todays lecture → ①

Linear Discriminant Analysis



So... What do we do in LDA and how do we do it?

- To build our classifier we will follow the hierarchy described before
- For Linear Discriminant Analysis:
  - we model the joint
  - End up with discriminant functions, $g_k(x)$
- For LDA we assume that class conditionals are:
  - Gaussian with a class-specific mean, common var

Few points...

- Want to classify x with highest discriminant $g_k(x)$
- For Bayes:
  - Choose posterior probs ||
  - joint dist. ||
  - Log scale

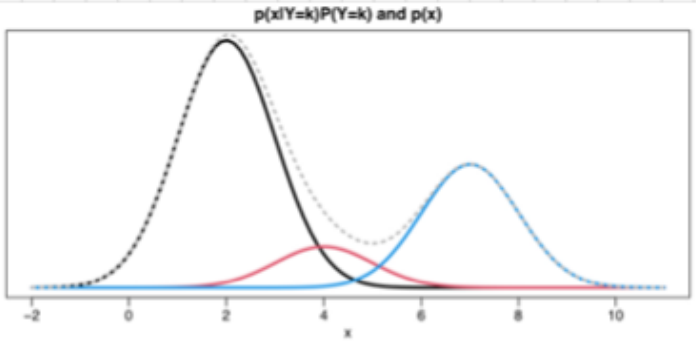
Let's build Bayes classifier:

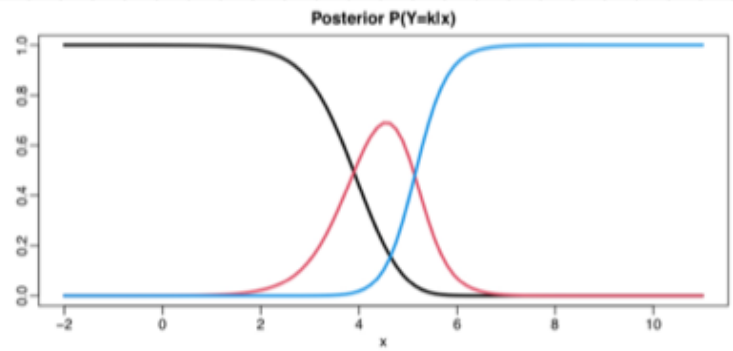
- We can derive it under the following assumption:
  - we know the true distribution of data.
- Gaussian class conditionals:
  - we assume feature x to have a univariate Gaussian distribution:

$p(x|y = \text{BLACK}) = \mathcal{N}(2,1) \longrightarrow \pi_{\text{BLACK}} = 0.6$

$p(x|y = \text{RED}) = \mathcal{N}(4,1) \longrightarrow \pi_{\text{RED}} = 0.1$

$p(x|y = \text{BLUE}) = \mathcal{N}(9,1) \longrightarrow \pi_{\text{BLUE}} = 0.3$





Discriminant for LDA with one feature:

- Look at the log of the joint distribution,

$$\log p(x,y) = \log P(y=k) + \log p(x|y=k)$$

$$= \log \pi_k + \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right\} \text{ (Our gaussian assumption)}$$

$$= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$$

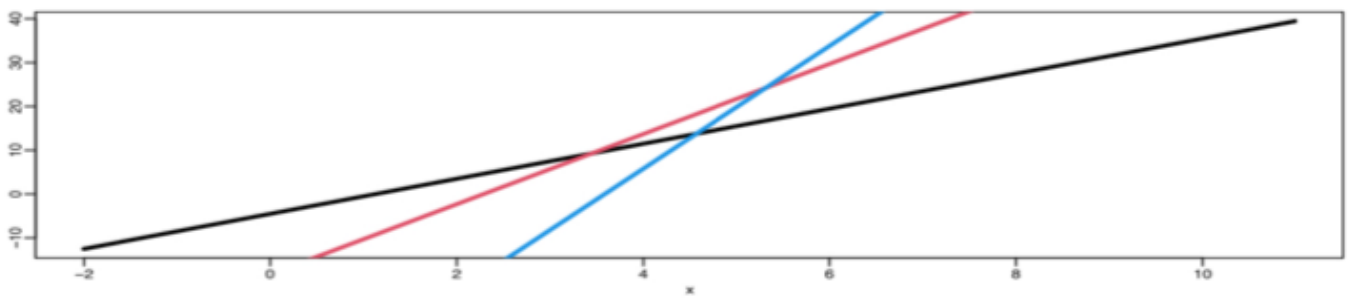
$$= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x-\mu_k)^2}{2\sigma^2}$$

$$= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2 - 2\mu_k x + \mu_k^2}{2\sigma^2}$$

$$= \log \pi_k + \underbrace{\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2}}_{\text{same for all k}} + \frac{2\mu_k x}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

- This means that the Bayes classifier chooses the class with the highest

$$g_k(x) = \frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \longrightarrow \text{linear in } x$$



(Note that we modelled the joint distribution as Gaussian (class-specific mean, common variance). This makes this a generative model).

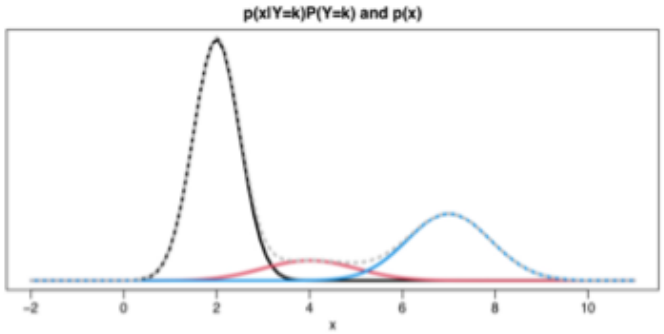
Quadratic Discriminant Analysis (QDA)

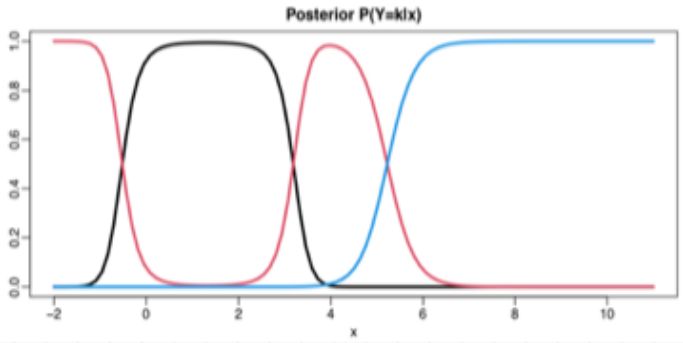
- The assumption of a common variance in LDA may be too strict
- In QDA we assume class conditionals are:
  - Gaussian with both class-specific means & class-specific variance
- QDA introduce class-specific variances:

$$p(x|y=\text{black}) = N(2, 0.25) \longrightarrow \pi_{\text{black}} = 0.6$$

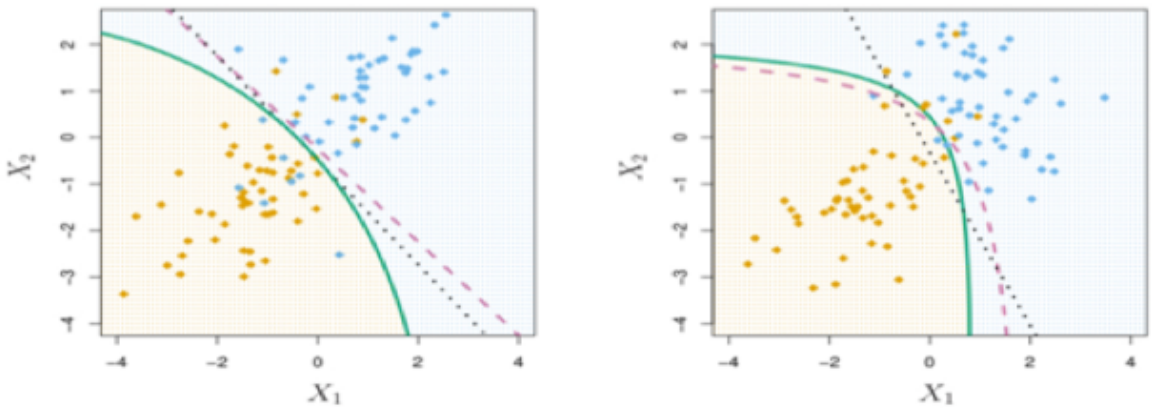
$$p(x|y=\text{red}) = N(4, 1) \longrightarrow \pi_{\text{red}} = 0.1$$

$$p(x|y=\text{blue}) = N(7, 0.81) \longrightarrow \pi_{\text{blue}} = 0.3$$





LDA VS QDA



Purple: Bayes boundary, black: LDA, green: QDA.

QDA gives more flexibility (variance!). LDA vs QDA is a bias-variance tradeoff.

- LDA & QDA ARE PLUG-IN CLASSIFIERS
- THE "GOAL STANDARD" BAYES CLASSIFIER IS DERIVED:
  - By ASSUMING CLASS CONDITIONALS ARE GAUSSIAN
  - By ASSUMING WE KNOW ALL THEIR PARAMS (MEAN, VARIANCE)
- IN PRACTICE THOUGH:
  - CLASS CONDITIONALS ARE NOT TRULY GAUSSIAN
  - VARIANCE & MEAN ARE UNKNOWN.
  - TRUE CLASS PRIORS ARE UNKNOWN.

LDA & QDA APPROXIMATE THE BAYES CLASSIFIER

USUALLY WE DO NOT KNOW MODEL PARAMETERS:

- FOR CLASS PRIORS $\hat{\pi}_k$ WE USE MLE, THE EMPIRICAL FREQUENCIES IN TRAINING SET:

$$\hat{\pi}_k = \frac{n_k}{n}$$ ← # OF OBSERVATIONS W k

- FOR MEANS WE TYPICALLY USE MAX LIKELIHOOD ESTIMATES:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

- FOR VARIANCES WE HAVE DIFFERENT METHODS FOR LDA & QDA:

LDA:

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

QDA:

$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$