

# Strategies for building a classifier

# Discriminative models for classification

So far, our classification models have approximated posterior class probabilities

$$P(C_k | x),$$

in classification problems with classes  $C_1, C_2, \dots, C_K$ , and input  $x$ .

This is because we have used *discriminative models* to classify data.

# Discriminative models for classification

So far, our classification models have approximated posterior class probabilities

$$P(\mathcal{C}_k | x),$$

in classification problems with classes  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_K$ , and input  $x$ .

This is because we have used *discriminative models* to classify data.

Discriminative models model posterior class probabilities directly from data.

With posterior class probabilities,  $P(Y = \mathcal{C}_k | x)$ , we can build a good classifier (one minimising posterior expected loss, thus also expected loss)

# Discriminative models for classification

Discriminative models so far:

- Logistic regression

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

- K-nearest neighbours

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

In the following, we will use the notation  $C_k = k$ .

# Generative models for classification

Another approach is to take one step further back and model the joint distribution of features, which gives also posterior probabilities:

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

# Generative models for classification

Another approach is to take one step further back and model the joint distribution of features, which gives also posterior probabilities:

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

The joint distribution of features  $X$  and class  $Y$  can be specified in two parts

$$p(x,y) = P(Y = k)p(x|Y = k)$$

$P(Y = k)$ : The class prior describes the probability of obtaining a data point from class  $k$  (irrespective of feature value)  
Class priors are often denoted  $\pi_k$ .

$p(x|Y = k)$ : The class conditional describes for each class  $k$  the distribution of features for data within class  $k$ .

# Generative models for classification

Another approach is to take one step further back and model the joint distribution of features, which gives also posterior probabilities:

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

The joint distribution of features  $X$  and class  $Y$  can be specified in two parts

$$p(x,y) = P(Y = k)p(x|Y = k)$$

So... This approach, *Generative Models*, will eventually give us the posterior class distributions which we use for classification in discriminative models.

But... Generative models also give estimates of the joint distribution  $p(x,y)$ , which can be used for, e.g. simulating data.

# Strategies for building a classifier

We can characterise three distinct strategies for building a classifier:

1. Model the full joint distribution (generative models)

The joint distribution gives posterior probabilities

2. Model class posterior probabilities (discriminative model)

Posterior probabilities can be used as discriminant functions

3. Model directly a *discriminant function*, which is the function  $d(x)$  that maps inputs into predictions.



# Strategies for building a classifier

We can characterise three distinct strategies for building a classifier:

1. Model the full joint distribution (generative models)

The joint distribution gives posterior probabilities

2. Model class posterior probabilities (discriminative model)

Posterior probabilities can be used as discriminant functions

3. Model directly a *discriminant function*, which is the function  $d(x)$  that maps inputs into predictions.

Note the hierarchy of strategies.

# Strategies for building a classifier

We can characterise three distinct strategies for building a classifier:

1. Model the full joint distribution (generative models)

The joint distribution gives posterior probabilities

- ▶ Today: Linear Discriminant Analysis (LDA)
- ▶ Today: Quadratic Discriminant Analysis (QDA)

2. Model class posterior probabilities (discriminative model)

Posterior probabilities can be used as discriminant functions

- ▶ Logistic Regression, K-nearest neighbors

3. Model directly a *discriminant function*, which is the function  $d(x)$  that maps inputs into predictions.

Note the hierarchy of strategies.

# Linear Discriminant Analysis (LDA)

## THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

BY R. A. FISHER, Sc.D., F.R.S.

### I. DISCRIMINANT FUNCTIONS

WHEN two or more populations have been measured in several characters,  $x_1, \dots, x_s$ , special interest attaches to certain linear functions of the measurements by which the populations are best discriminated. At the author's suggestion use has already been made of this fact in craniometry (*a*) by Mr E. S. Martin, who has applied the principle to the sex differences in measurements of the mandible, and (*b*) by Miss Mildred Barnard, who showed how to obtain from a series of dated series the particular compound of cranial measurements showing most distinctly a progressive or secular trend. In the present paper the application of the same principle will be illustrated on a taxonomic problem; some questions connected with the precision of the processes employed will also be discussed.




# Linear Discriminant Analysis (LDA)

## THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

By R. A. FISHER, Sc.D., F.R.S.

### I. DISCRIMINANT FUNCTIONS

WHEN two or more populations have been measured in several characters,  $x_1, \dots, x_s$ , special interest attaches to certain linear functions of the measurements by which the populations are best discriminated. At the author's suggestion use has already been made of this fact in craniometry (a) by Mr E. S. Martin, who has applied the principle to the sex differences in measurements of the mandible, and (b) by Miss Mildred Barnard, who showed how to obtain from a series of dated series the particular compound of cranial measurements showing most distinctly a progressive or secular trend. In the present paper the application of the same principle will be illustrated on a taxonomic problem; some questions connected with the precision of the processes employed will also be discussed.

2. <sup>a</sup> <sup>b</sup> Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems"  (PDF). *Annals of Eugenics*. 7 (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x . hdl:2440/15227 .

# Linear Discriminant Analysis (LDA)

## THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

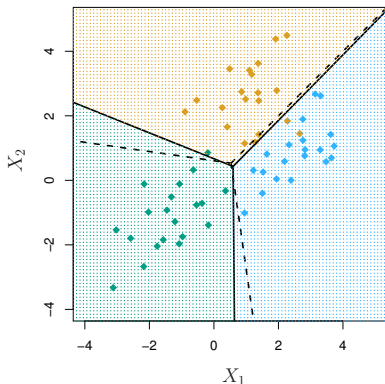
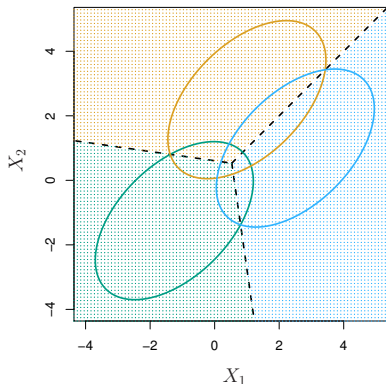
BY R. A. FISHER, Sc.D., F.R.S.

### I. DISCRIMINANT FUNCTIONS

WHEN two or more populations have been measured in several characters,  $x_1, \dots, x_s$ , special interest attaches to certain linear functions of the measurements by which the populations are best discriminated. At the author's suggestion use has already been made of this fact in craniometry (*a*) by Mr E. S. Martin, who has applied the principle to the sex differences in measurements of the mandible, and (*b*) by Miss Mildred Barnard, who showed how to obtain from a series of dated series the particular compound of cranial measurements showing most distinctly a progressive or secular trend. In the present paper the application of the same principle will be illustrated on a taxonomic problem; some questions connected with the precision of the processes employed will also be discussed.

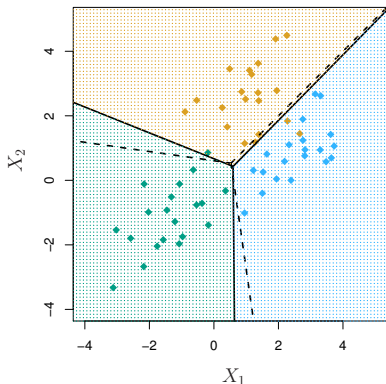
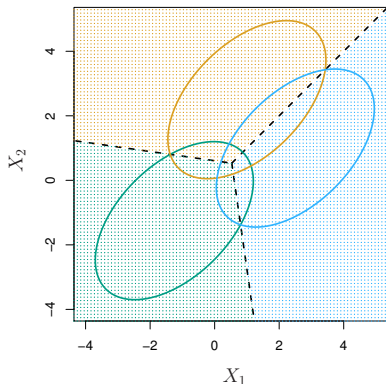
# Linear Discriminant Analysis (LDA)

Generative model that gives linear decision boundaries in classification problems.



# Linear Discriminant Analysis (LDA)

Generative model that gives linear decision boundaries in classification problems.



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

# Strategies for building a classifier

Remember the hierarchy of strategies.

We can characterise three distinct strategies for building a classifier:

1. Model the full joint distribution (generative models)  
The joint distribution gives posterior probabilities
2. Model class posterior probabilities (discriminative model)  
Posterior probabilities can be used as discriminant functions
3. Model directly a *discriminant function*, which is the function  $d(x)$  that maps inputs into predictions.
  - A  $K$ -class discriminant comprises of  $K$  discriminant functions  $g_1(x), \dots, g_K(x)$ ; we classify to the class as,

$$d(x) = \arg \max_k g_k(x).$$



# Building discriminants

For Linear Discriminant Analysis, we will model the joint distribution  $p(x, y)$ , and end up with discriminant functions  $g_k(x)$ .

A few points on the discriminant functions. . .

## Building discriminants

Want to classify to the class  $k$  with the highest discriminant  $g_k(x)$ .

For the Bayes classifier, we could choose

## Building discriminants

Want to classify to the class  $k$  with the highest discriminant  $g_k(x)$ .

For the Bayes classifier, we could choose  
... the posterior probabilities

$$g_k(x) = P(Y = k | x)$$

## Building discriminants

Want to classify to the class  $k$  with the highest discriminant  $g_k(x)$ .

For the Bayes classifier, we could choose  
... the posterior probabilities

$$g_k(x) = P(Y = k | x)$$

... or the joint distribution

$$g_k(x) = P(Y = k)p(x | Y = k) = P(Y = k | x)p(x)$$

## Building discriminants

Want to classify to the class  $k$  with the highest discriminant  $g_k(x)$ .

For the Bayes classifier, we could choose  
... the posterior probabilities

$$g_k(x) = P(Y = k | x)$$

... or the joint distribution

$$g_k(x) = P(Y = k)p(x | Y = k) = P(Y = k | x)p(x)$$

... or go to log scale

$$g_k(x) = \log P(Y = k | x)$$

Anything that ensures that  $\arg \max_{k=1,\dots,K} g_k(x)$  stays the same.

For Linear Discriminant Analysis, the log picture will be convenient.

# LDA: A generative model for classification

In *Linear discriminant analysis (LDA)*, we assume that the class conditionals are

Gaussian with a class-specific mean and a *common variance*.

# Let us build Bayes classifier based on this model

We can derive the optimal solution (Bayes classifier) under the assumption that we know the true distribution of data.

For LDA that means:

- Class conditionals are Gaussian
- We know all parameters (mean, variance) needed to fully specify the Gaussian distributions.
- We know the class priors.

## Gaussian class conditionals: Class-specific mean, but common variance

Conditionally on the class we assume the feature  $x$  to have a univariate Gaussian distribution as

$$p(x | Y = \text{black}) = \mathcal{N}(2, 1)$$

$$p(x | Y = \text{red}) = \mathcal{N}(4, 1)$$

$$p(x | Y = \text{blue}) = \mathcal{N}(7, 1)$$

The class probabilities we take to be

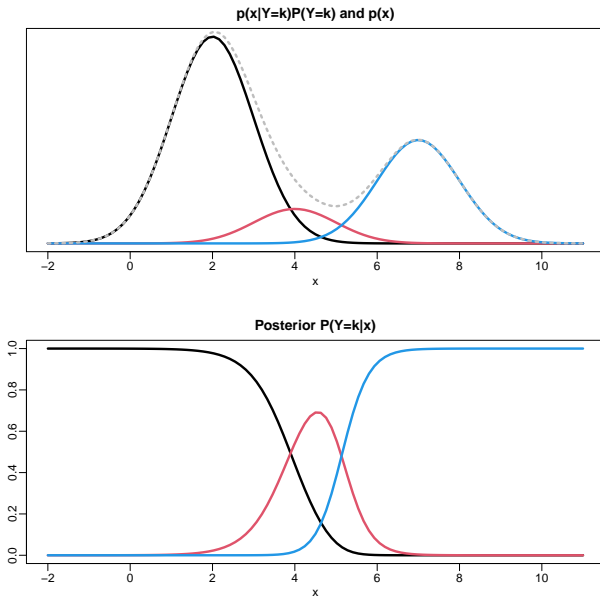
$$\pi_{\text{black}} = 0.6$$

$$\pi_{\text{red}} = 0.1$$

$$\pi_{\text{blue}} = 0.3$$



# Gaussian class conditionals, common variance.



# Discriminant for LDA with one feature

Look at the log of the joint distribution,

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

# Discriminant for LDA with one feature

Look at the log of the joint distribution,

$$\begin{aligned}\log p(x, k) &= \log P(Y = k) + \log p(x | Y = k) \\ &= \log \pi_k + \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right\} \quad (\text{Our gaussian assumption})\end{aligned}$$

# Discriminant for LDA with one feature

Look at the log of the joint distribution,

$$\begin{aligned}\log p(x, k) &= \log P(Y = k) + \log p(x | Y = k) \\ &= \log \pi_k + \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right\} \quad (\text{Our gaussian assumption}) \\ &= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}\end{aligned}$$

# Discriminant for LDA with one feature

Look at the log of the joint distribution,

$$\begin{aligned}\log p(x, k) &= \log P(Y = k) + \log p(x | Y = k) \\&= \log \pi_k + \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right\} \text{ (Our gaussian assumption)} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x - \mu_k)^2}{2\sigma^2}\end{aligned}$$

# Discriminant for LDA with one feature

Look at the log of the joint distribution,

$$\begin{aligned}\log p(x, k) &= \log P(Y = k) + \log p(x | Y = k) \\&= \log \pi_k + \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right\} \quad (\text{Our gaussian assumption}) \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x - \mu_k)^2}{2\sigma^2} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2 - 2\mu_k x + \mu_k^2}{2\sigma^2}\end{aligned}$$

# Discriminant for LDA with one feature

Look at the log of the joint distribution,

$$\begin{aligned}\log p(x, k) &= \log P(Y = k) + \log p(x | Y = k) \\&= \log \pi_k + \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right\} \quad (\text{Our gaussian assumption}) \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x - \mu_k)^2}{2\sigma^2} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2 - 2\mu_k x + \mu_k^2}{2\sigma^2} \\&= \log \pi_k + \underbrace{\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2}}_{\text{same for all k}} + \frac{2\mu_k x}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}\end{aligned}$$

# Discriminant for LDA with one feature

Look at the log of the joint distribution,

$$\begin{aligned}\log p(x, k) &= \log P(Y = k) + \log p(x | Y = k) \\&= \log \pi_k + \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right\} \quad (\text{Our gaussian assumption}) \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x - \mu_k)^2}{2\sigma^2} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2 - 2\mu_k x + \mu_k^2}{2\sigma^2} \\&= \log \pi_k + \underbrace{\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2}}_{\text{same for all } k} + \frac{2\mu_k x}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}\end{aligned}$$

Thus Bayes classifier chooses the class with the highest

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



# Discriminant for LDA with one feature

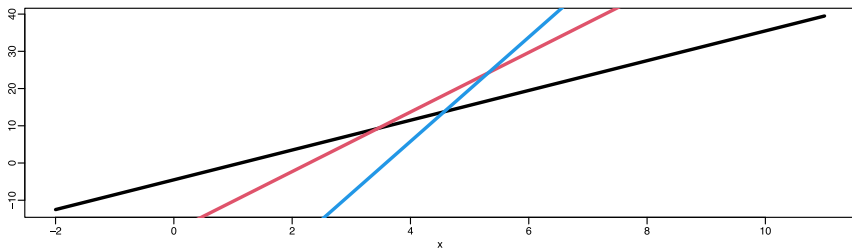
Look at the log of the joint distribution,

$$\begin{aligned}\log p(x, k) &= \log P(Y = k) + \log p(x | Y = k) \\&= \log \pi_k + \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right\} \quad (\text{Our gaussian assumption}) \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x - \mu_k)^2}{2\sigma^2} \\&= \log \pi_k + \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2 - 2\mu_k x + \mu_k^2}{2\sigma^2} \\&= \log \pi_k + \underbrace{\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2}}_{\text{same for all } k} + \frac{2\mu_k x}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}\end{aligned}$$

Thus Bayes classifier chooses the class with the highest

$$g_k(x) = \frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \quad (\text{LINEAR in } x \text{ (hence name)})$$

# The Linear Discriminant Functions



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

# Quadratic Discriminant Analysis

The assumption of a common variance for the class conditionals in LDA may be too strict.

In *quadratic discriminant analysis (QDA)* we assume that the class conditionals are **Gaussian with both class-specific means and class-specific variance**.

## QDA Example: Introducing class-specific variances

Conditionally on the class we assume the feature  $x$  to have a univariate Gaussian distribution as

$$p(x | Y = \text{black}) = \mathcal{N}(2, 0.25)$$

$$p(x | Y = \text{red}) = \mathcal{N}(4, 1)$$

$$p(x | Y = \text{blue}) = \mathcal{N}(7, 0.81)$$

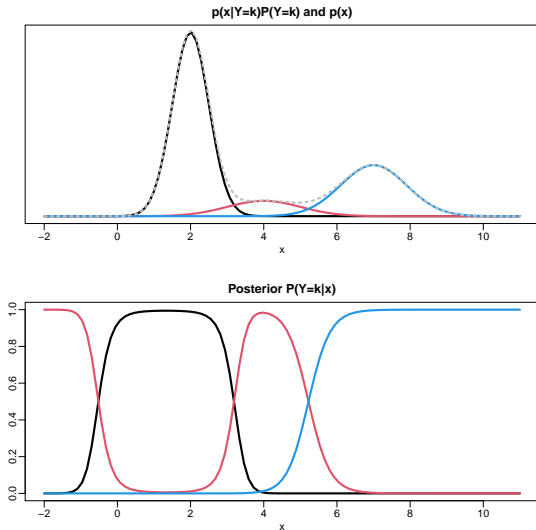
The class probabilities we take to be

$$\pi_{\text{black}} = 0.6$$

$$\pi_{\text{red}} = 0.1$$

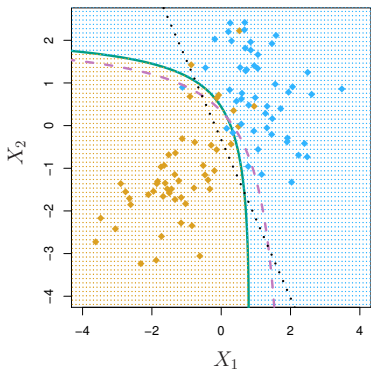
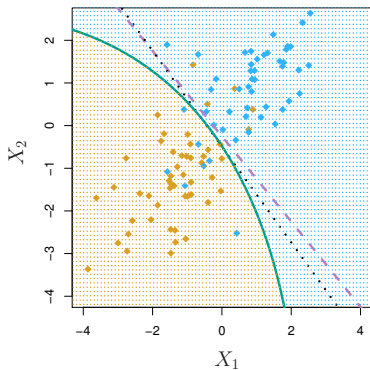
$$\pi_{\text{blue}} = 0.3$$

# Gaussian class conditionals, class-specific variances.



Central observation: Greater variance of red results in highest posterior in **2** intervals. QDA discriminant function can pick this up.

## Taste: LDA vs QDA in several dimensions



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

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

## LDA and QDA are plug-in classifiers

The “gold standard” Bayes classifier is derived assuming not only that class conditionals are Gaussian, but also that we know all their parameters (mean, variance) as well as the class prior probabilities.

In practice, we will introduce approximations at various levels:

- Perhaps class conditionals are not truly Gaussian.
- Perhaps we know that the Gaussian model is correct, but its model parameters – variance and mean – are unknown.
- Perhaps the true class priors are unknown.

LDA and QDA approximate the Bayes classifier.

## Usually we do not know model parameters

For **class probabilities**  $\hat{\pi}_k$  we use typically the MLE which are the empirical frequencies in the training set

$$\hat{\pi}_k = \frac{n_k}{n}$$

where  $n_k$  is the number of observations in class  $k$ .

The probabilities need not come from training data  
— perhaps you know that test data has a different class distribution than your training data?



# Means for the Gaussian class conditionals

We typically use the maximum likelihood estimates – the sample mean within each class

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

## Variances for the Gaussian class conditionals

For LDA we use a sample variance reflecting that observations vary around class means  $\mu_k$ , but that the variation is the same within each class.

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

For QDA we use the sample variance within each class  $k$ ,

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

Each variance is based only on the  $n_k$  observations in class  $k$ .