

DATASCI 347 Machine Learning

Lecture 4: LDA and QDA

Ruoxuan Xiong

Suggested reading: ISL Chapter 4

Lecture plan

- LDA and QDA

Discriminative vs generative methods

- Discriminative methods
 - Directly model $P(Y = k|X = x)$ and classify
 - E.g., logistic regression, KNN
 - E.g., $\log \left[\frac{P(Y = 1|X)}{P(Y = 0|X)} \right] = \beta' \tilde{X} = \beta_0 + \beta_1 \cdot \text{student} + \beta_2 \cdot \text{balance} + \beta_3 \cdot \text{income}$
- Generative methods
 1. Model the joint probability $P(X = x, Y = k) = P(Y = k)P(X = x | Y = k)$
 2. Estimate class probability $P(Y = k)$
 3. Estimate conditional distribution of X given $Y = k$, $P(X = x|Y = k)$
 4. Bayes theorem is applied to obtain $P(Y = k|X = x)$ and classify
 - E.g., linear discriminant analysis (LDA), quadratic discriminant analysis (QDA)

Bayes theorem

- Given
 - Conditional distribution of X given $Y = k$: $P(X = x | Y = k)$
 - The **prior** probabilities for each possible class k : $P(Y = k)$
- Bayes theorem to obtain $P(Y = k | X = x)$

$$\begin{aligned} P(Y = k | X = x) &= \frac{P(Y = k, X = x)}{P(X = x)} \\ &= \frac{P(X=x|Y=k)P(Y=k)}{\sum_j P(X=x|Y=j)P(Y=j)} \end{aligned}$$

Law of total probability: $P(X = x) = \sum_j P(X = x, Y = j)$

Example: An iris data set

- Perhaps the best known database in the pattern recognition literature
- Predict class of iris plant
- There are three classes

Collected by Ronald
Fisher (1936)



Iris Versicolor



Iris Setosa



Iris Virginica

Sepal and petal of iris

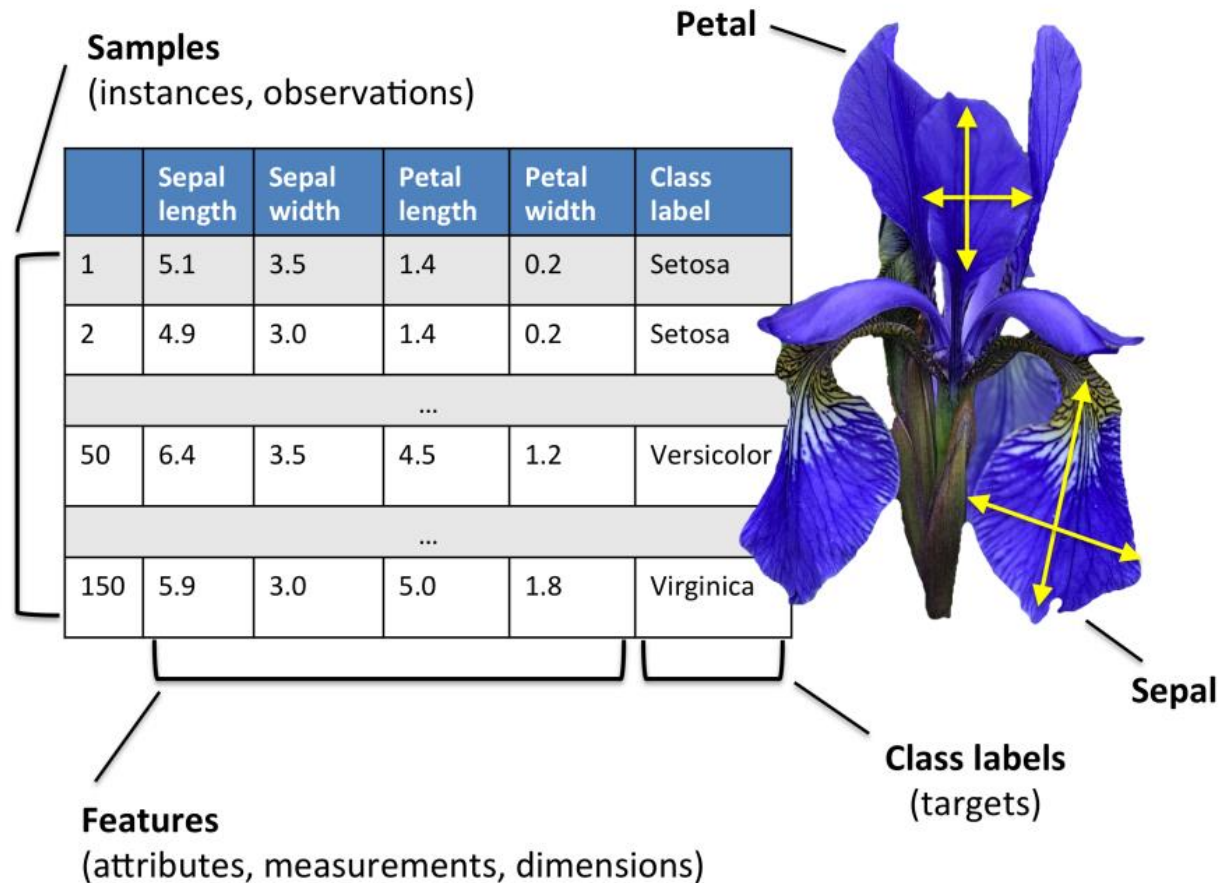
Sepal: outermost layer
of a flower



Petal: surrounds the plant's
reproductive organs

A data snapshot

- 50 samples from each of three class of *Iris* (*versicolor*, *setosa*, *virginica*)
- Four features: sepal length, sepal width, petal length, petal width



Estimating $\pi_k = P(Y = k)$

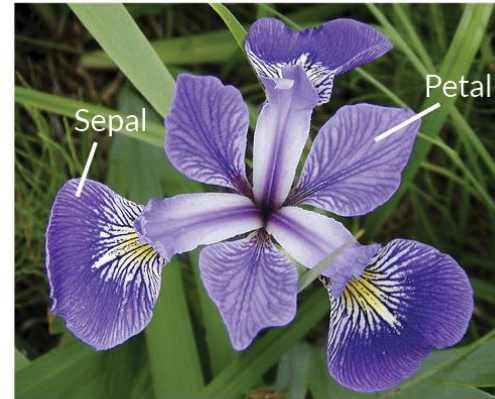
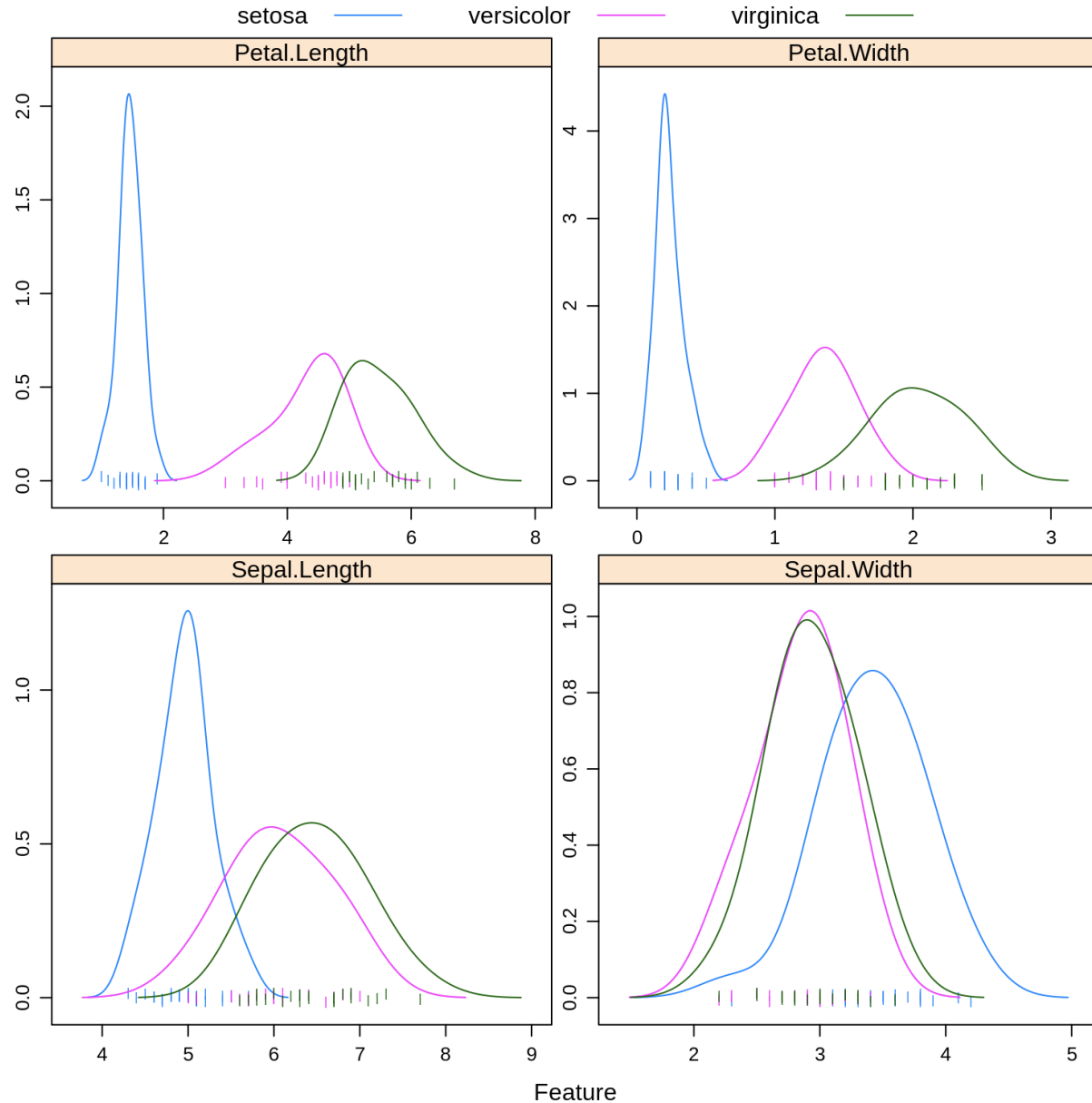
- The fraction of training samples of class k

$$\hat{\pi}_k = \hat{P}(Y = k) = \frac{\#\{i: y_i = k\}}{n}$$

- Iris data: 50 samples from each of three class of *Iris* (*versicolor*, *setosa*, *virginica*). Then

$$\hat{\pi}_{setosa} = \hat{\pi}_{versicolor} = \hat{\pi}_{virginica} = \frac{50}{50 + 50 + 50} = \frac{1}{3}$$

Distribution of features $P(X = x \mid Y = k)$



Iris Versicolor



Iris Setosa



Iris Virginica

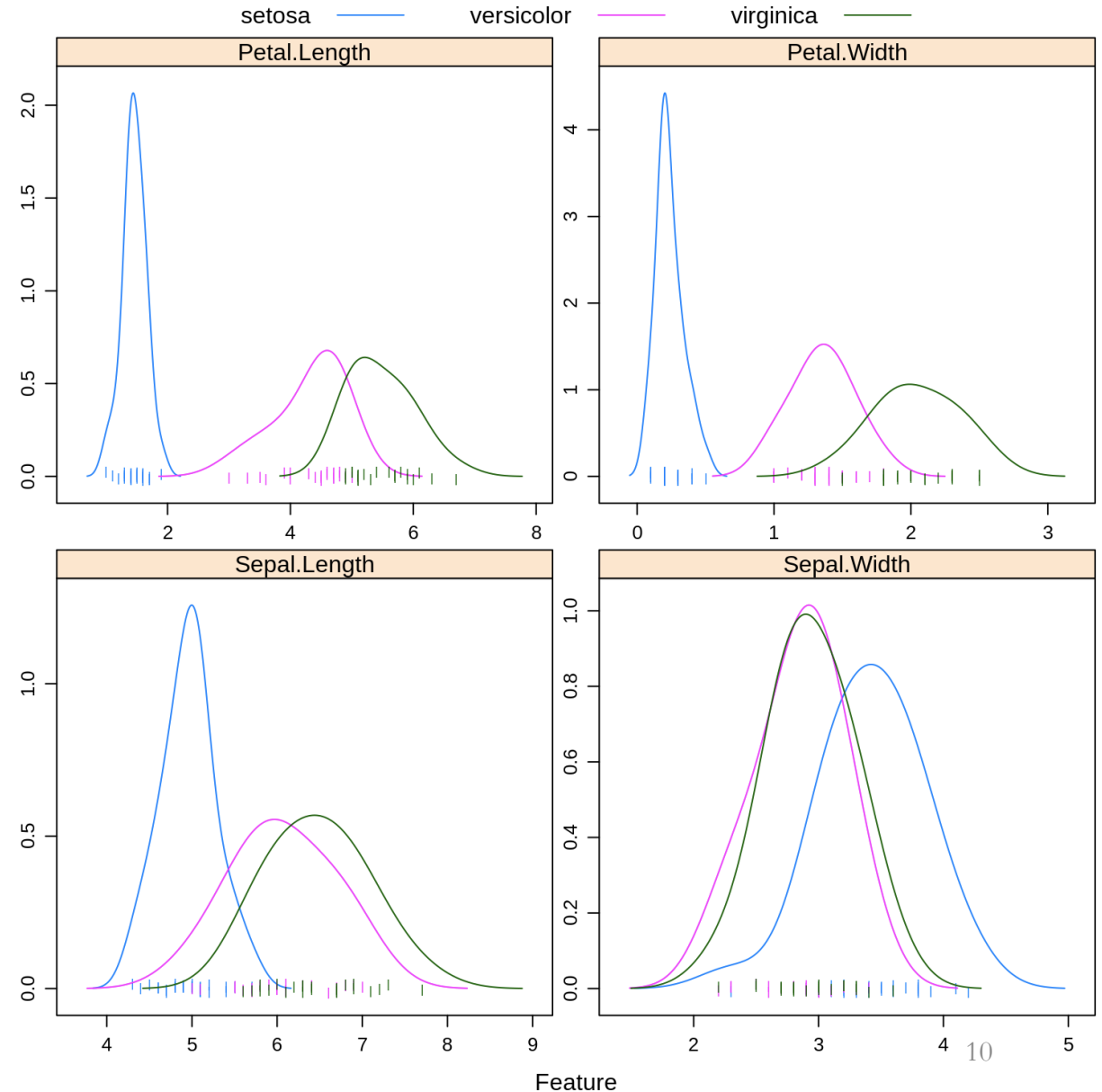
Linear discriminant analysis

- Model $P(X = x \mid Y = k)$

- $X = \begin{bmatrix} \text{sepal length} \\ \text{sepal width} \\ \text{petal length} \\ \text{petal width} \end{bmatrix}$

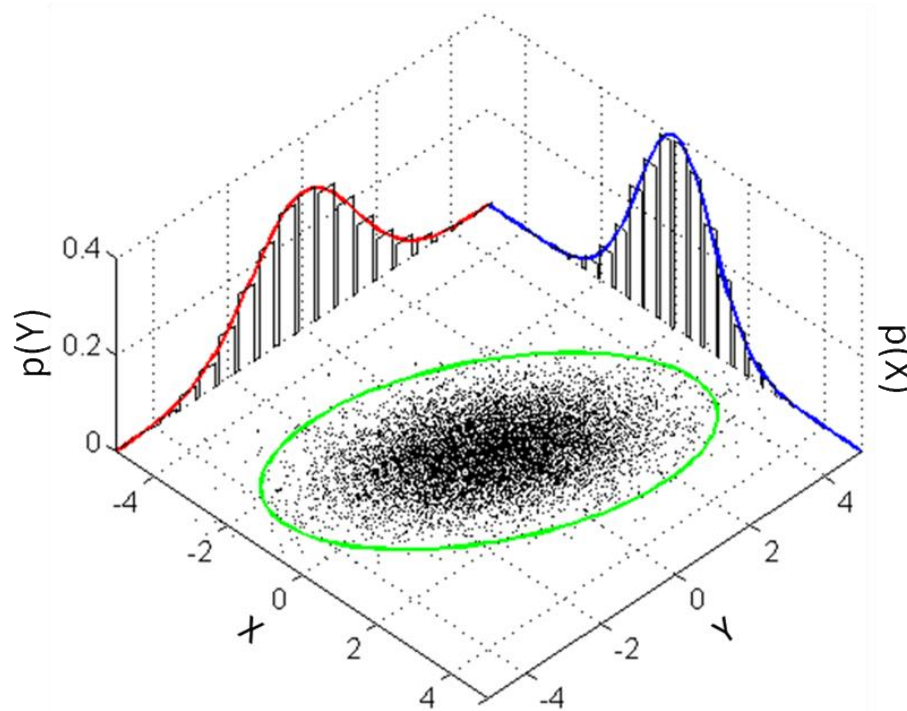
- $Y \in \{\text{versicolor}, \text{setosa}, \text{virginica}\}$

by a *Multivariate Normal Distribution* $N(\mu_k, \Sigma)$
with mean μ_k , covariance matrix Σ



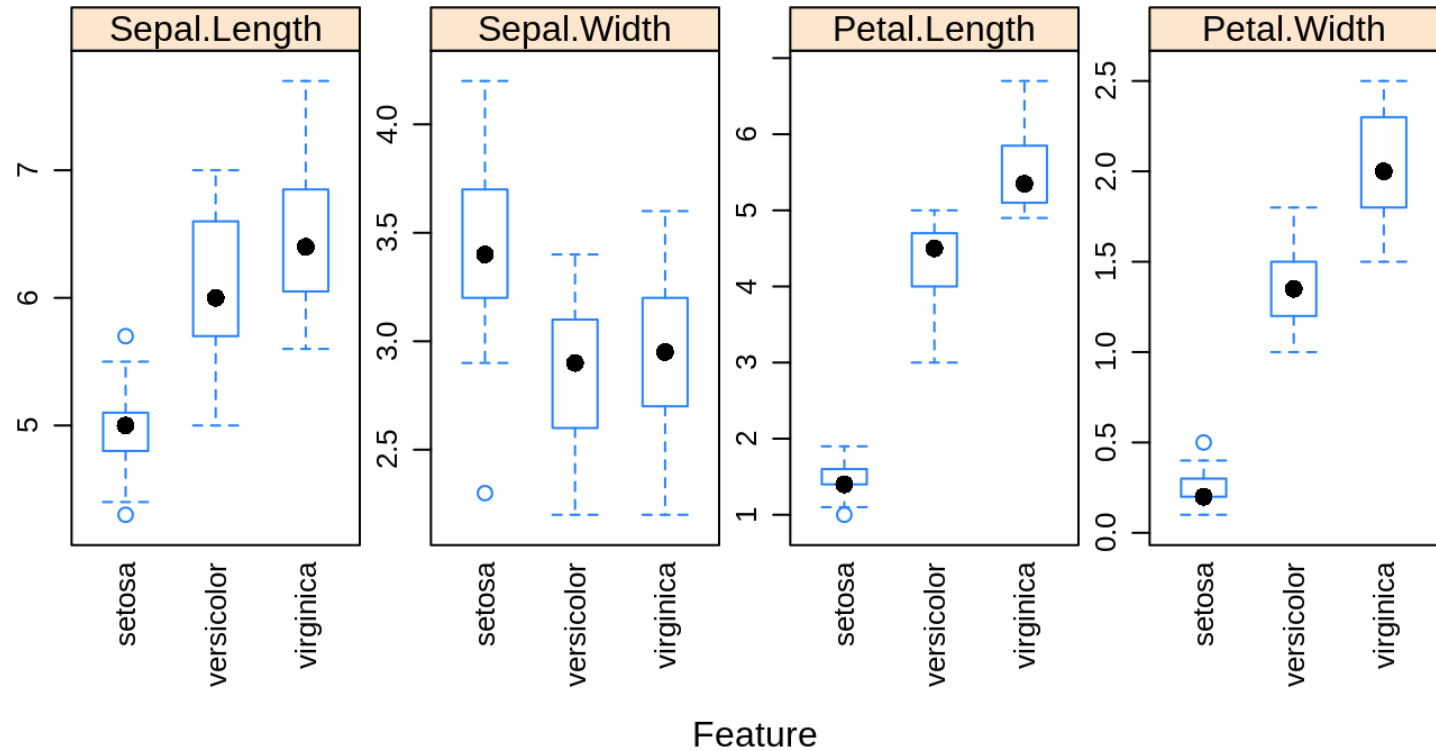
Multivariate normal distribution

- Illustration of a **two-dimensional multivariate normal distribution**
 - Projection to every dimension (**blue** and **red**) is still a Gaussian
 - Question: What are the means of X and Y ? Is the correlation between X and Y positive or negative?



μ_k in linear discriminant analysis

- $\mu_{setosa} = \begin{bmatrix} \text{setosa sepal length} \\ \text{setosa sepal width} \\ \text{setosa petal length} \\ \text{setosa petal width} \end{bmatrix}$
- Bar represents average value
- Black dots of setosa in the box plots

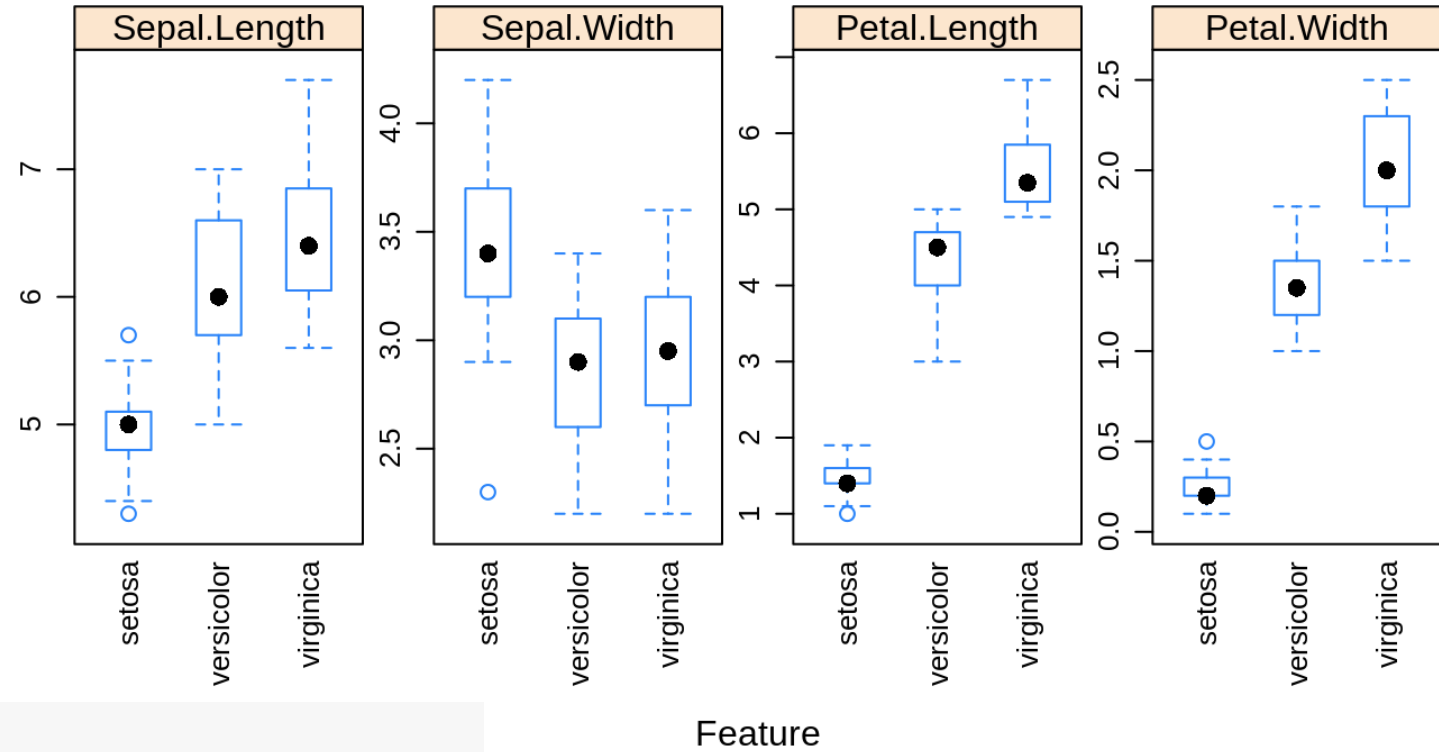


Estimating the center μ_k

- Estimate the **center** of each class μ_k :

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

where $n_k = \#\{i: y_i = k\}$

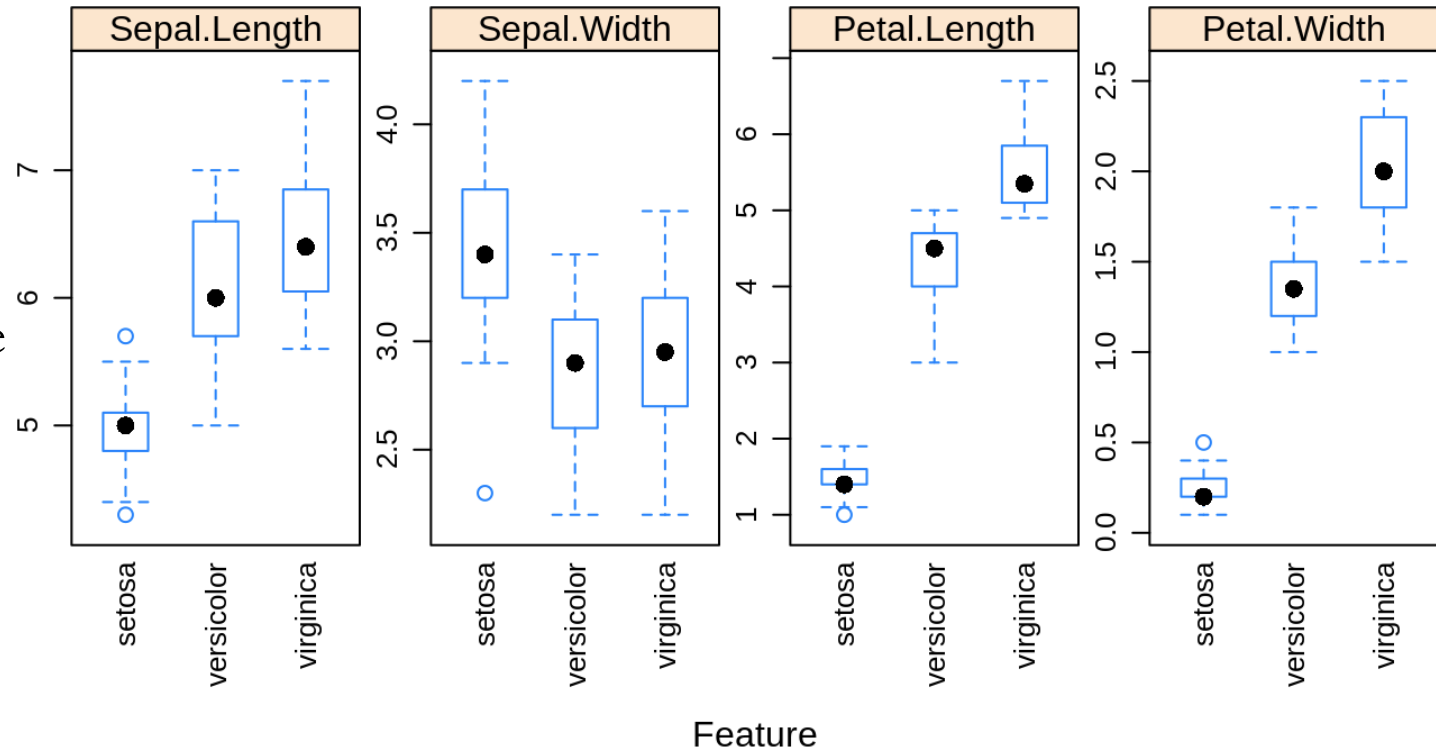


Group means:

##	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
## setosa	4.958621	3.420690	1.458621	0.237931
## versicolor	6.063636	2.845455	4.318182	1.354545
## virginica	6.479167	2.937500	5.479167	2.045833

Σ in linear discriminant analysis

- Σ is the same for *versicolor*, *setosa*, *virginica*
 - Diagonal entries equal to variance of each feature for all classes
 - Proportional to the width of the box plots
 - Off-diagonal entries equal to covariance between two features for all classes
- Is this a good model?



Quadratic discriminant analysis

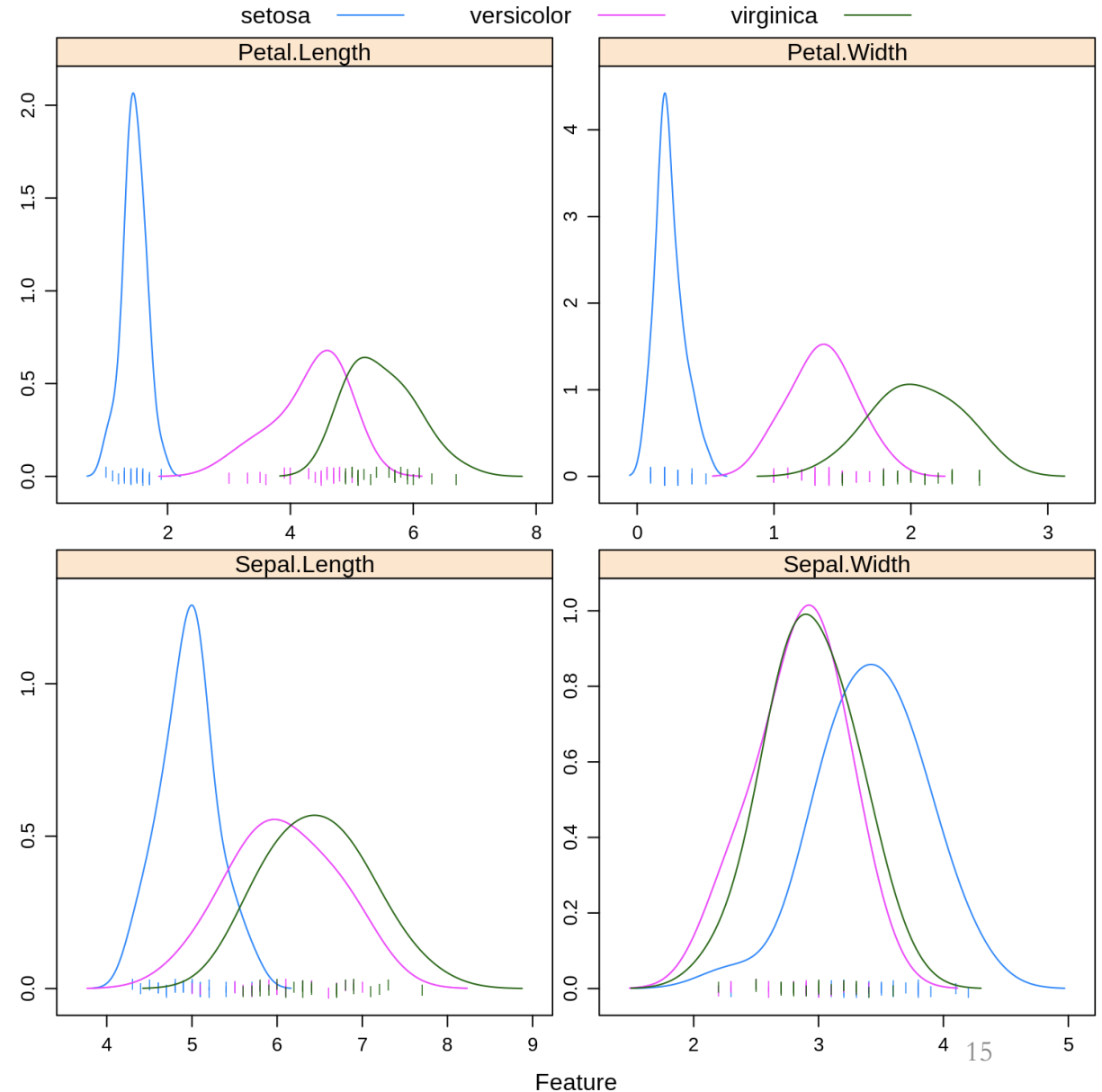
- Model $P(X = x \mid Y = k)$

- $X = \begin{bmatrix} \text{sepal length} \\ \text{sepal width} \\ \text{petal length} \\ \text{petal width} \end{bmatrix}$

- $Y \in \{\text{versicolor}, \text{setosa}, \text{virginica}\}$

by a *Multivariate Normal Distribution*

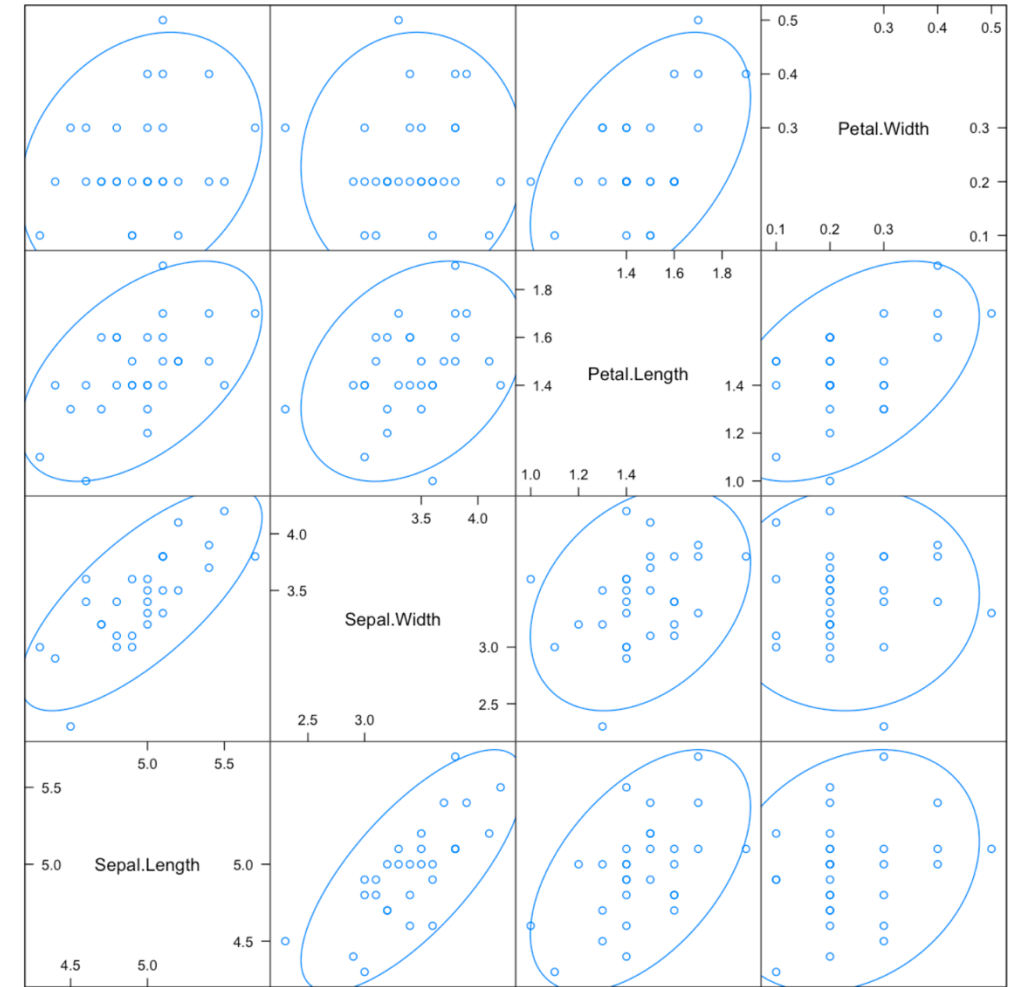
$N(\mu_k, \Sigma_k)$ with mean μ_k , covariance matrix Σ_k



Σ_k in quadratic discriminant analysis

- Σ_{setosa}

- Diagonal entries equal to variance of each feature for setosa
- Off-diagonal entries equal to covariance between two features for setosa
- Question: What are signs of correlation between any two features? Which two features have strongest correlation?



Scatter Plot Matrix

Estimating the covariance Σ_k in QDA

- Estimate the covariance Σ_k

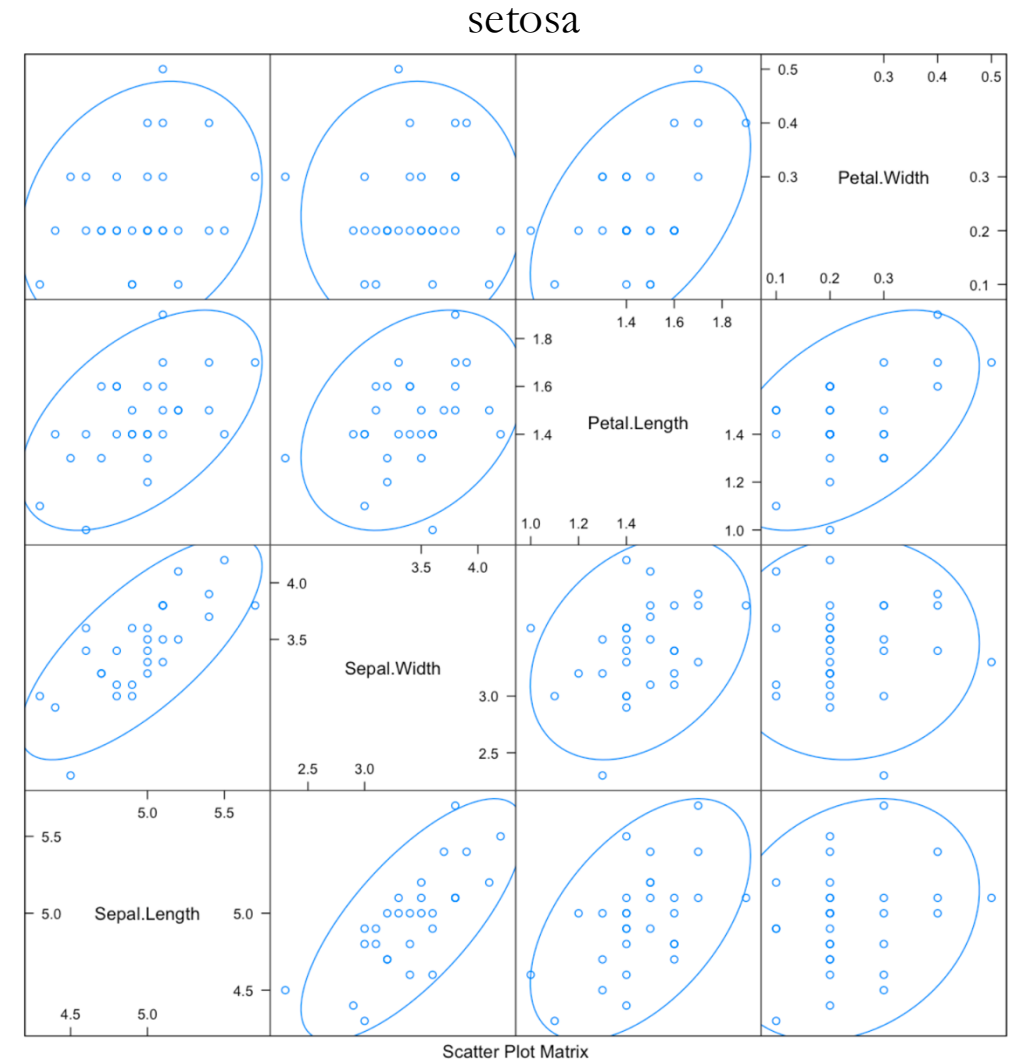
$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i: y_i = k} (x_i - \hat{\mu}_k) \cdot (x_i - \hat{\mu}_k)^\top$$

where $n_k = \#\{i: y_i = k\}$

- Example: Σ_{setosa}

```
iris_trn_setosa <- iris_trn[iris_trn$Species == "setosa",]  
cov(iris_trn_setosa[,c(1:4)])
```

```
##          Sepal.Length Sepal.Width Petal.Length Petal.Width  
## Sepal.Length 0.103226601 0.095172414 0.031798030 0.007697044  
## Sepal.Width  0.095172414 0.160985222 0.025172414 0.001687192  
## Petal.Length 0.031798030 0.025172414 0.035369458 0.009125616  
## Petal.Width  0.007697044 0.001687192 0.009125616 0.009581281
```



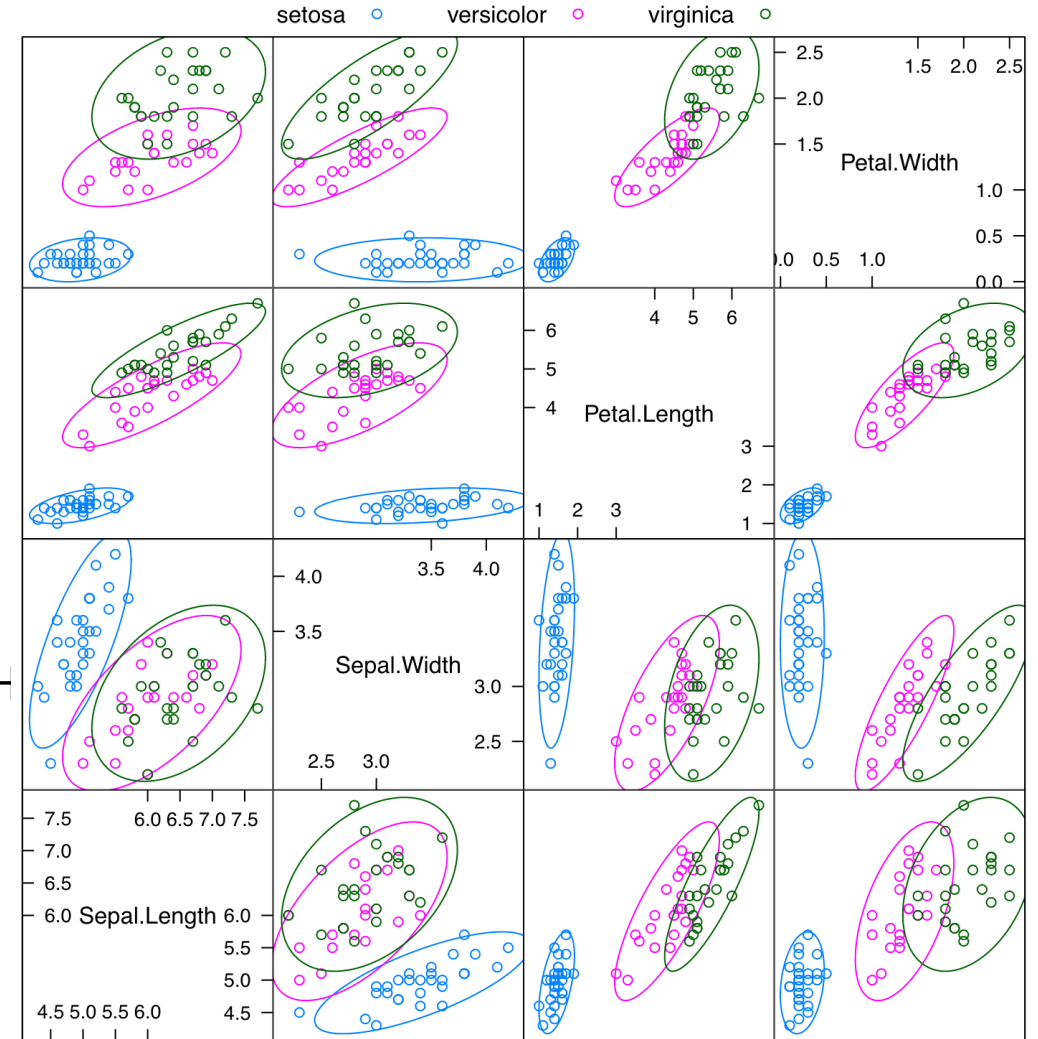
Estimating the covariance Σ in LDA

- Estimate the covariance Σ

$$\hat{\Sigma} = \sum_{k=1}^K \frac{n_k - 1}{n - K} \cdot \hat{\Sigma}_k$$

where $n_k = \#\{i: y_i = k\}$

- Example: $\hat{\Sigma} = \frac{n_{setosa}-1}{n-3} \cdot \hat{\Sigma}_{setosa} + \frac{n_{versicolor}-1}{n-3} \cdot \hat{\Sigma}_{versicolor} + \frac{n_{virginica}-1}{n-3} \cdot \hat{\Sigma}_{virginica}$



Summary of LDA

- For each class k , we model $P(X = x|Y = k) = f_k(x)$ as a *Multivariate Normal Distribution* $N(\mu_k, \Sigma)$ with mean μ_k and covariance matrix Σ
- We estimate $\hat{P}(X = x|Y = k)$ as $N(\hat{\mu}_k, \hat{\Sigma})$ and $\hat{P}(Y = k) = \hat{\pi}_k$
- We apply to Bayes theorem to obtain $P(Y = k | X = x)$

$$\hat{P}(Y = k | X = x) = \frac{\hat{P}(Y = k, X = x)}{\hat{P}(X = x)} = \frac{\hat{P}(X = x | Y = k)\hat{P}(Y = k)}{\sum_j \hat{P}(X = x | Y = j)\hat{P}(Y = j)}$$

Summary of QDA

- For each class k , we model $P(X = x|Y = k) = f_k(x)$ as a *Multivariate Normal Distribution* $N(\mu_k, \Sigma_k)$ with mean μ_k and covariance matrix Σ_k
- We estimate $\hat{P}(X = x|Y = k)$ as $N(\hat{\mu}_k, \hat{\Sigma}_k)$ and $\hat{P}(Y = k) = \hat{\pi}_k$
- We apply to Bayes theorem to obtain $P(Y = k | X = x)$

$$\hat{P}(Y = k | X = x) = \frac{\hat{P}(Y = k, X = x)}{\hat{P}(X = x)} = \frac{\hat{P}(X = x | Y = k)\hat{P}(Y = k)}{\sum_j \hat{P}(X = x | Y = j)\hat{P}(Y = j)}$$

Comparison between LDA and QDA

- **Decision boundary:** the set of points in which 2 classes do just as well
 - LDA has *linear* decision boundary
 - QDA has *quadratic* decision boundary
- **Bias-variance tradeoff**
 - LDA is less flexible but has a smaller variance. Small sample size n : LDA
 - QDA requires more parameters. Large sample size n : QDA

Question: What is the shape of decision boundary of logistic regression $\log \frac{P(Y = 1|X)}{P(Y = 0|X)}$ $\beta' \tilde{X}$? What about KNN?

