SB2b Statistical Machine Learning
Hilary Term 2017

Seth Flaxman

http://www.stats.ox.ac.uk/~flaxman/course_ml.html

Linear Discriminant Analysis

- **LDA** is the most well-known and simplest example of plug-in classification.
- Assume multivariate normal conditional density $g_k(x)$ for each class k:

$$X|Y = k \sim \mathcal{N}(\mu_k, \Sigma),$$

$$g_k(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)\right),$$

  - each class can have a **different mean** μ_k,
  - all classes share the **same covariance** Σ.
- For an observation x, the k-th log-discriminant function is

$$\log \pi_k g_k(x) = c + \log \pi_k - \frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)$$

  The quantity $(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)$ is the squared **Mahalanobis distance** between x and μ_k.
- If $\Sigma = I_p$ and $\pi_k = \frac{1}{K}$, LDA simply chooses the class k with the nearest (in the Euclidean sense) class mean.

Linear Discriminant Analysis

- Expanding the term $(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)$,

$$\log \pi_k g_k(x) = c + \log \pi_k - \frac{1}{2}\left(\mu_k^\top \Sigma^{-1}\mu_k - 2\mu_k^\top \Sigma^{-1}x + x^\top \Sigma^{-1}x\right)$$
$$= c' + \log \pi_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \mu_k^\top \Sigma^{-1}x$$

- Setting $a_k = \log(\pi_k) - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k$ and $b_k = \Sigma^{-1}\mu_k$, we obtain

$$\log \pi_k g_k(x) = c' + a_k + b_k^\top x$$

  i.e. a **linear** discriminant function in x.
- Consider choosing class k over k':

$$a_k + b_k^\top x > a_{k'} + b_{k'}^\top x \qquad \Leftrightarrow \qquad a_\star + b_\star^\top x > 0$$

  where $a_\star = a_k - a_{k'}$ and $b_\star = b_k - b_{k'}$.
- The Bayes classifier thus partitions $\mathcal{X}$ into regions with the same class predictions via **separating hyperplanes**.
- The Bayes classifier under these assumptions is more commonly known as the **LDA classifier**.

Parameter Estimation

- How to estimate the parameters of the LDA model?
- We can achieve this by maximum likelihood (EM algorithm is not needed here since the class variables y_i are observed!).
- Let $n_k = \#\{j : y_j = k\}$ be the number of observations in class k.

$$\ell(\pi, (\mu_k)_{k=1}^K, \Sigma) = \log p\left((x_i, y_i)_{i=1}^n \,|\pi, (\mu_k)_{k=1}^K, \Sigma\right) = \sum_{i=1}^n \log \pi_{y_i} g_{y_i}(x_i)$$

$$= c + \sum_{k=1}^K \sum_{j:y_j=k} \log \pi_k - \frac{1}{2}\left(\log|\Sigma| + (x_j - \mu_k)^\top \Sigma^{-1}(x_j - \mu_k)\right)$$

ML estimates:

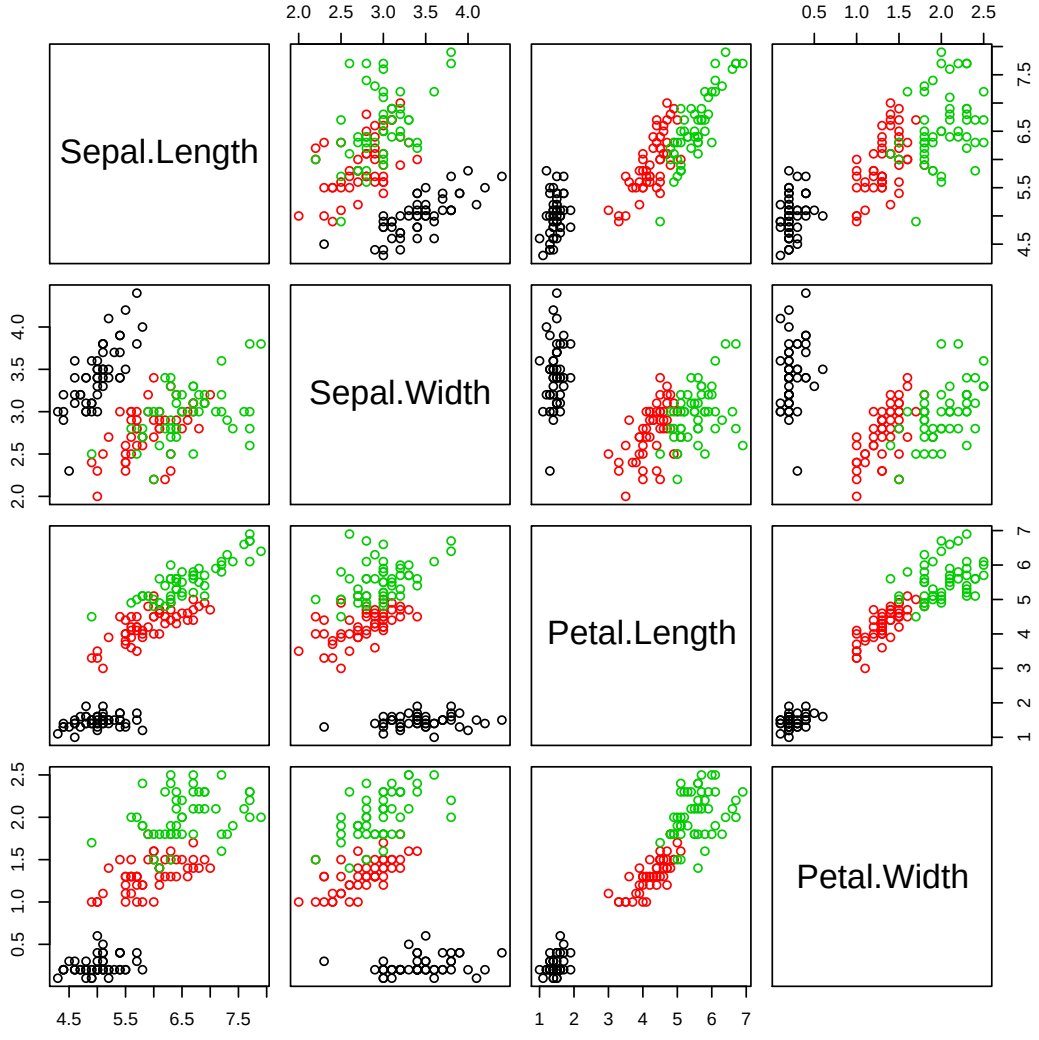
$$\hat{\pi}_k = \frac{n_k}{n} \qquad \hat{\mu}_k = \frac{1}{n_k}\sum_{j:y_j=k} x_j$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^K \sum_{j:y_j=k} (x_j - \hat{\mu}_k)(x_j - \hat{\mu}_k)^\top$$

- Note: the ML estimate of Σ is biased. For an unbiased estimate we need to divide by $n - K$.

Iris Dataset

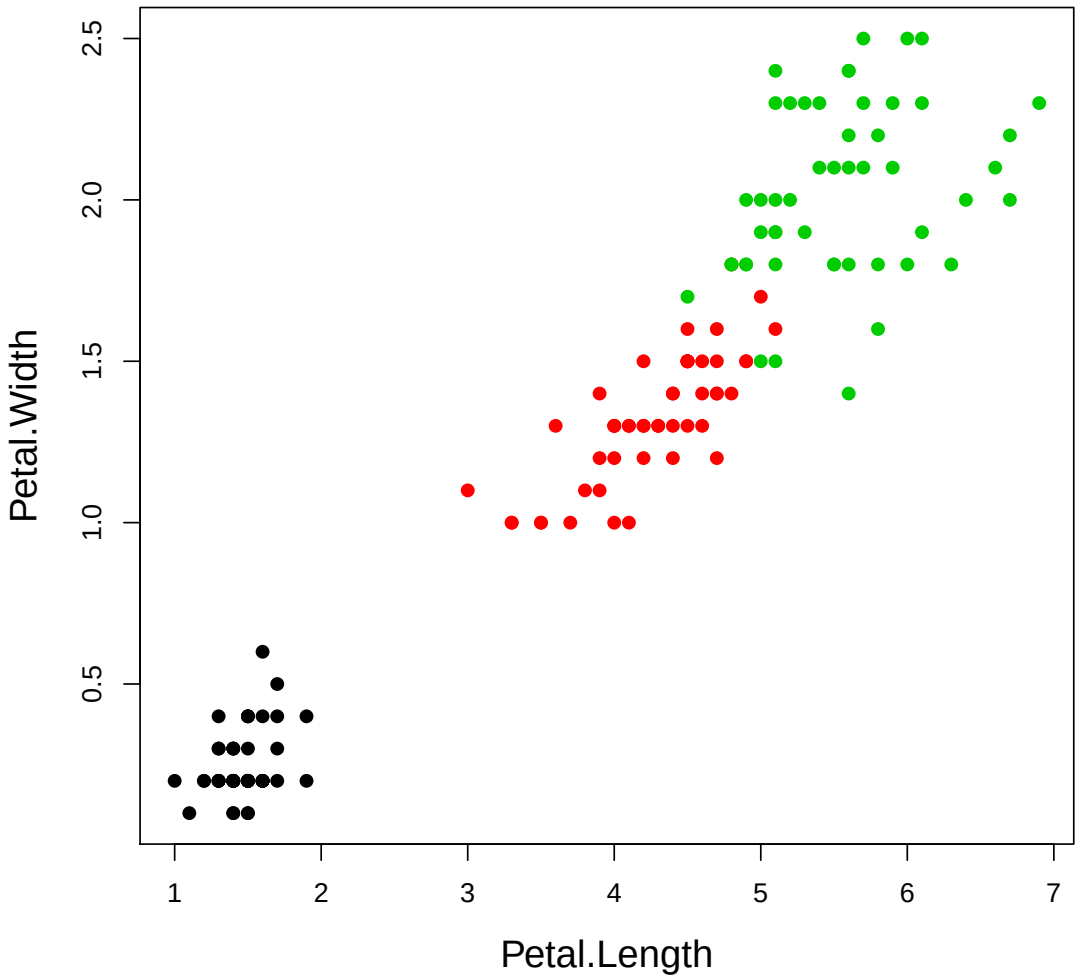
```
library(MASS)
data(iris)
##save class labels
ct <- unclass(iris$Species)
##pairwise plot
pairs(iris[,1:4],col=ct)
```



Iris Dataset

Just focus on two predictor variables.

```
iris.data <- iris[,3:4]
plot(iris.data,col=ct,pch=20,cex=1.5,cex.lab=1.4)
```

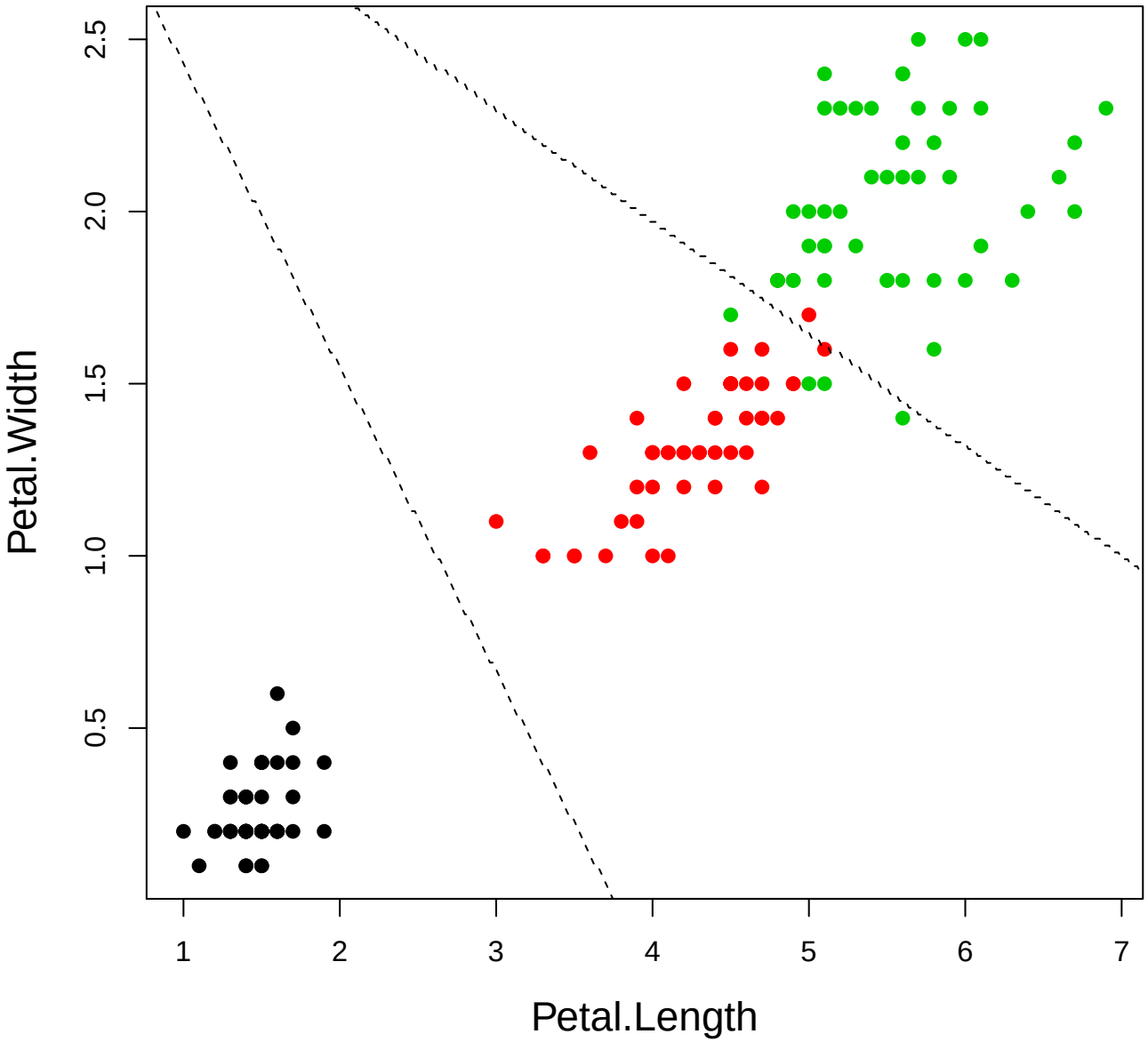


Iris Dataset

Computing and plotting the LDA boundaries.

```
##fit LDA
iris.lda <- lda(x=iris.data,grouping=ct)

##create a grid for our plotting surface
x <- seq(0,8,0.02)
y <- seq(0,3,0.02)
m <- length(x)
n <- length(y)
z <- as.matrix(expand.grid(x,y),0)

##classes are 1,2 and 3, so set contours at 1.5 and 2.5
iris.ldp <- predict(iris.lda,z)$class
contour(x,y,matrix(iris.ldp,m,n),
        levels=c(1.5,2.5), add=TRUE, d=FALSE, lty=2)
```

Iris Dataset



Crabs Dataset

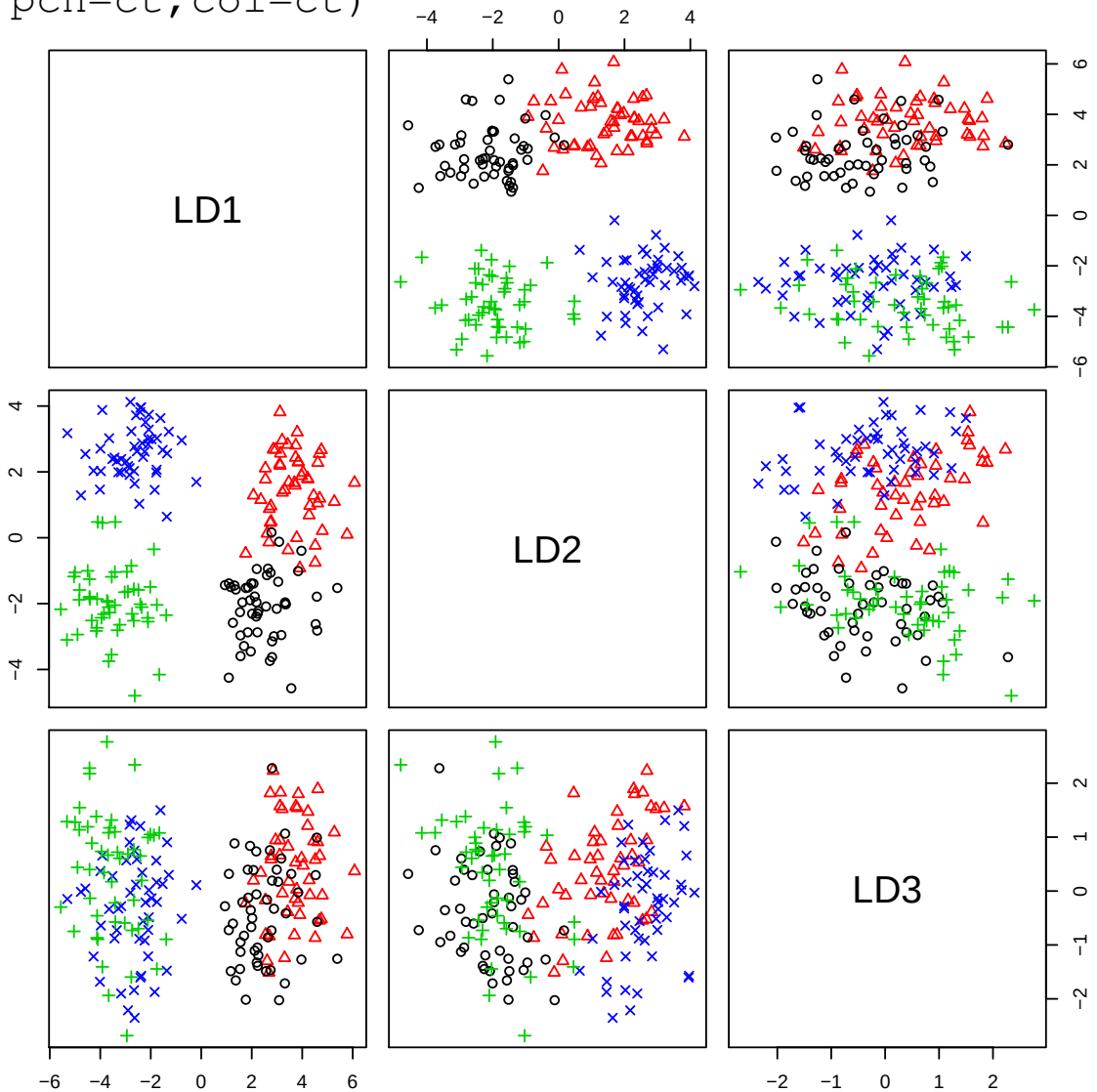
```
library(MASS)
data(crabs)

## create class labels (species+sex)
crabs$spsex=factor(paste(crabs$sp,crabs$sex,sep=""))
ct <- unclass(crabs$spsex)

## LDA on crabs in log-domain
cb.lda <- lda(log(crabs[,4:8]),ct)
```

Crabs Dataset

```
> cb.lda
Call:
lda(log(crabs[, 4:8]), ct)

Prior probabilities of groups:
   1    2    3    4
0.25 0.25 0.25 0.25

Group means:
        FL       RW       CL       CW       BD
1 2.564985 2.475174 3.312685 3.462327 2.441351
2 2.672724 2.443774 3.437968 3.578077 2.560806
3 2.852455 2.683831 3.529370 3.649555 2.733273
4 2.787885 2.489921 3.490431 3.589426 2.701580

Coefficients of linear discriminants:
          LD1        LD2        LD3
FL -31.217207  -2.851488  25.719750
RW  -9.485303 -24.652581  -6.067361
CL  -9.822169  38.578804 -31.679288
CW  65.950295 -21.375951  30.600428
BD -17.998493   6.002432 -14.541487

Proportion of trace:
   LD1    LD2    LD3
0.6891 0.3018 0.0091
```

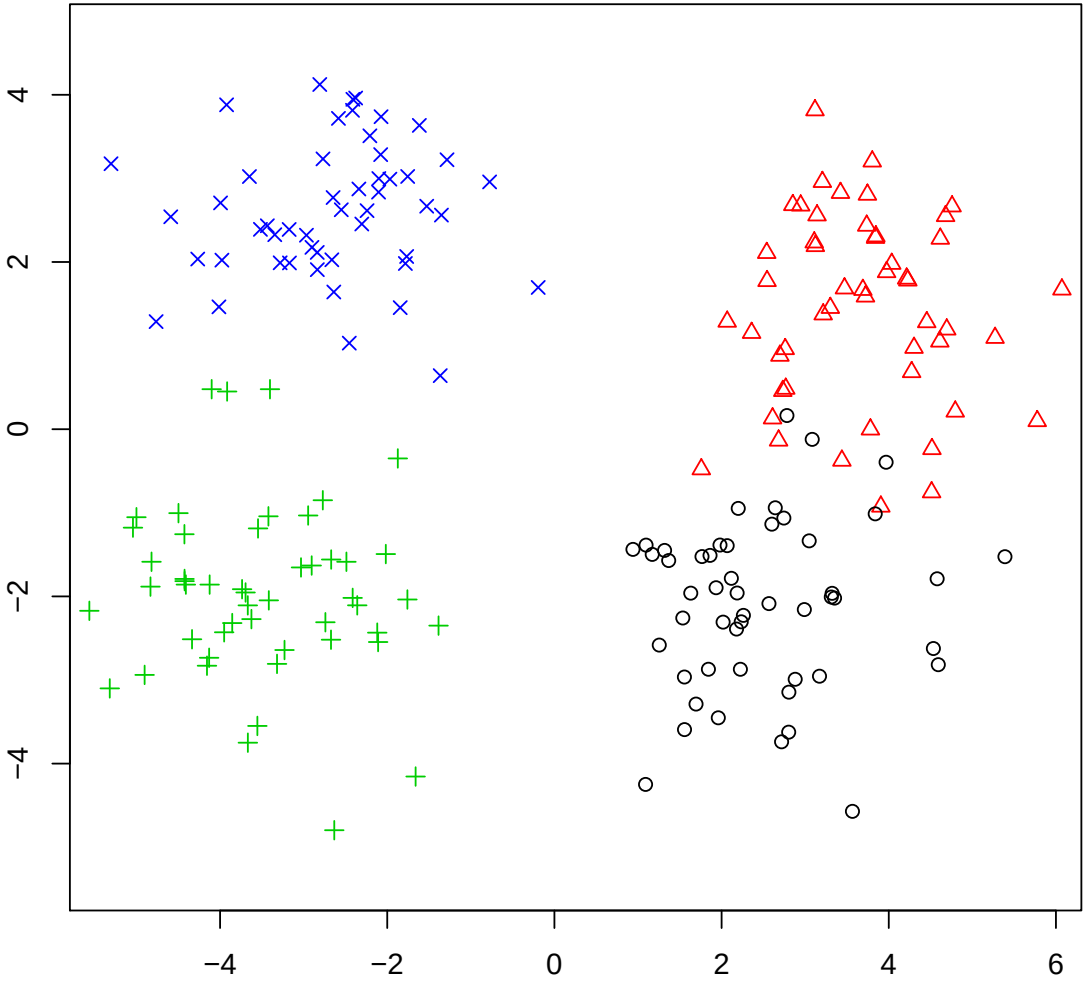
Crabs Dataset

```
cb.ldp <- predict(cb.lda)
pairs(cb.ldp$x,pch=ct,col=ct)
```



Crabs Dataset

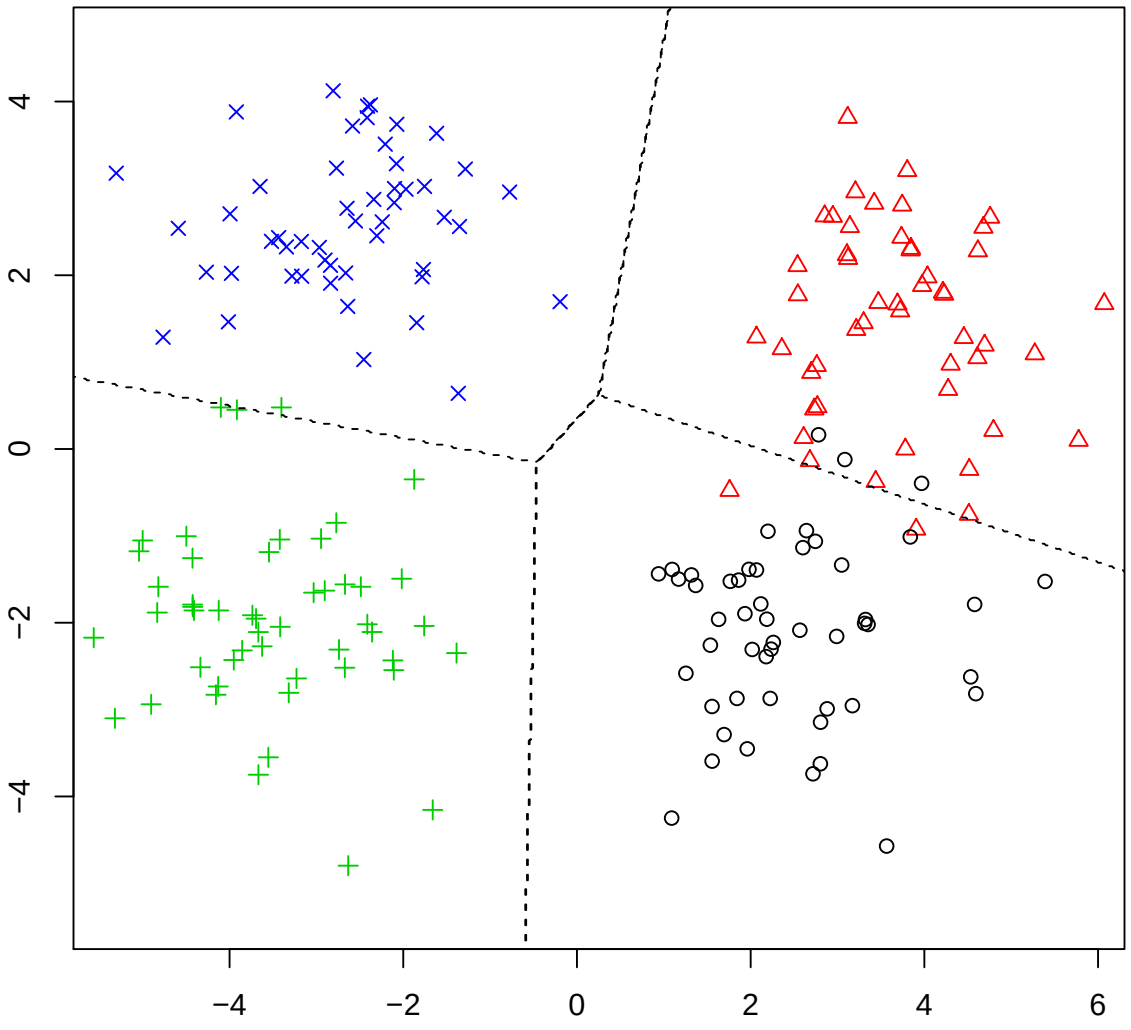
```
cb.ldp12 <- cb.ldp$x[,1:2]
eqscplot(cb.ldp12,pch=ct,col=ct)
```



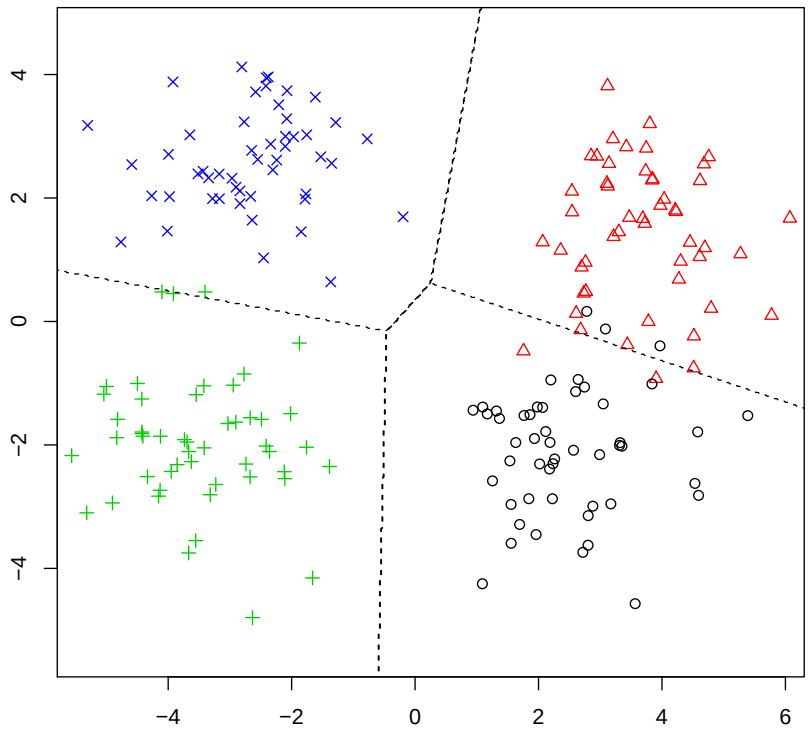
Crabs Dataset

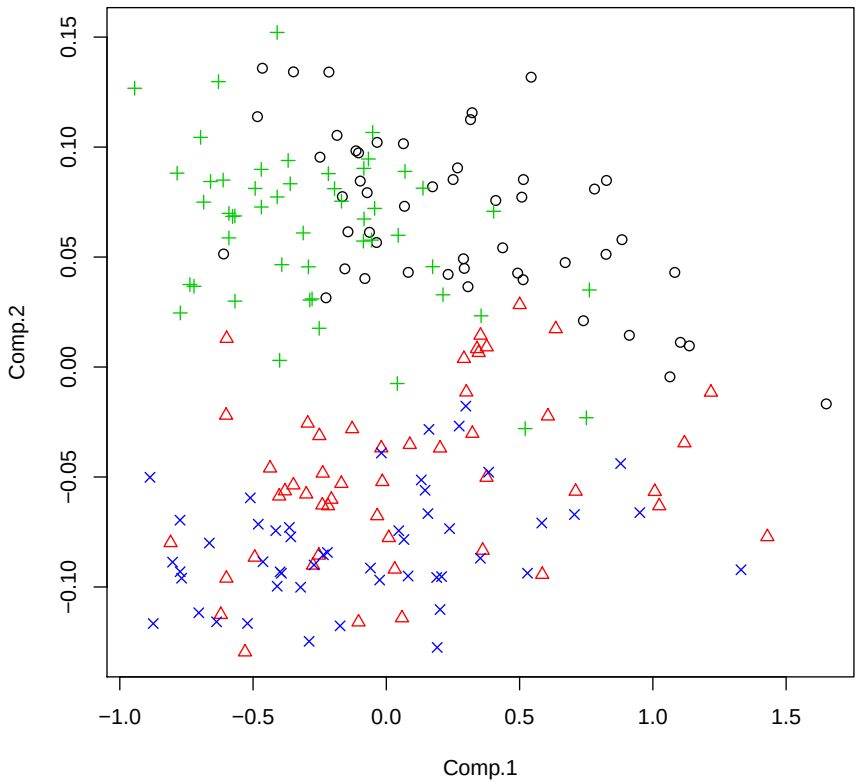
```
## display the decision boundaries
## take a lattice of points in LD-space
x <- seq(-6,7,0.02)
y <- seq(-6,7,0.02)
z <- as.matrix(expand.grid(x,y))
m <- length(x)
n <- length(y)

## perform LDA on first two discriminant directions
cb.lda_new <- lda(cb.ldp12,ct)
## predict onto the grid
cb.ldpp <- predict(cb.lda_new,z)$class

## classes are 1,2,3 and 4 so set contours
## at 1.5,2.5 and 3.5
contour(x,y,matrix(cb.ldpp,m,n),
        levels=c(1.5,2.5,3.5),
        add=TRUE,d=FALSE,lty=2)
```

Crabs Dataset



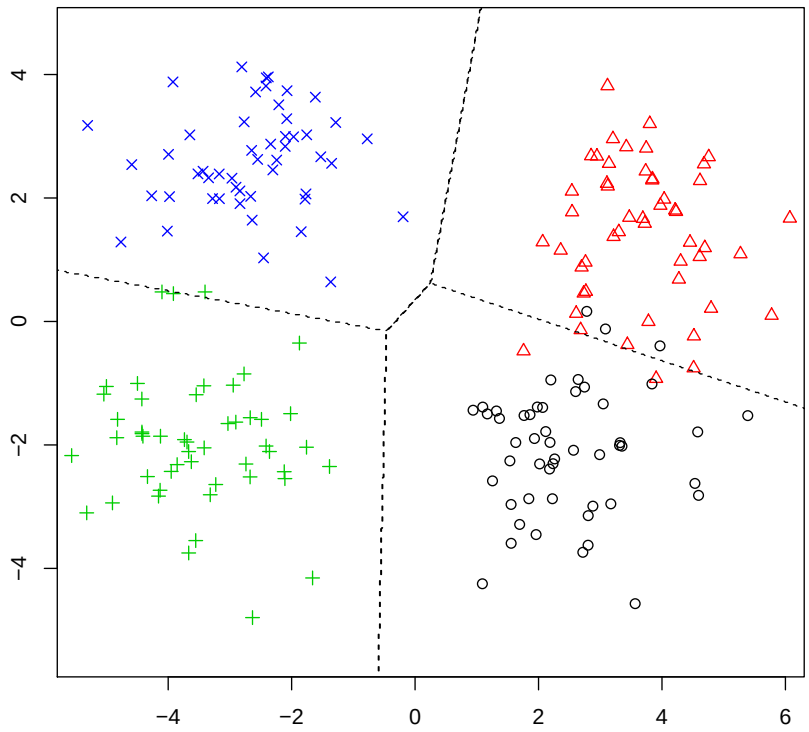
LDA vs PCA projections

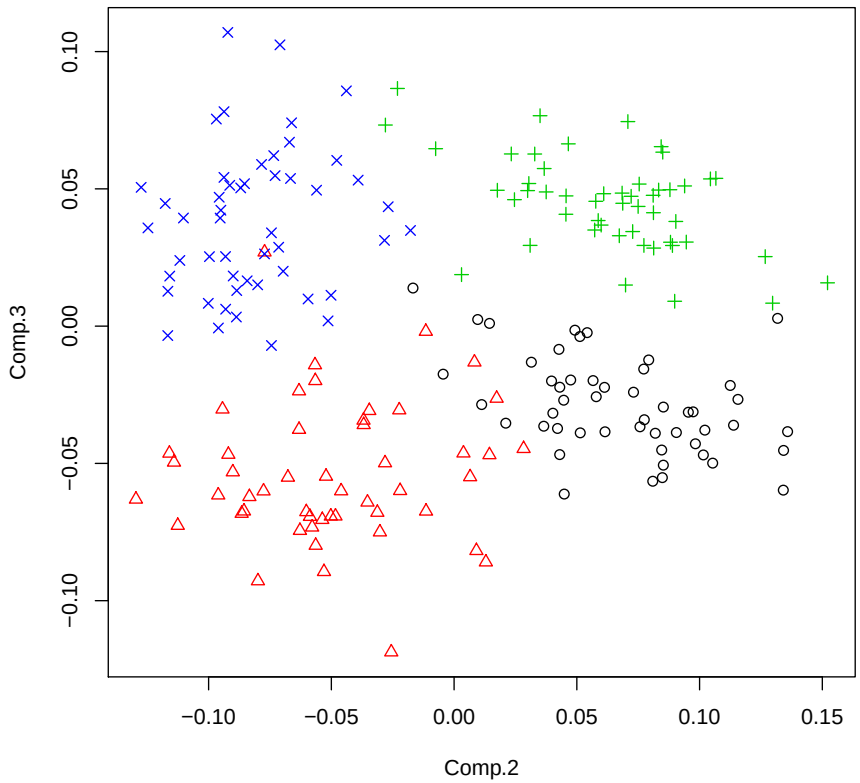




LDA separates the groups better.

LDA vs PCA projections





LDA separates the groups better.

Conditional densities with different covariances

Given training data with K classes, assume a parametric form for conditional density $g_k(x)$, where for each class

$$X|Y = k \quad \sim \quad \mathcal{N}(\mu_k, \Sigma_k),$$

i.e., instead of assuming that every class has a different mean μ_k with the **same** covariance matrix Σ (LDA), we now allow each class to have its own covariance matrix.
Considering $\log \pi_k g_k(x)$ as before,

$$\begin{aligned}
\log \pi_k g_k(x) &= \text{const} + \log(\pi_k) - \frac{1}{2}\left(\log|\Sigma_k| + (x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right) \\
&= \text{const} + \log(\pi_k) - \frac{1}{2}\left(\log|\Sigma_k| + \mu_k^T\Sigma_k^{-1}\mu_k\right) \\
&\qquad + \mu_k^T\Sigma_k^{-1}x - \frac{1}{2}x^T\Sigma_k^{-1}x \\
&= a_k + b_k^T x + x^T c_k x.
\end{aligned}$$

A **quadratic** discriminant function instead of linear.

Quadratic decision boundaries

Again, by considering that we choose class k over k',

$$\begin{aligned} a_k + b_k^T x + x^T c_k x - (a_{k'} + b_{k'}^T x + x^T c_{k'} x) \\ = a_\star + b_\star^T x + x^T c_\star x > 0 \end{aligned}$$

we see that the decision boundaries of the Bayes Classifier are quadratic surfaces.

- The plug-in Bayes Classifer under these assumptions is known as the **Quadratic Discriminant Analysis** (QDA) Classifier.

QDA

LDA classifier:

$$f_{\mathsf{LDA}}(x) = \underset{k\in\{1,\ldots,K\}}{\arg\min} \left\{ (x-\hat{\mu}_k)^T \hat{\Sigma}^{-1} (x-\hat{\mu}_k) - 2\log(\hat{\pi}_k) \right\}$$

QDA classifier:

$$f_{\mathsf{QDA}}(x) = \underset{k\in\{1,\ldots,K\}}{\arg\min} \left\{ (x-\hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x-\hat{\mu}_k) - 2\log(\hat{\pi}_k) + \log(|\hat{\Sigma}_k|) \right\}$$

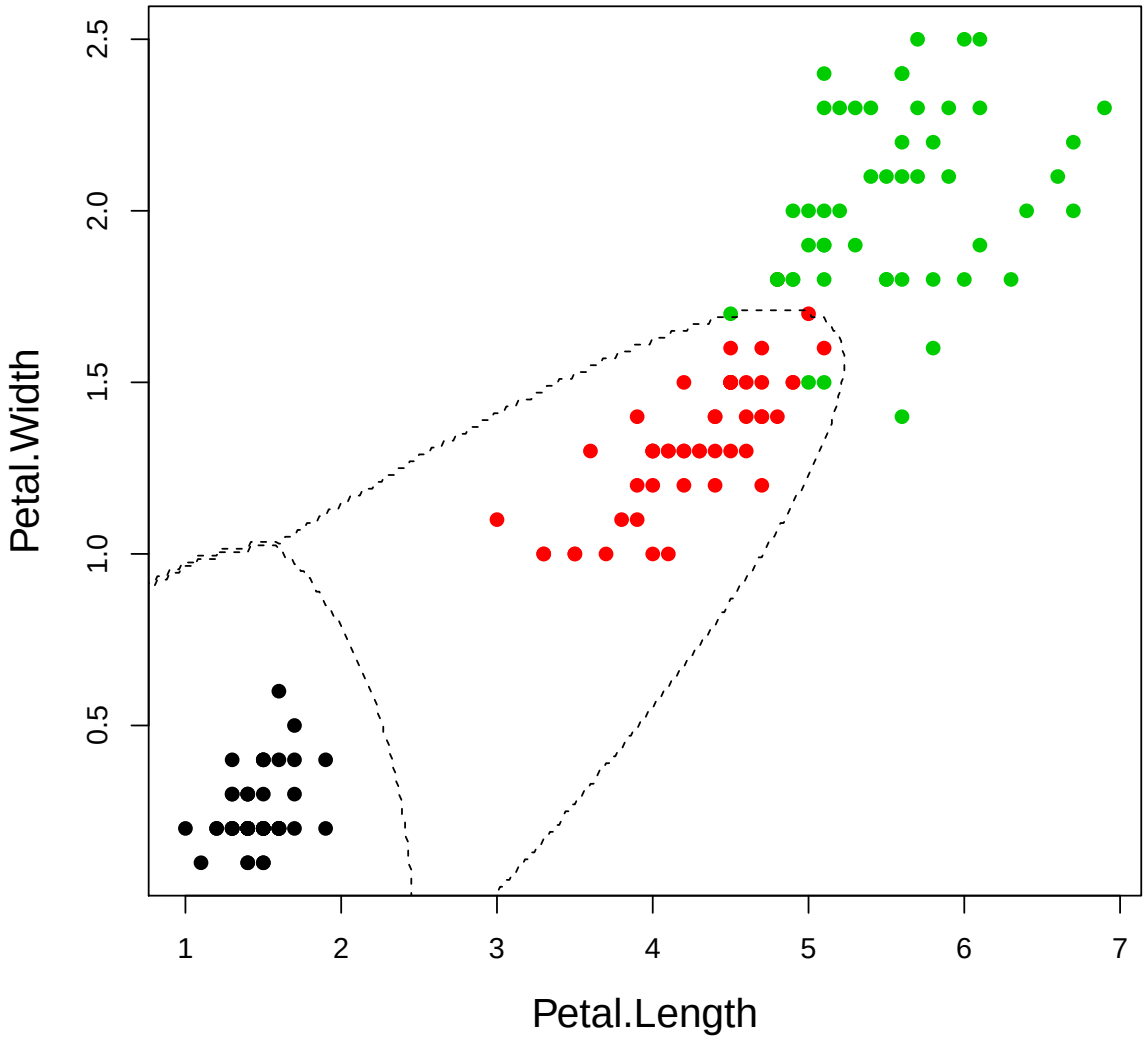
for each point $x \in \mathcal{X}$ where the plug-in estimate $\hat{\mu}_k$ is as before and $\hat{\Sigma}_k$ is (in contrast to LDA) estimated for each class $k = 1, \ldots, K$ separately:

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{j:y_j=k} (x_j - \hat{\mu}_k)(x_j - \hat{\mu}_k)^T.$$

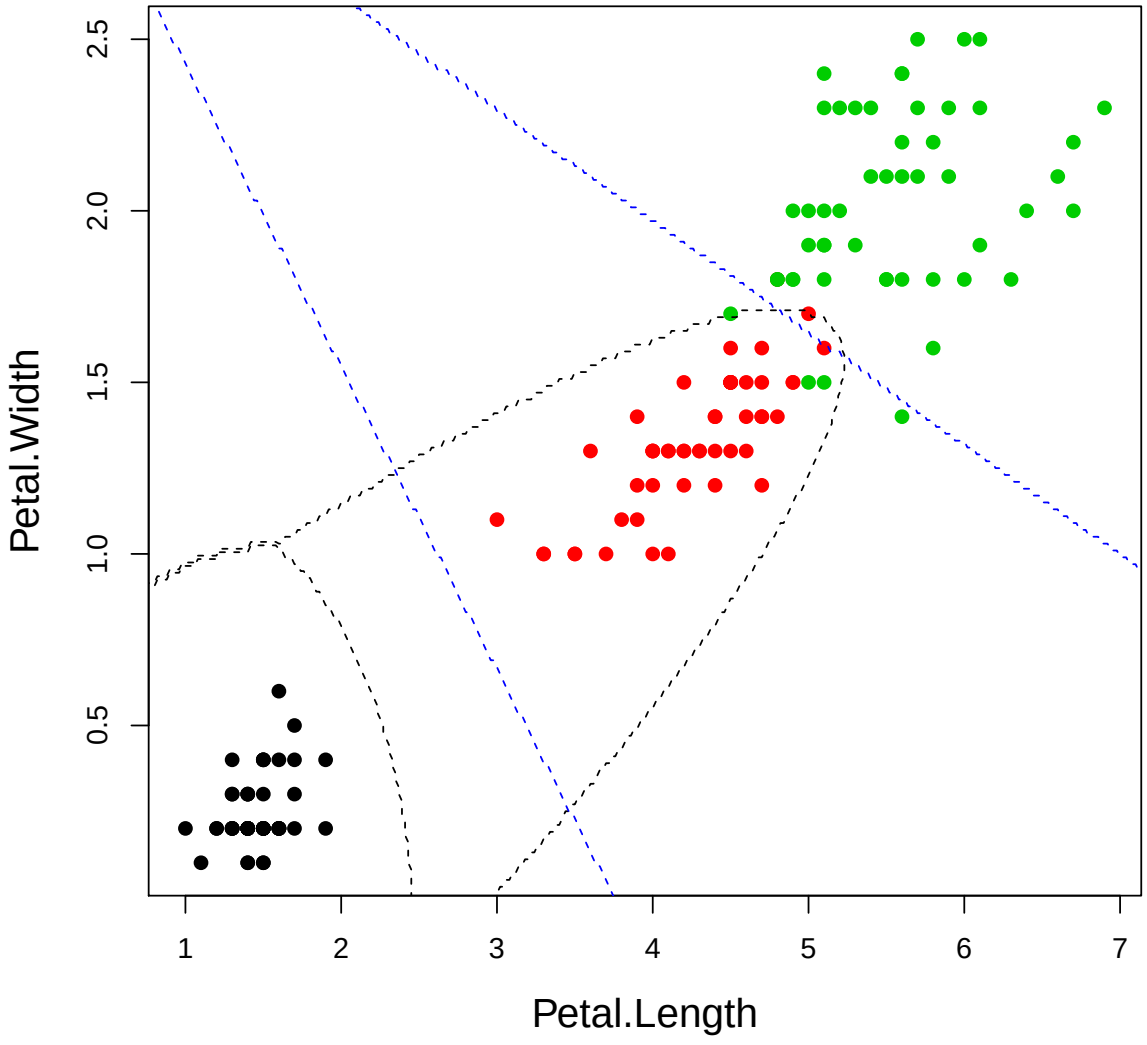
Computing and plotting the QDA boundaries.

```
##fit QDA
iris.qda <- qda(x=iris.data,grouping=ct)

##create a grid for our plotting surface
x <- seq(-6,6,0.02)
y <- seq(-4,4,0.02)
z <- as.matrix(expand.grid(x,y),0)
m <- length(x)
n <- length(y)


iris.qdp <- predict(iris.qda,z)$class
contour(x,y,matrix(iris.qdp,m,n),
        levels=c(1.5,2.5), add=TRUE, d=FALSE, lty=2)
```

Iris example: QDA boundaries



Iris example: QDA boundaries



LDA or QDA?

- Having seen both LDA and QDA in action, it is natural to ask which is the "better" classifier.
- If the covariances of different classes are very distinct, QDA will probably have an advantage over LDA.
- Parametric models are only ever approximations to the real world, allowing **more flexible decision boundaries** (QDA) may seem like a good idea. However, there is a price to pay in terms of increased variance and potential **overfitting**.