Linear Discriminant Analysis

-Applied Analytics-

APM chapter 12.3 & ISL. chapter 4.4

Lecturer: Darren Homrighausen, PhD

Preamble:

- Discuss an odd limitation to logistic regression
- Introduce linear discriminant analysis (LDA) as a classification procedure
- Compare LDA and logistic regression

Logistic regression

Logistic regression estimates the following model

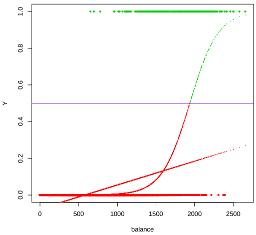
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

Logistic regression is a very popular model based on its simplicity and interpretability

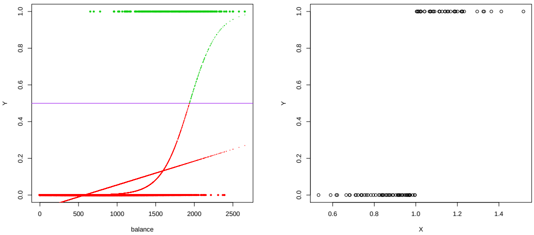
There are several drawbacks to logistic regression:

- If the classes are separable, logistic regression is unstable (or undefined)
- It is somewhat awkward to use when the supervisor has multiple levels
  (This is called multinomial logistic (if the supervisor is nominal) and proportional odds logistic regression (if the supervisor is ordinal))

Example of separability



Example of separability




```
> glm(Y~X,family='binomial')
(Intercept)            X
     -986.2        974.2
Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      138.3
Residual Deviance: 1.989e-08   AIC: 4
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA)

Suppose our supervisor can take on $C = 4$ different levels

$$Y = \begin{cases} C_1 = \text{ 'blood type O'} \\ \vdots \\ C_4 = \text{ 'blood type AB'} \end{cases}$$

Instead of directly modeling the odds, we can make a classifier via modeling the distribution of the features

An example

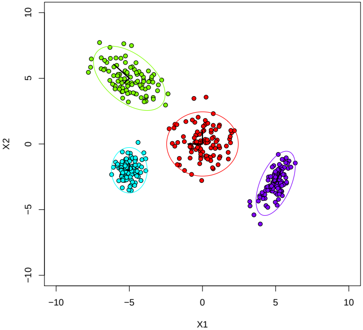
Building on the previous blood type example, suppose we measure

- the number of mosquito bites x_1
- the body temperature x_2

Then we can make a scatterplot of these features, indicating the blood type with colors

(with some transformations)

An example



We measure a new person's mosquito bites and temperature

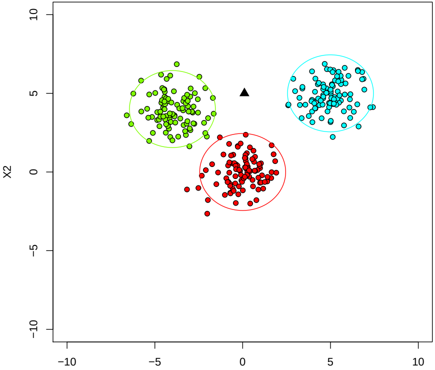
We want to predict the person's blood type

We could just classify an observation to the closest group

What do we mean by close? (Need to define distance)

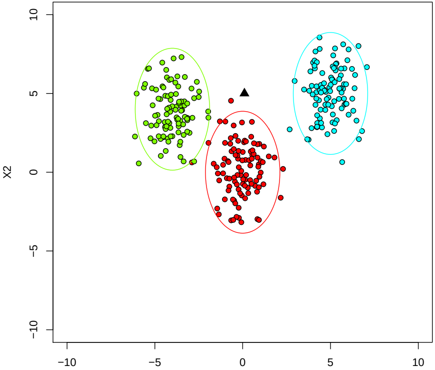
Intuition

What if the data looked like this?



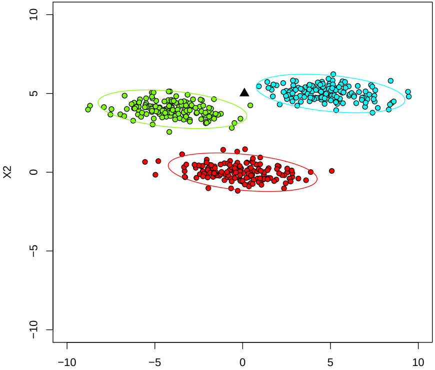
Intuition

Or this?



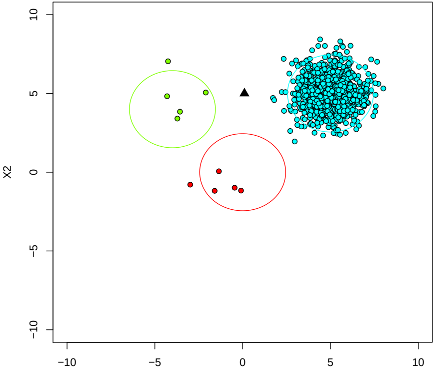
Intuition

How about this?



Intuition

What about now?



Intuition

All of these examples show that we need to take into account

- The shape of distribution of the features for each class
- The relative number of points in each class

These are the two main ingredients in LDA

LDA details

A bit more details

Reminder: If we had the true p_ℓ, then we could make the best possible classifier via choosing $Y_* = C_\ell$, where p_ℓ is the maximum

At different values of X, there are different p_ℓ, so we write $p_\ell(X)$

(Think of the default as a function of balance)

These probabilities get estimated via Bayes' theorem

(In this context, the p_ℓ are known as posterior probabilities)

The ingredients of this estimation procedure are

- Model the (X such that $Y = C_\ell$) as normal
- Use the training proportions: $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i = C_\ell)$

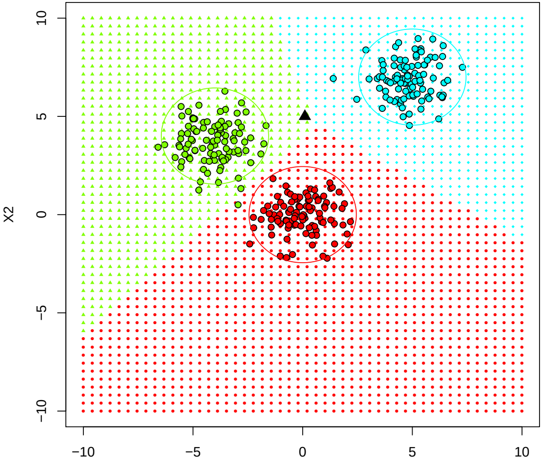
We can combine these together to form an estimate: $\hat{p}_\ell$

Note that sometimes this is (incorrectly) called Bayes' rule. See equation (12.4) in APM

Both APM and ISL give more details about how these probabilities are estimated

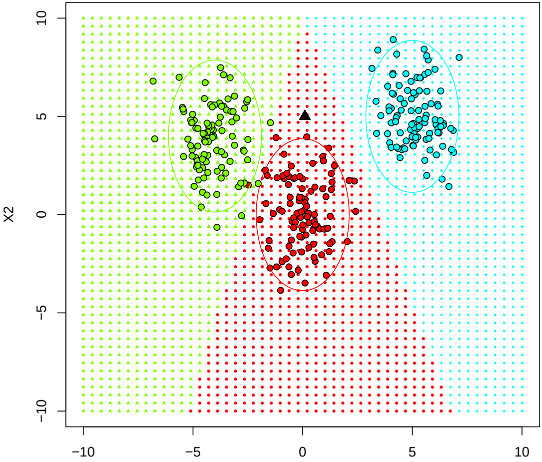
Probability estimates

What if the data looked like this?



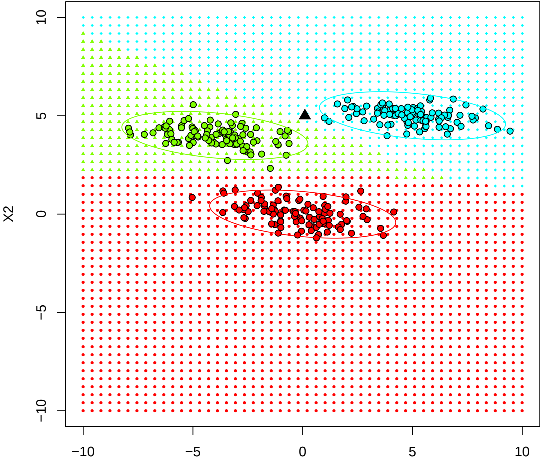
Probability estimates

Or this?



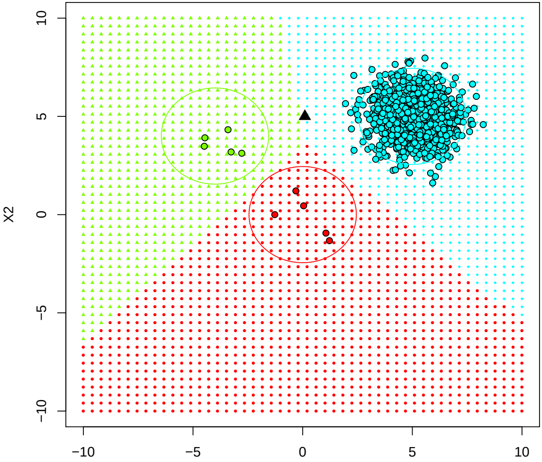
Probability estimates

How about this?



Probability estimates

What about now?



LDA in R

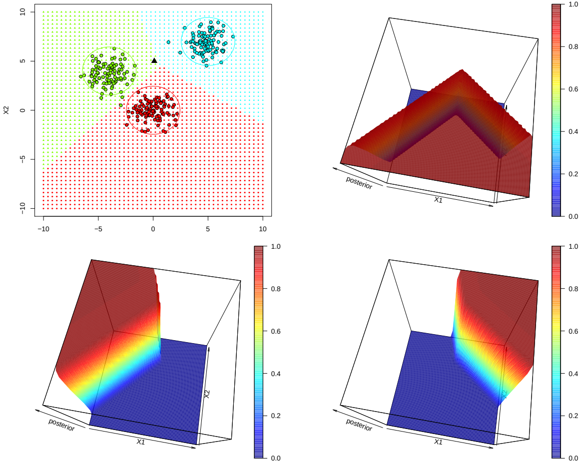
We can do this readily in R

```
library(MASS)
ldaOut  = lda(Y~.,data=X)

out = predict(ldaOut, Xtest)

> out$posterior[1:3,]
          1            2            3
1 0.9999908 9.215567e-06 1.504633e-55
2 0.9999977 2.341924e-06 1.664446e-54
3 0.9999994 5.951430e-07 1.841223e-53
```

What does posterior mean?




```
> print(predict(ldaOut,X_0)$posterior)
           1         2           3
1 0.04883796 0.9477494 0.003412639
```

Performance of LDA

The quality of the classifier produced by LDA depends on:

- The sample size n
  (This determines how accurate the parameter estimates are)
- How wrong the LDA assumptions are
  (Mainly: that X is multivariate normal)

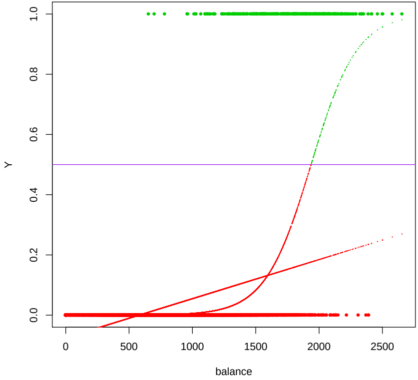
The decision boundary of a classifier determines what classification is ultimately made

It is always in terms of the features X, whereas the threshold is in terms of the probabilities

(For logistic and LDA, this corresponds to the values of X such that $\hat{p}_\ell(X) = \text{ threshold}$)

A linear decision boundary is when this set of values looks like a line

We've Already Seen Other Examples of Linear Decision Boundaries



LDA: under correct assumptions

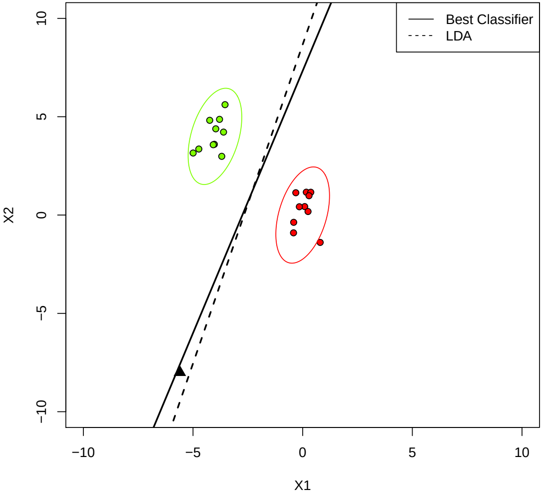


FIGURE: For $n = 20$

LDA: under correct assumptions

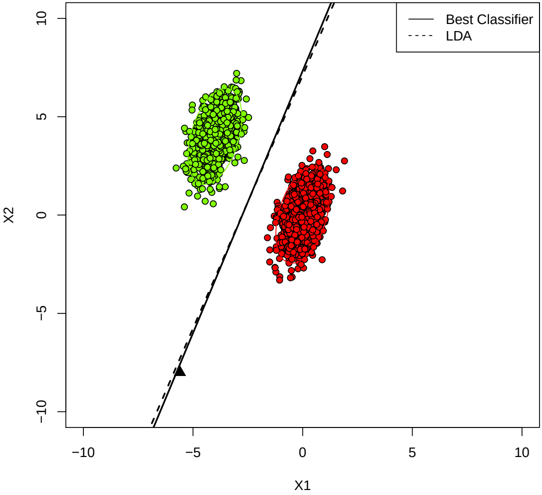


Figure: For $n = 2000$

LDA: mildly incorrect assumptions

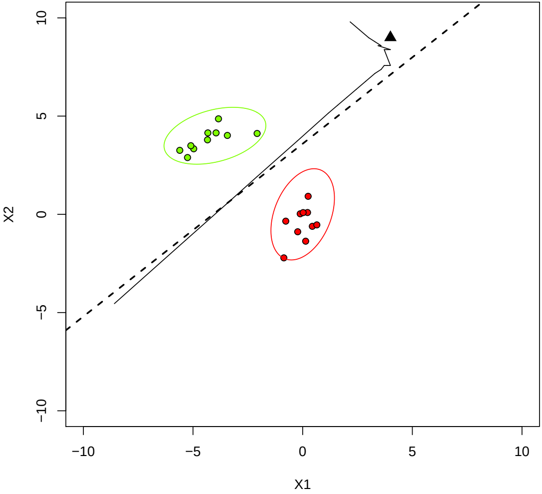


Figure: For $n = 20$

LDA: MILDLY INCORRECT ASSUMPTIONS

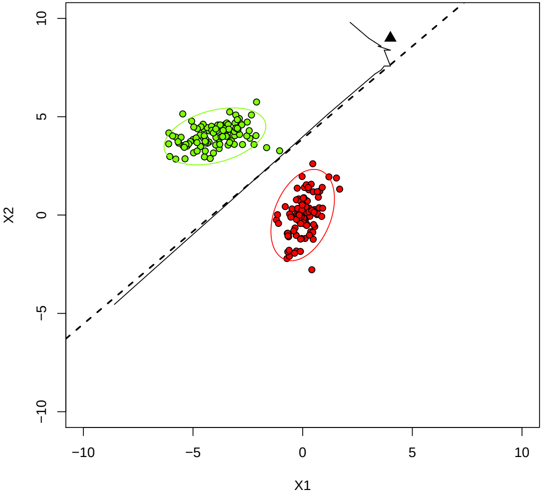


FIGURE: For $n = 200$

LDA: mildly incorrect assumptions

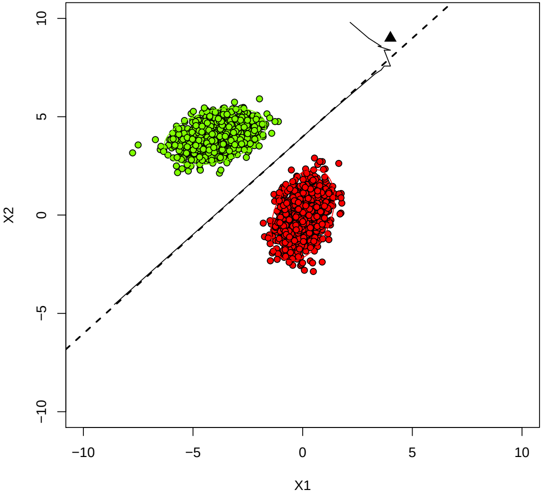


Figure: For $n = 2000$

LDA: very incorrect assumptions

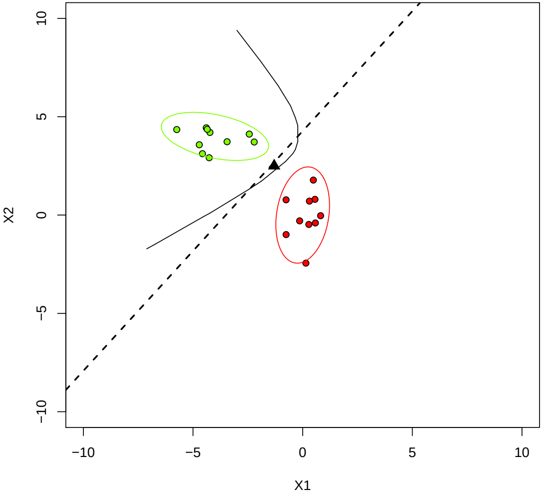


Figure: For $n = 20$

LDA: very incorrect assumptions

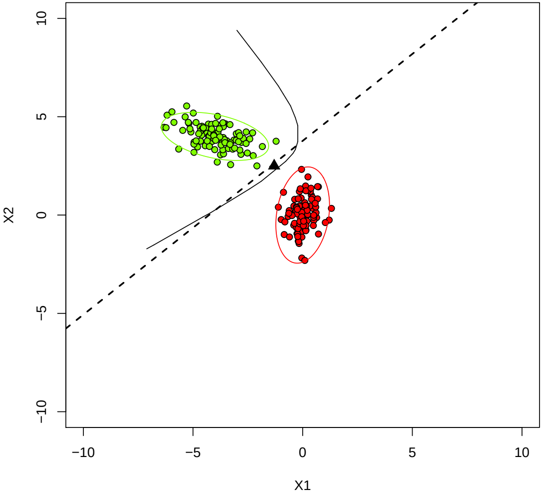


Figure: For $n = 200$

LDA: very incorrect assumptions

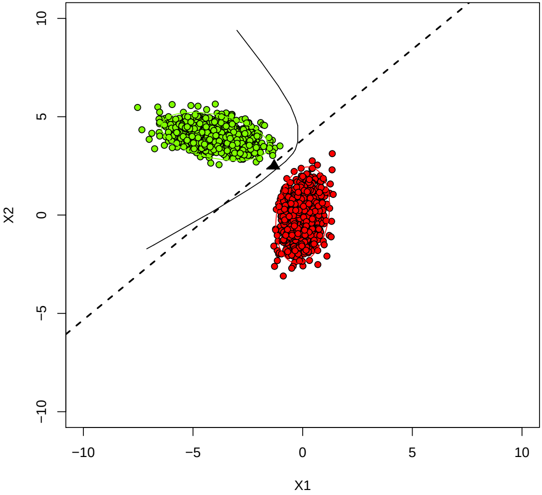


Figure: For $n = 200$

Postamble:

- Discuss an odd limitation to logistic regression

  (If the classes are perfectly linearly separable the iterative algorithm fails)
- Introduce linear discriminant analysis (LDA) as a classification procedure

  (The classifier attempts to formalize the intuition of 'assigning to nearest class')
- Compare LDA and logistic regression

  (LDA and logistic regression both model the log odds as a linear function. Logistic regression and LDA have very different assumptions. LDA assumes X is normal while logistic purely models the log odds as linear in the features)

See ISL chapter 4.5 for more on the relationship between LDA and logistic regression