ETC3250: Categorical response

Semester 1, 2020

Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 3 (a)

Categorical responses

In **classification**, the output Y is a categorical variable. For example,

- Loan approval: $Y \in \{\text{successful}, \text{unsuccessful}\}$
- Type of business culture: $Y \in \{\text{clan}, \text{adhocracy}, \text{market}, \text{hierarchical}\}$
- Historical document author: $Y \in \{\text{Austen}, \text{Dickens}, \text{Imitator}\}$
- Email: $Y \in \{\text{spam}, \text{ham}\}$

Map the categories to a numeric variable, or possibly a binary matrix.

A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?

A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?

An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user’s IP address, past transaction history, and so forth.

A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?

An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.

On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?

An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.

An email comes into the server. Should it be moved into the inbox or the junk mail box, based on header text, sender, origin, time of day, ...?

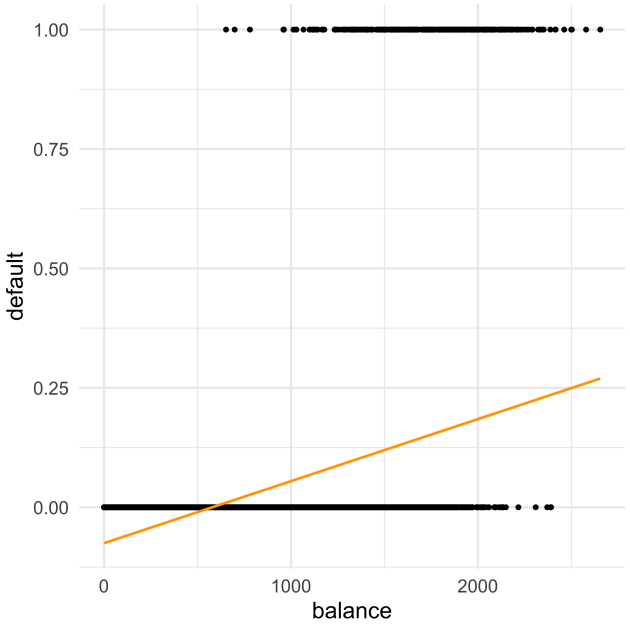
On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

When linear regression is not appropriate

Consider the following data `simcredit` which looks at the default status based on credit balance.

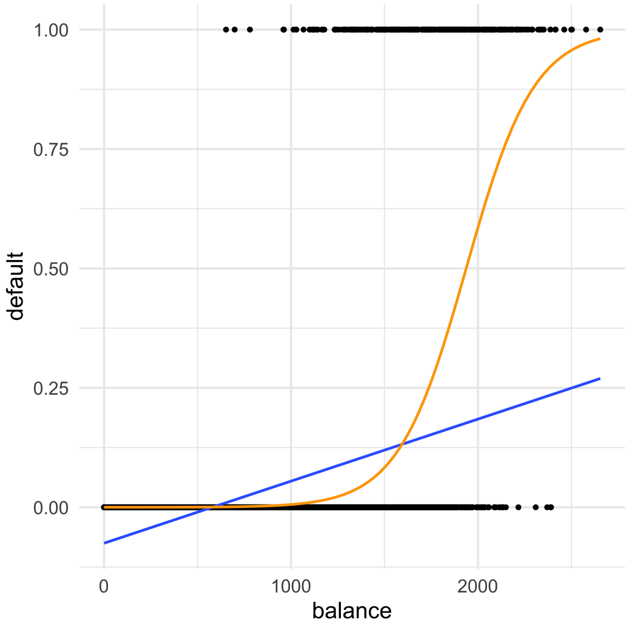
Question: Why is a linear model not appropriate for this data?

00:30



Modelling binary responses

- 📊 To model **binary data**, we need to link our **predictors** to our response using a *link function*.
- 📊 There are many different types of link functions we could use, but we will focus today on the logistic link function.



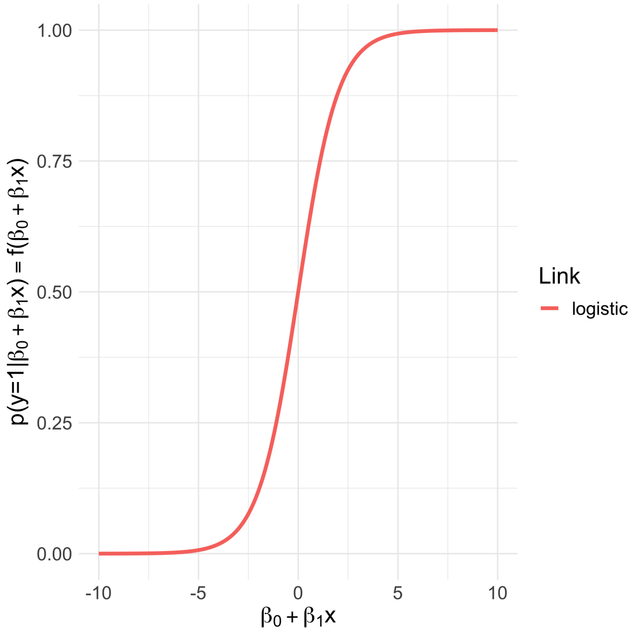
The logistic function

Instead of predicting the outcome directly, we instead predict the probability of being class 1, given the linear combination of predictors, using the logistic link function.

$$p(y = 1|\beta_0 + \beta_1 x) = f(\beta_0 + \beta_1 x)$$

where

$$f(\beta_0 + \beta_1 x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



Logistic function

Transform the function:

$$f(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

$\rightarrow f(x) = \frac{1}{1/e^{\beta_0+\beta_1 x}+1}$

$\rightarrow 1/f(x) = 1/e^{\beta_0+\beta_1 x} + 1$

$\rightarrow 1/f(x) - 1 = 1/e^{\beta_0+\beta_1 x}$

$\rightarrow \frac{1}{1/f(x)-1} = e^{\beta_0+\beta_1 x}$

$\rightarrow \quad \ldots$

$\rightarrow \quad \ldots$

$\rightarrow \frac{f(x)}{1-f(x)} = e^{\beta_0+\beta_1 x}$

$\rightarrow \log_e \frac{f(x)}{1-f(x)} = \beta_0 + \beta_1 x$

The left-hand side $\log_e \frac{f(x)}{1-f(x)}$ is called the log-odds ratio or logit.

The logistic regression model

The fitted model, where $P(Y = 0|X) = 1 - P(Y = 1|X)$, is then written as:

$$\log_e \frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 X$$

▶ Details

Interpretation

- **Linear regression**
  - β_1 gives the average change in Y associated with a one-unit increase in X
- **Logistic regression**
  - Increasing X by one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}
  - However, because the model is not linear in X, β_1 does not correspond to the change in response associated with a one-unit increase in X

Maximum Likelihood Estimation

Given the logistic $p(x_i) = \frac{1}{e^{-(\beta_0+\beta_1 x_i)}+1}$

We choose parameters β_0, β_1 to maximize the likelihood of the data given the model. The likelihood function is

$$l_n(\beta_0, \beta_1) = \prod_{y_i=1,i}^{n} p(x_i) \prod_{y_i=0,i}^{n} (1 - p(x_i)).$$

It is more convenient to maximize the *log-likelihood*:

$$\max_{\beta_0,\beta_1} \log l_n(\beta_0, \beta_1) = \max_{\beta_0,\beta_1} - \sum_{i=1}^{n} \log\left(1 + e^{-(\beta_0+\beta_1 x_i)}\right)$$

Making predictions

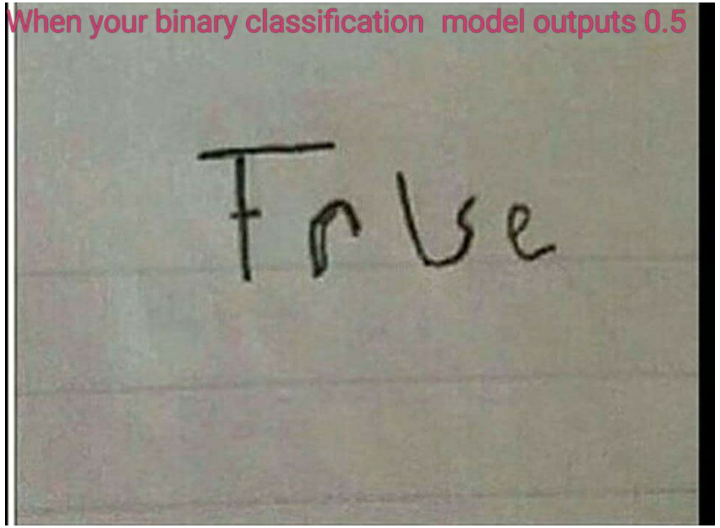
With estimates from the model fit, $\hat{\beta}_0, \hat{\beta}_1$, we can predict the **probability of belonging to class 1** using:

$$p(y = 1|\hat{\beta}_0 + \hat{\beta}_1 x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

This probability can be rounded to 0 or 1 for class prediction.

In `R`, we simply use the `predict()` function.

Of course, probabilities close to 0.5 are hard to classify!



Source: Statistical Statistics Memes

Fitting credit data in R

We use the `glm` function in R to fit a logistic regression model. The `glm` function can support many response types, so we specify `family="binomial"` to let R know that our response is *binary*.

```r
library(broom)
fit <- glm(default~balance,
           data=simcredit, family="binomial")
tidy(fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) -10.7      0.361       -29.5 3.62e-191
## 2 balance       0.00550  0.000220     25.0 1.98e-137
```

Fitting credit data in R

We can use the `predict()` function to predict the probability of default, given credit balance. We then round these probabilities to predict default status.

```
probs <- predict(fit, simcredit ,type="response")
head(probs, 4)
```

```
##            1            2            3            4
## 0.0013056797 0.0021125949 0.0085947405 0.0004344368
```

```
head(round(probs), 4)
```

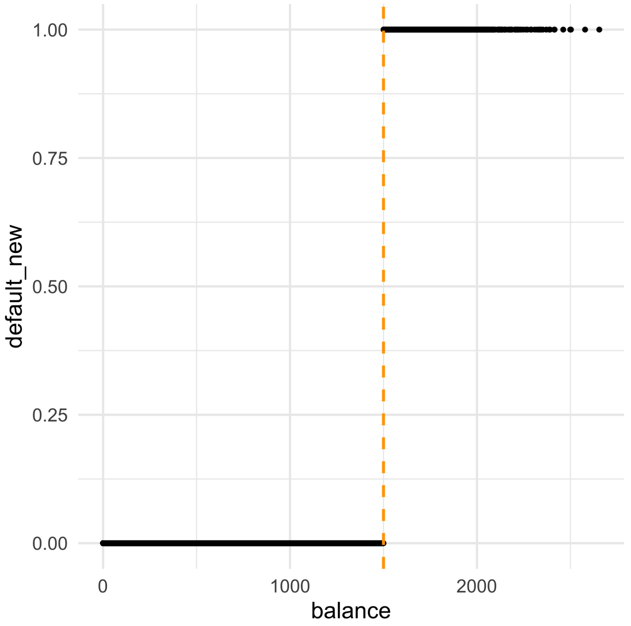
```
## 1 2 3 4
## 0 0 0 0
```

A warning for using GLMs!

When logistic regression fails

Consider the case when the data is *perfectly* separated.

Here, we can see that all balances above $1500 default.



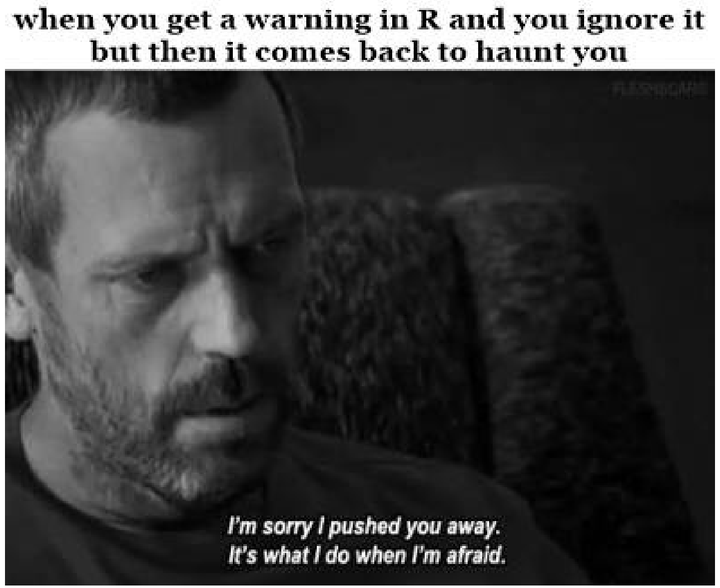
When logistic regression fails

If we fit a `glm` model to this data, the MLE fit will try and fit a step-wise function to this graph, pushing coefficients sizes towards infinity and produce large standard errors. R will warn us that the algorithm does not converge.

```
fit <- glm(default_new~balance,
           data=simcredit, family="binomial")
tidy(fit)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)  -41692.    57581.    -0.724   0.469
2 balance        27.8      38.4      0.724   0.469
```

Take home message - take note of R warnings



Source: R Memes for Statistical Fiends

Linear Discriminant Analysis

Logistic regression involves directly modeling $P(Y = k|X = x)$ using the logistic function. Rounding the probabilities produces class predictions, in two class problems; selecting the class with the highest probability produces class predictions in multi-class problems.

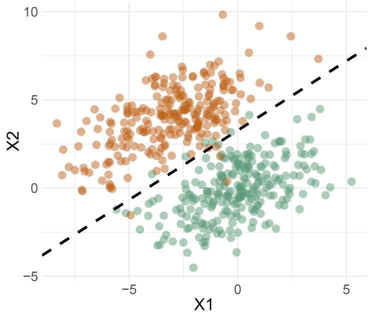
Another approach for building a classification model is linear discriminant analysis. This involves directly estimating the distribution of the predictors, separately for each class.

Compare the pair

Logistic Regression

Goal - directly estimate $P(Y|X)$ (*the dashed line*)

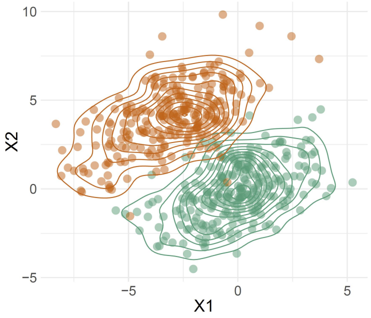
Assumptions - no assumptions on predictor space



Linear Discriminant Analysis

Goal - estimate $P(X|Y)$ (*the contours*) to then deduce $P(Y|X)$

Assumptions - predictors are normally distributed



Assumptions are critical in LDA

Source: Statistical Statistics Memes

Bayes Theorem

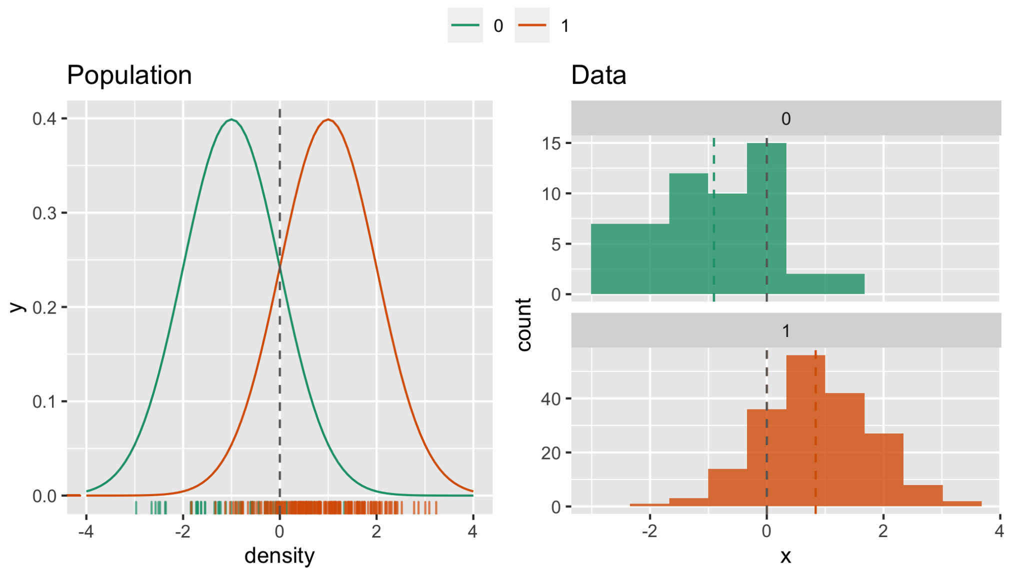
Let $f_k(x)$ be the density function for predictor x for class k. If f is small, the probability that x belongs to class k is small, and conversely if f is large.

Bayes theorem (for K classes) states:

$$P(Y = k|X = x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_k f_k(x)}$$

where $\pi_k = P(Y = k)$ is the prior probability that the observation comes from class k.

LDA with $p = 1$ predictors



LDA with $p = 1$ predictors

We assume $f_k(x)$ is univariate Normal (Gaussian):

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

where μ_k and σ_k^2 are the mean and variance parameters for the kth class. Further assume that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$; then the conditional probabilities are

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

LDA with $p = 1$ predictors

The Bayes classifier is assign new observation $X = x_0$ to the class with the highest $p_k(x_0)$. A simplification of $p_k(x_0)$ yields the discriminant functions:

$$\delta_k(x_0) = x_0 \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$$

and the rule Bayes classifier will assign x_0 to the class with the largest value.

LDA with $p = 1$ predictors

If $K = 2$ and $\pi_1 = \pi_2$, we assign x_0 to class if

$$\delta_1(x_0) > \delta_2(x_0)$$

$$x_0 \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi) > x_0 \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi)$$

which simplifies to $x_0 > \frac{\mu_1+\mu_2}{2}$.

This is estimated on the data with $x_0 > \frac{\bar{x}_1+\bar{x}_2}{2}$.

Multivariate LDA

To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution with $E[X] = \mu$ and $\mathrm{Cov}(X) = \Sigma$, we write $X \sim N(\mu, \Sigma)$.

The multivariate normal density function is:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\}$$

with x, μ are p-dimensional vectors, Σ is a $p \times p$ variance-covariance matrix.

Multivariate LDA

The discriminant functions are:

$$\delta_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log(\pi_k)$$

and Bayes classifier is assign a new observation x_0 to the class with the highest $\delta_k(x_0)$.

When $K = 2$ and $\pi_1 = \pi_2$ this reduces to

Assign observation x_0 to class 1 if

$$x_0'\Sigma^{-1}(\mu_1 - \mu_2) > \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$

Multivariate LDA

Discriminant space: a benefit of LDA is that it provides a low-dimensional projection of the p-dimensional space, where the groups are the most separated. For $K = 2$, this is

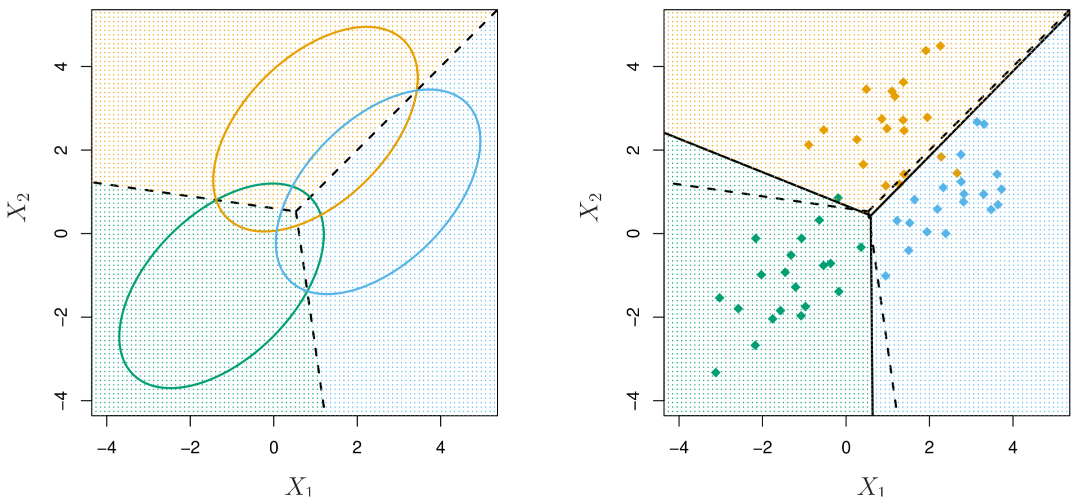
$$\Sigma^{-1}(\mu_1 - \mu_2)$$

For $K > 2$, the discriminant space is found be taking an eigen-decomposition of $\Sigma^{-1}\Sigma_B$, where

$$\Sigma_B = \frac{1}{K}\sum_{i=1}^{K}(\mu_i - \mu)(\mu_i - \mu)'$$

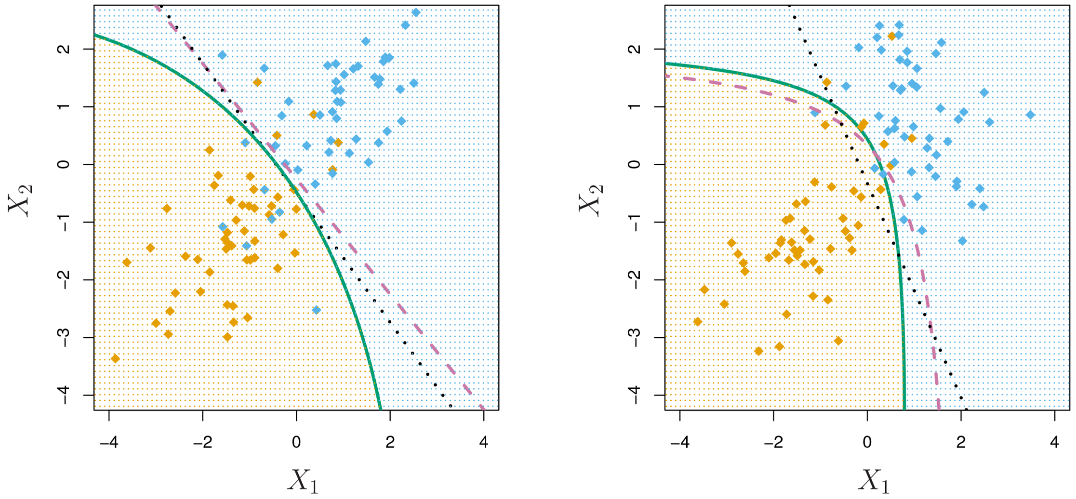
Multivariate LDA

The dashed lines are the Bayes decision boundaries. Ellipses that contain 95% of the probability for each of the three classes are shown. Solid line corresponds to the class boundaries from the LDA model fit to the sample.



Quadratic DA (QDA)

A quadratic boundary is obtained by relaxing the assumption of equal variance-covariance, and assume that
$\Sigma_k \neq \Sigma_l, \quad k \neq l, k, l = 1, \ldots, K$



🧑‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.