ETC3250: Classification

Semester 1, 2019

Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 3 (a)

Outline

- Classification
  - Categorical response

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

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- Historical document author: $Y \in \{Austen, Dickens, Imitator\}$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

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.

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?

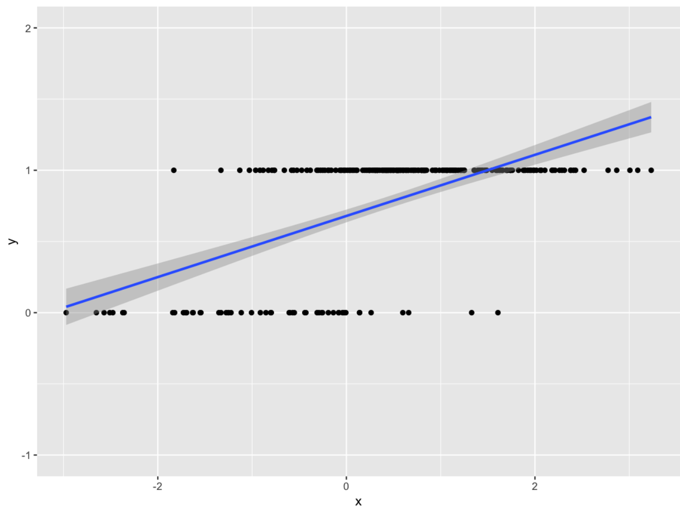
An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.

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.

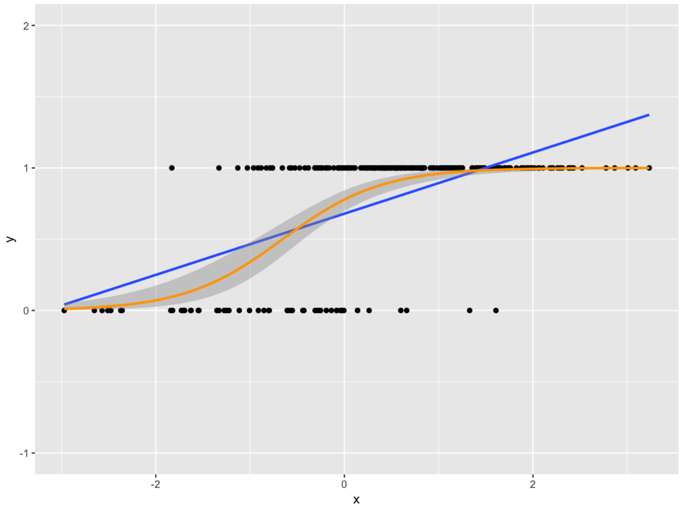
Outline

- Classification
  - Categorical response
  - Why not linear reg?



Outline

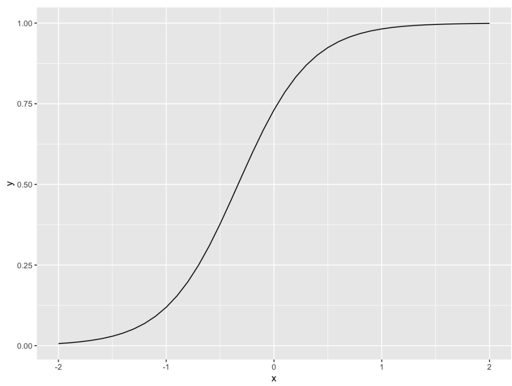
- Classification
- Logistic regression



Outline

- Classification
- Logistic regression
  - Logistic function

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



Transform the function:

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

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

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

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

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

$\rightarrow \ldots$

$\rightarrow \ldots$

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

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

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

The fitted model is then written as:

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

and then

$$P(Y=0|X) = 1 - P(Y=1|X)$$

Multiple categories: This formula can be extended to more than binary response variables. Writing the equation is not simple, but follows from the above, extending it to provide probabilities for each level/category. The sum of all probabilities is 1.

Outline

- Classification
- Logistic regression
  - Logistic function
  - Logistic model - GLM
  - 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

Outline

- Classification
- Logistic regression
  - Logistic function
  - Logistic model - GLM
  - Interpretation
  - Estimation

Maximum Likelihood Estimation

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

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*:

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

Outline

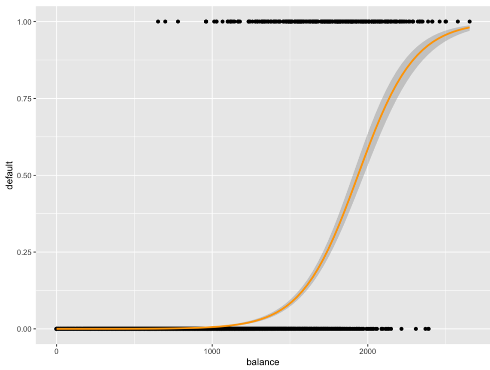
With estimates from the model fit, $\hat{\beta}_0, \hat{\beta}_1$, predict the response using:

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

It is a proportion, but can be rounded to 0 or 1 for class prediction.

Outline

Simulated data to predict which customers will default on their credit card debt.



Outline


```
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
```

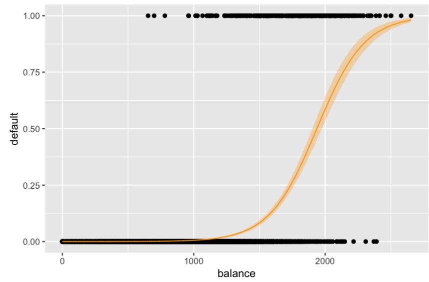

Outline


```
glance(fit)
```

```
## # A tibble: 1 x 7
##   null.deviance df.null logLik   AIC   BIC deviance df.residual
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int>
## 1         2921.    9999  -798. 1600. 1615.    1596.        9998
```

Outline


```
simcredit_fit <- augment(fit, simcredit, type.predict="response")
ggplot(simcredit_fit, aes(x=balance, y=default_bin)) +
  geom_point() +
  geom_ribbon(aes(ymin=.fitted-2*.se.fit, ymax=.fitted+2*.se.fit),
              fill = "orange", alpha=0.3) +
  geom_line(aes(y=.fitted), colour = "orange") +
  ylab("default")
```



Outline

Logistic regression involves directly modeling $Pr(Y = k|X = x)$ 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.

Outline

- Classification
- Logistic regression
- Linear discriminant analysis
  - 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:

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

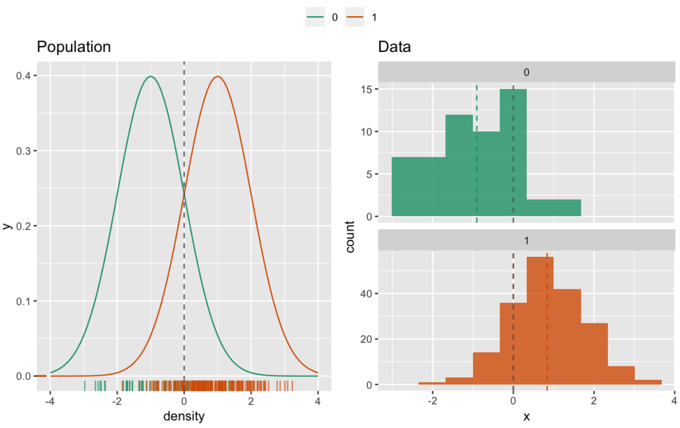
where

$$\pi_k = \Pr(Y = k)$$

is the prior probability that the observation comes from class, *k*.

Outline

- Classification
- Logistic regression
- Linear discriminant analysis
  - Bayes theorem
  - When $p = 1$



Outline

- Classification
- Logistic regression
- Linear discriminant analysis
  - Bayes theorem
  - When $p = 1$

We assume $f_k(x)$ is Normal or 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

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

Outline

- Classification
- Logistic regression
- Linear discriminant analysis
  - Bayes theorem
  - When $p = 1$

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.

Outline

- Classification
- Logistic regression
- Linear discriminant analysis
  - Bayes theorem
  - When $p = 1$

When $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier is :

Assign x_0 to class 1, 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 $x_0 > \frac{\bar{x}_1 + \bar{x}_2}{2}$.

Outline

- Classification
- Logistic regression
- Linear discriminant analysis
  - Bayes theorem
  - When $p = 1$
  - Multivariate

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.

Outline

The discriminant functions are:

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

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

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

Outline

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

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

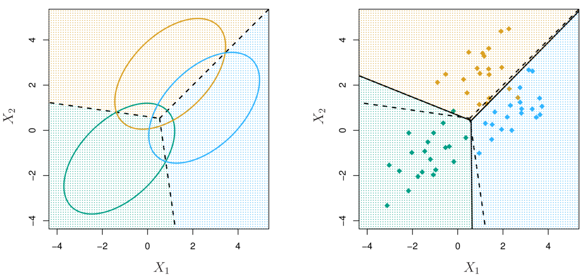
For $K > 2$, the discriminant space is found be taking an eigen-decomposition of $\Sigma^{-1}\Sigma_B$, where

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

Outline

- Classification
- Logistic regression
- Linear discriminant analysis
  - Bayes theorem
  - When $p=1$
  - Multivariate

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.

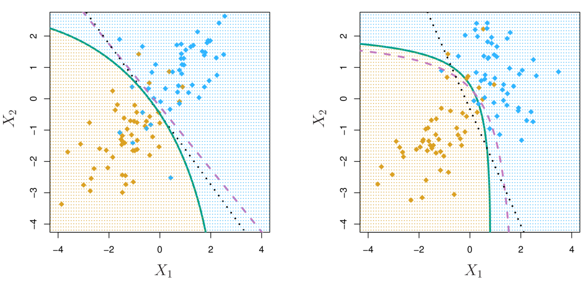


(Chapter4/4.6.pdf)

Outline

- Classification
- Logistic regression
- Linear discriminant analysis
  - Bayes theorem
  - When $p=1$
  - Multivariate
  - Quadratic

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



(Chapter4/4.9.pdf)

👩‍💻 Made by a human with a computer

Slides at https://monba.dicook.org.

Code and data at https://github.com/dicook/Business_Analytics.

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