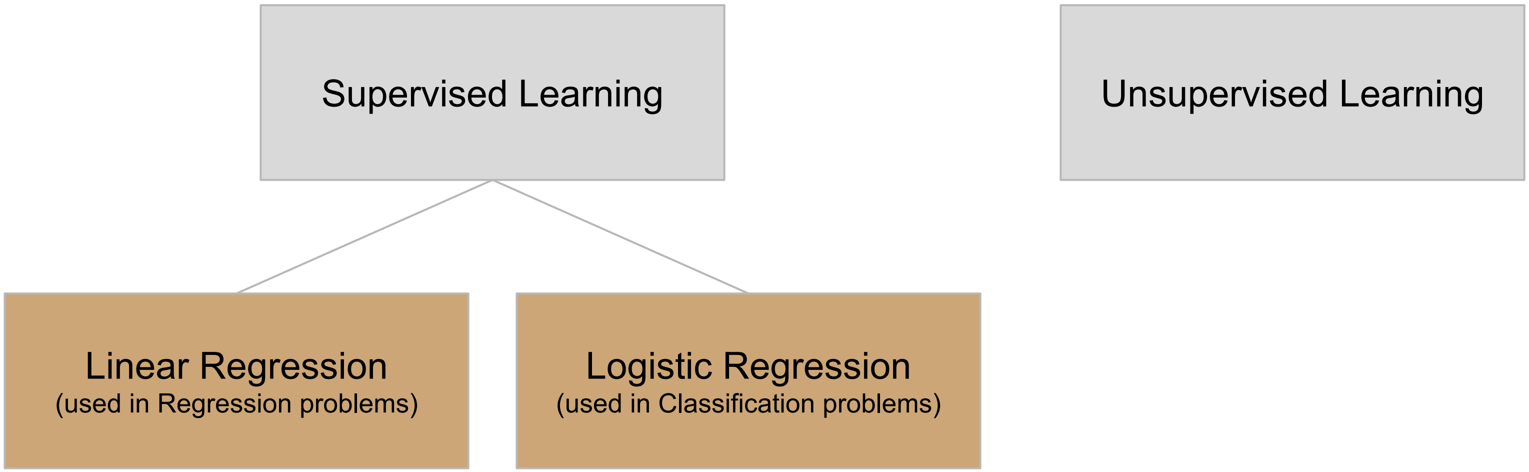
Machine Learning & Data Mining @ NuIEEE

Linear Regression and Logistic Regression

Miguel Sandim & Paula Fortuna

24-02-2016

0 - Intro


Supervised Learning
Unsupervised Learning
Linear Regression
(used in Regression problems)
Logistic Regression
(used in Classification problems)

1 - Linear Regression

Simple Linear Regression

- Only one independent variable:
- Model formula:

$$Y \approx \beta_0 + \beta_1 X$$

- Goal: predict dependent variable values given the attributes.

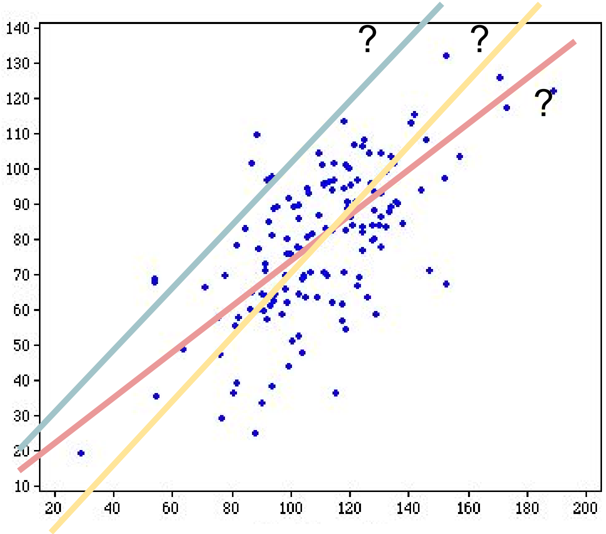
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

1 - Linear Regression

Simple Linear Regression

- Problem: How to discover the coefficients?

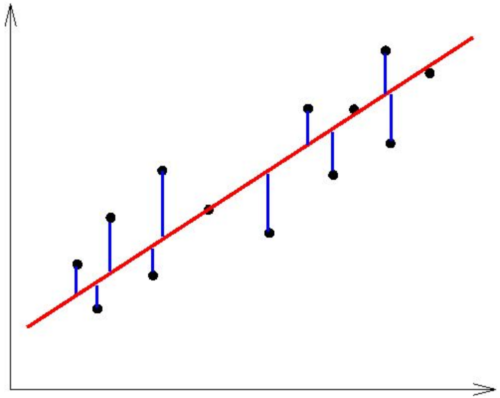
$$Y \approx \beta_0 + \beta_1 X$$



1 - Linear Regression

Simple Linear Regression

- Problem: How to discover the coefficients?
  - Minimize the distance from the line to each point;
  - Least Squares Fit;
  - Minimize Residual Sum of Squares:



$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

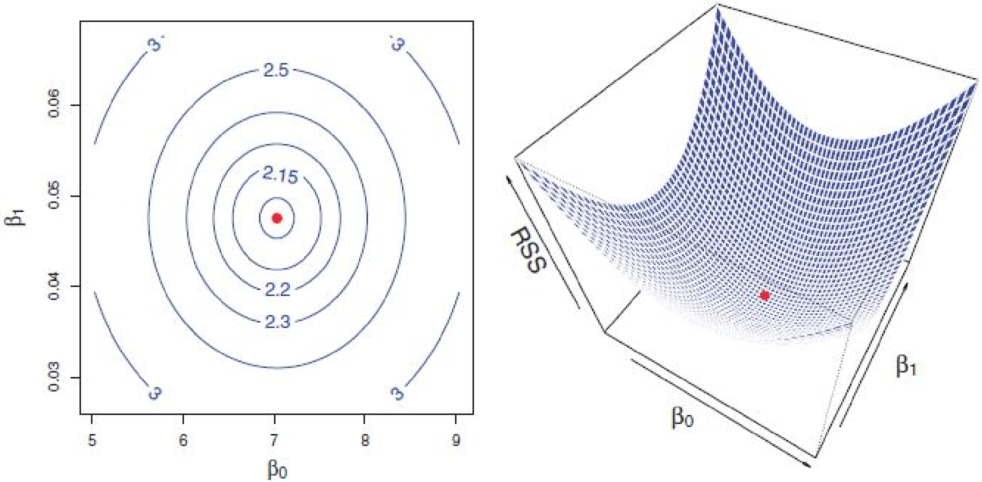
$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

1 - Linear Regression

Simple Linear Regression

- Problem: How to discover the coefficients?

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$



1 - Linear Regression

Simple Linear Regression

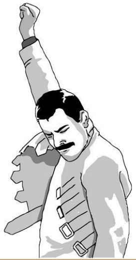
- We are testing an hypothesis:
    - Our sample is from a population that can be described by the model found.

$$H_0 : \text{There is no relationship between } X \text{ and } Y \longrightarrow H_0 : \beta_1 = 0$$

$$H_a : \text{There is some relationship between } X \text{ and } Y \longrightarrow H_a : \beta_1 \neq 0.$$

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \longrightarrow p\text{-value.}$$

$p<0.05$



**p-value* is the probability for the null hypothesis to be true

1 - Linear Regression

Simple Linear Regression

- How to assess the Accuracy of the Model?

    - the residual standard error (RSE)
    - R^2

Linear Regression

Simple Linear Regression

- How to assess the Accuracy of the Model?
  - The **RSE** is an estimate of the standard deviation of the error. It is the average amount that the response will deviate from the true regression line.
  - The RSE is considered a measure of the lack of fit of the model.

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Linear Regression

Simple Linear Regression

- How to assess the Accuracy of the Model?
    - R^2 - measures the proportion of variance in Y that can be explained using X

Residuals Sum of Squares:
Errors predicting with the model

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots$$

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

Total Sum of Squares:
Errors predicting with the average

$$R^2 = r^2$$

Linear Regression

Simple Linear Regression - R

Linear Regression

Multiple Linear Regression

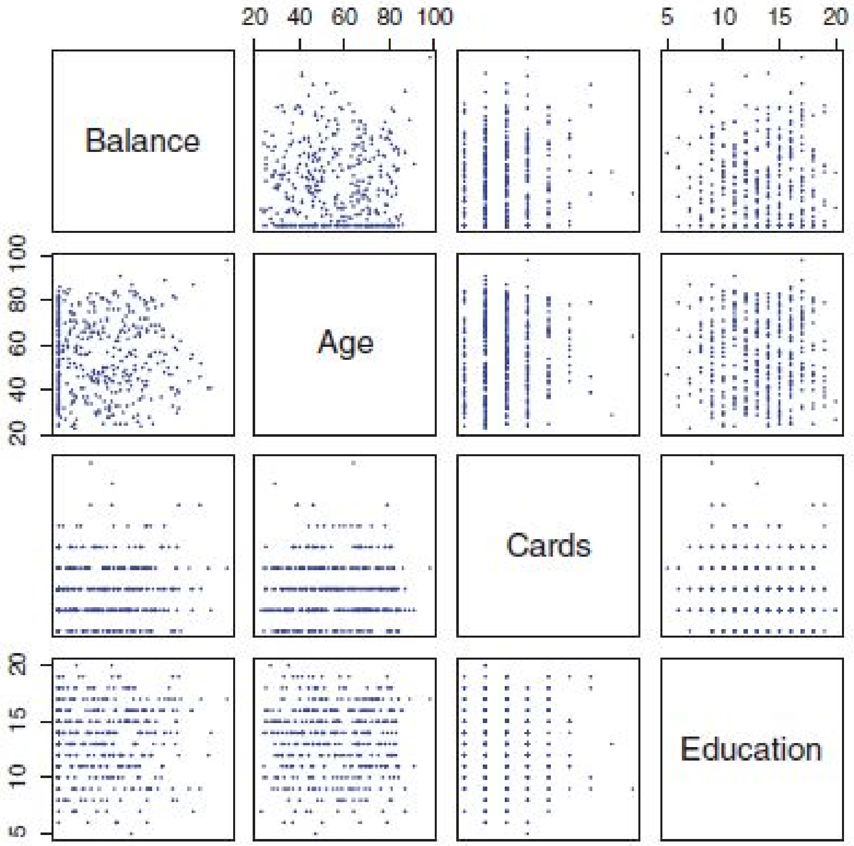
- Generalization of the simple linear regression.
- Difference: computes *F* instead of *t*

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)}$$

Linear Regression

Some questions

- Qualitative predictors
  - Create dummy variable (-1,1)



Linear Regression

Assumptions:

- Additive
    - Instead of an additive effect we could have an interaction between both variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

$$\Downarrow$$

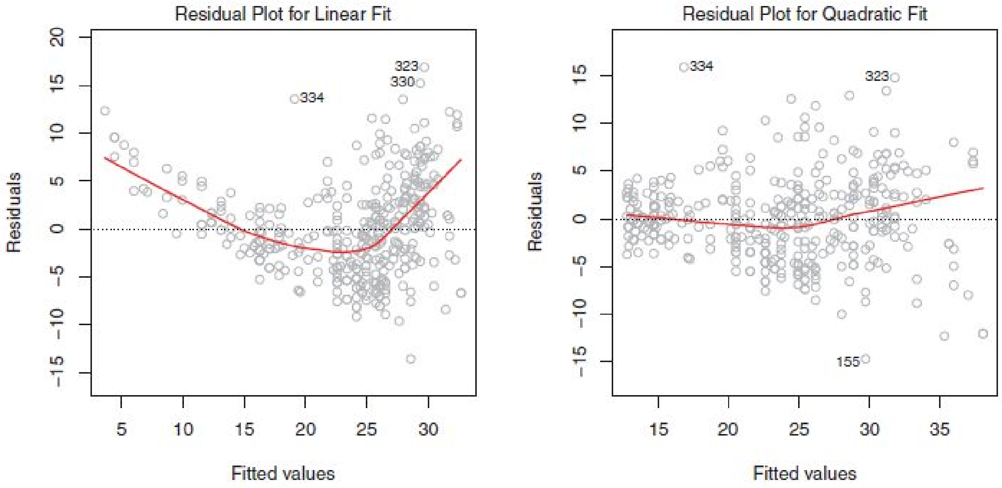
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

- Non-linear relation
    - Polynomial regression

Linear Regression

What do I need to take into account? Potential Problems:

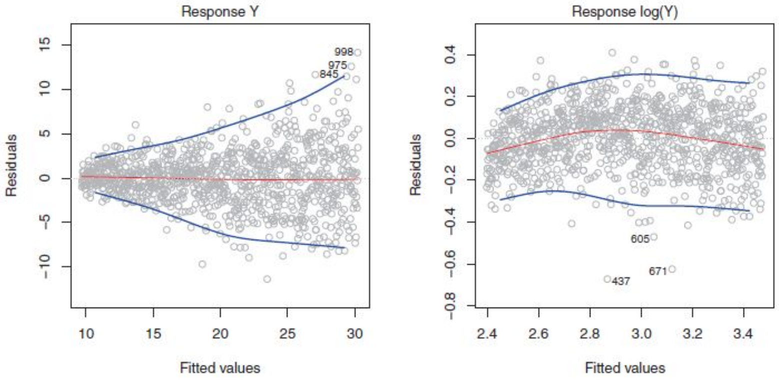
- What if the relationship between predictor and variable is non-linear?
  - Check the residuals plots



Linear Regression

What do I need to take into account? Potential Problems:

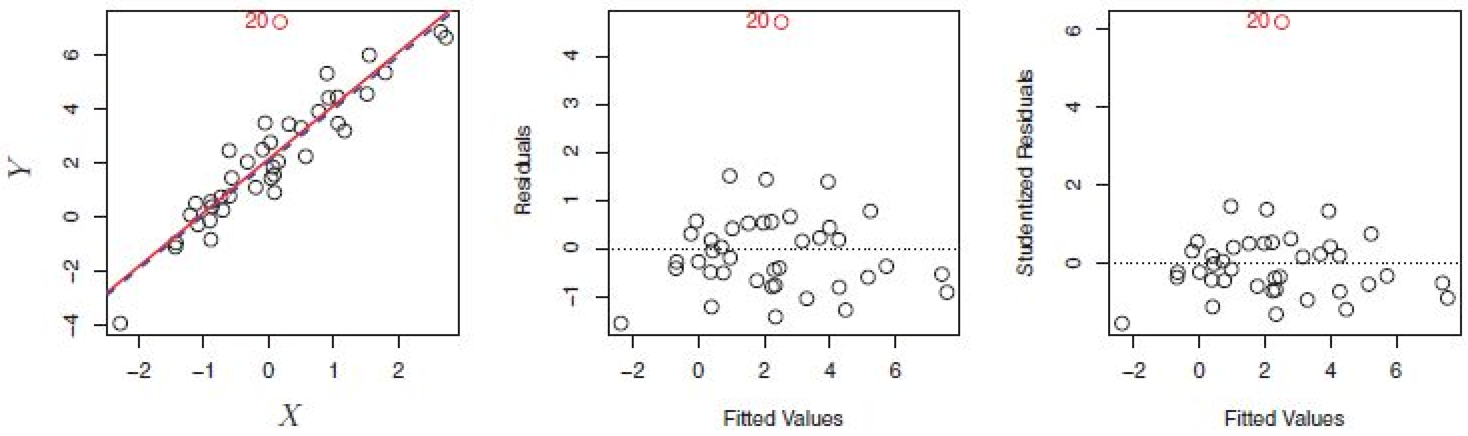
- Non-constant variance of error terms.
  - assumption of the linear regression model is that the error terms have a constant variance (homoscedasticity)
  - transform function -> apply sqrt or log



Linear Regression

What do I need to take into account? Potential Problems:

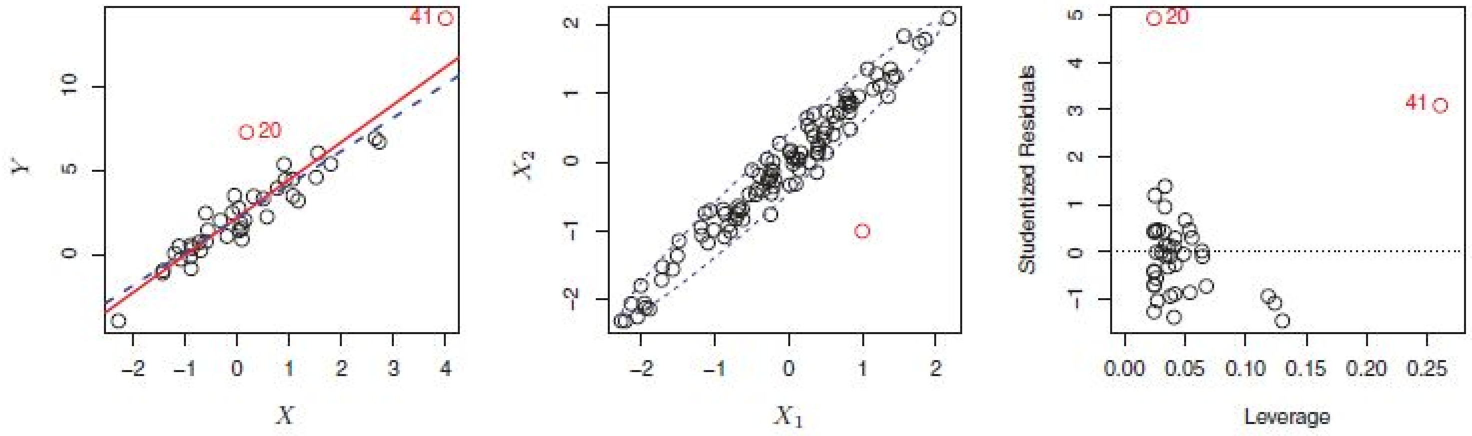
- Outliers



Linear Regression

What do I need to take into account? Potential Problems:

- High-leverage points.



Linear Regression

What do I need to take into account? Potential Problems:

- Collinearity
  - The larger the set of predictor the harder it is to validate our hypothesis (p value)
  - The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.
  - Check correlation matrix to see if the attributes are correlated;
  - Multicollinearity - variance inflation factor (VIF).

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

  - Solution - delete redundant variables from the model

Linear Regression

When to use?

- Quantitative Dependent Variable
- Difference between correlation and regression;

2 - Logistic Regression

- When the variable we're trying to predict is qualitative, rather than quantitative, we have a classification problem.

- We can interpret the output variable as the probability of a given example belonging to a specific class (SPAM, specific disease, etc...).

- Linear regression is no longer a viable solution, since the output can fall out of the interval [0 ,1].

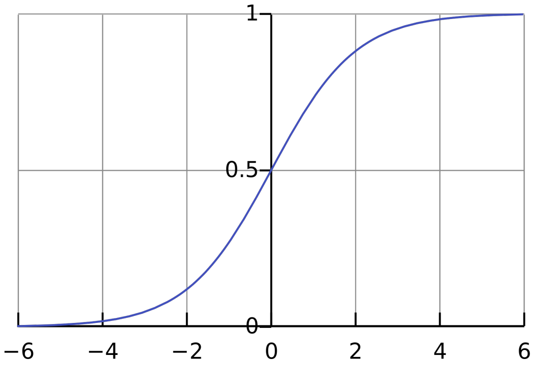
2.1 - How to model this problem?

- We need a mathematical function that outputs values between [0, 1] for all values of its input;

- The chosen function is the logistic function;

- We now model the probability of the output Y to be 1, given an example X ($p(Y=1|X)$) as follows:

$$p(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$



2.1 - How to model this problem?

- With a little manipulation we get: $$\frac{p(Y=1|X)}{1-p(Y=1|X)} = e^{\beta_0+\beta_1 X}$$

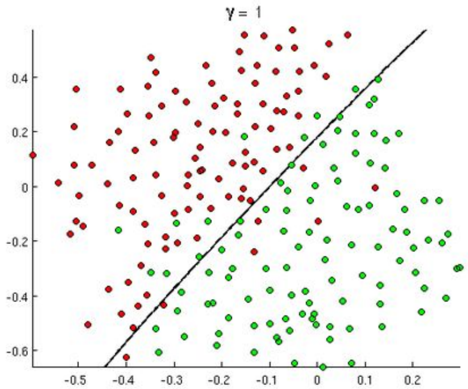
- The left-side of the equation above is called *odd* (varies between 0 and infinity).

$$log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right) = \beta_0 + \beta_1 X$$

- The left-side of equation above is called *log-odd* or *logit*. This can be extended if we have several parameters X, by having several β coefficients. This is the "Binomial" version, since we're trying to predict a binary variable (2 class problem). "Multinomial logistic regression" allows to solve problems with more than 2 classes.

2.2 - How to fit this model to our data and how to predict?

- <u>Fit the model:</u> using the *Maximum Likelihood* method, and maximizing a *likelihood function*:
  - This function takes a number close to 1 for all the examples that are classified as Y = 1, and a number close to 0 for all the examples that are classified as Y = 0.



- <u>Predict:</u> classify an example as belonging to the class if the probability is higher than 0.5. We can adapt the threshold according to the our domain knowledge (for example, if we're trying to detect credit card fraud we might want to reduce the threshold).

2.3 - How to access the accuracy of the model?

- The best way of assessing logistic regression's accuracy (or another classification algorithm) is by having a test set;

- There are several ways of analyzing the performance of a classifier:
    - Overall error rate: $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$
    - Confusion matrix
    - ROC and AUC

2.3 - How to access the accuracy of the model?

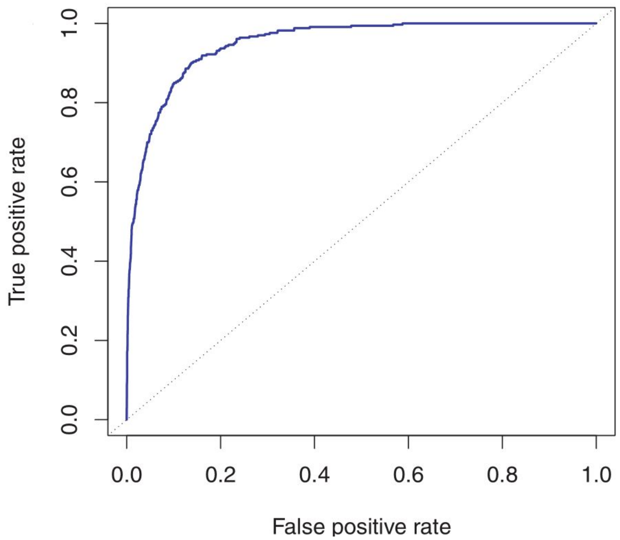
- Confusion Matrix

		Actual Class	
		p	n
Predicted Class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Totals:		P	N

- Sensitivity = TP/P → Measures how well our algorithm is good detecting the positive examples
- Specificity = TN/N → Measures how well our algorithm is good detecting the negative examples

2.3 - How to access the accuracy of the model?

- ROC and AUC



- The ROC curve shows the true positive rate and false positive rate for several possible thresholds.
- The area under the ROC curve is called AUC. The AUC measures the performance of a classifier summarized over all the thresholds.

3. Other resources

- Courses
    - “Machine Learning” by Andrew Ng - Coursera;
    - “Machine Learning” by Pedro Domingos - Coursera;
    - “The Analytics Edge” by MIT - edX.

- Data Science Bible (Trello)

- Books
    - “Introduction to Statistical Learning with applications in R” (where most of the contents of this presentation are available);
    - “Applied Predictive Modelling”.