Advanced Data Modelling

MA4128 2016 Week 2

Kevin O'Brien

Kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Spring Semester 2016

Hypothesis Testing

Recall from Hypothesis Testing: the inferential step to conclude that the null hypothesis is false goes as follows:
The data (or data more extreme) are very unlikely given that the null hypothesis is true. This means that:

(1) a very unlikely event occurred or

(2) the null hypothesis is false.

The inference usually made is that the null hypothesis is false. Importantly it doesnt prove the null hypothesis to be false.

Type I and II errors

There are two kinds of errors that can be made in hypothesis testing:

(1) a true null hypothesis can be incorrectly rejected

(2) a false null hypothesis can fail to be rejected.

The former error is called a ***Type I error*** and the latter error is called a ***Type II error***.

The probability of Type I error is always equal to the level of significance α (alpha) that is used as the standard for rejecting the null hypothesis .

Type II Error

- The probability of a Type II error is designated by the Greek letter beta (β).
- A Type II error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost.
- It is not an error in the sense that an incorrect conclusion was drawn since no conclusion is drawn when the null hypothesis is not rejected.

Hypothesi Testing

Statistical Power

- The power of any test of statistical significance is defined as the probability that it will reject a false null hypothesis.
- Statistical power is inversely related to beta or the probability of making a Type II error.
- In short, power $= 1 - \beta$.

Types of Error

- A Type I error, on the other hand, is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true.
- Therefore, Type I errors are generally considered more serious than Type II errors.
- The probability of a Type I error (α) is set by the experimenter.
- There is a trade-off between Type I and Type II errors. The more an experimenter protects himself or herself against Type I errors by choosing a low level, the greater the chance of a Type II error.

Types of Error

- Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected.
- However, it increases the chance that a false null hypothesis will not be rejected, thus increasing the likelihood of Type II error.
- The Type I error rate is almost always set at 0.05 or at 0.01, the latter being more conservative since it requires stronger evidence to reject the null hypothesis at the 0.01 level then at the 0.05 level.
- **Important**In this module, the significance level α can be assumed to be 0.05, unless explicitly stated otherwise.

Type I and II errors

These two types of errors are defined in the table below.

	True State: H0 True	True State: H0 False
Decision: Reject H0	Type I error	Correct
Decision: Do not Reject H0	Correct	Type II error

Binary Classification

- Statistical classification in general is one of the problems studied in computer science, in order to automatically learn classification systems.
- Some methods suitable for learning binary classifiers include the decision trees, Bayesian networks, support vector machines, neural networks, probit regression, and logit regression.

Binary Classification

- **Binary** or **Binomial classification** is the task of classifying the elements of a given set into two groups on the basis of a classification rule.
- Algorithms for Binary Classification Include:
  1. Logistic Regression
  2. Support Vector Machines

Binary Classification

. Some typical binary classification tasks are:

- medical testing to determine if a patient has certain disease or not the classification property is the presence of the disease;
- A "**pass or fail**" test method or quality control in factories; i.e. deciding if a specification has or has not been met: a Go/no go classification.
- An item may have a qualitative property; it does or does not have a specified characteristic
- information retrieval, namely deciding whether a page or an article should be in the result set of a search or not the classification property is the relevance of the article, or the usefulness to the user.

Medical testing example

- True positive = Sick people correctly diagnosed as sick
- False positive= Healthy people incorrectly identified as sick
- True negative = Healthy people correctly identified as healthy
- False negative = Sick people incorrectly identified as healthy.

Accuracy Rate

The accuracy rate calculates the proportion ofobservations being allocated to the **correct** group by the predictive model. It is calculated as follows:

$$\frac{\text{Number of Correct Classifications}}{\text{Total Number of Classifications}}$$

$$= \frac{TP + TN}{TP + FP + TN + FN}$$

Misclassification Rate

The misclassification rate calculates the proportion ofobservations being allocated to the **incorrect** group by the predictive model. It is calculated as follows:

$$\frac{\text{Number of Incorrect Classifications}}{\text{Total Number of Classifications}}$$

$$= \frac{FP + FN}{TP + FP + TN + FN}$$

Binary Classification

Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function:

- **Sensitivity** (also called the true positive rate, or the recall in some fields) measures the proportion of positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).
- **Specificity** (also called the true negative rate) measures the proportion of negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition). Thus sensitivity quantifies the avoiding of false negatives, as specificity does for false positives

Binary Classification

Recall the possible outcomes of a hypothesis test procedure. In particular recall the two important types of error. Importantly the binary classification prediction procedure can yield wrong predictions.

	Null hypothesis (H_0) true	Null hypothesis (H_0) false
Reject null hypothesis	**Type I error** False positive	**Correct outcome** True positive
Fail to reject null hypothesis	**Correct outcome** True negative	**Type II error** False negative

Accuracy, Precision and Recall

Confusion Matrix

- A confusion matrix is a table with two rows and two columns that reports the number of false positives, false negatives, true positives, and true negatives.
- This allows more detailed analysis than mere proportion of correct guesses (accuracy).
- Accuracy is not a reliable metric for the real performance of a classifier, because it will yield misleading results if the data set is unbalanced (that is, when the number of samples in different classes vary greatly).

Accuracy, Precision and Recall

	Predicted Negative	Predicted Positive
Observed Negative	True Negative (TN)	False Positive (FP)
Observed Positive	False Negative (FN)	True Positive (TP)

(Notice that "Negative" precedes "Positive")

Precision and Recall

- **Precision** is the number of correct positive results divided by the number of ***predicted positive*** results.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall** is the number of correct positive results divided by the number of ***actual positive*** results.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Accuracy, Precision and Recall

Important metrics for determining how usefulness of the prediction procedure are : **Accuracy**, **Recall** and **Precision**.

Accuracy, Precision and Recall are defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Accuracy, Precision and Recall

Another measure is the F-measure (or F-score), which is computed as

$$F = 2 \cdot \left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \right)$$

Questions

	Predicted Negative	Predicted Positive
Negative Cases	TN: 9,700	FP: 165
Positive Cases	FN: 35	TP: 100

Accuracy, Precision and Recall

With reference to the table on the previous slide, compute each of the following appraisal metrics.

a. Accuracy

b. Precision

c. Recall

d. F measure

Accuracy, Precision and Recall

- Why is the accuracy value so high?
- Why is the F-measure so low?
  - This is the **class-imbalance** problem: more "negative" outcomes which skews the statistic, but these outcomes are the least relevant.
  - The F-measure disregards the irrelevant "true negatives, and concentrates on the more relevant potential outcomes.

The F-score

- The F-score or F-measure is a single measure of a classification procedure's usefulness.
- The F-score considers both the ***Precision*** and the ***Recall*** of the procedure to compute the score.
- The higher the F-score, the better the predictive power of the classification procedure.
- A score of 1 means the classification procedure is perfect. The lowest possible F-score is 0.

$$0 \leq F \leq 1$$

The F-score

The F-score is the Harmonic mean of Precision and Recall.

$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

Alternatively

$$F = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

Class Imbalance

Class Imbalance Problem

- It is the problem in machine learning where the total number of one class of data (**positive**) is far less than the total number of another class of data (**negative**).
- This problem is extremely common in practice and can be observed in various disciplines including fraud detection, anomaly detection, medical diagnosis, oil spillage detection, facial recognition, etc.
- Most machine learning algorithms and works best when the number of instances of each classes are roughly equal.
- When the number of instances of one class far exceeds the other, problems arise (*Type I and Type II error*).

Number of cases: **100,000**

	Predicted Negative	Predicted Positive
Observed Negative	True Negative 97750	False Positive 150
Observed Positive	False Negative 330	True Positive 1770

- **Accuracy** = 0.9952
- **Recall** = 0.8428
- **Precision** = 0.9218

The F-score

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F = 2 \times \frac{0.9218 \times 0.8428}{0.9218 + 0.8428} = 2 \times \left(\frac{0.9218 \times 0.8428}{0.9218 + 0.8428} \right)$$

$$F = 2 \times \left(\frac{0.7770}{1.7646} \right) = 2 \times 0.4402$$

$$F = 0.8804$$