MODULE 1 UNIT 3

Statistical methods

UNIVERSITY OF CAPE TOWN

IN COLLABORATION WITH getsmarter™

Table of contents

Tel: +27 21 447 7565 | **Fax:** +27 21 447 8344
Website: www.getsmarter.com | **Email:** info@getsmarter.com

Learning outcomes:

LO4: Determine the type of learning problem for a given situation.

LO5: Determine the appropriate type of task for a given supervised problem.

LO6: Justify statistical model selection using specified criteria.

1. Introduction

Earlier in this module, you learnt about the different types of statistical learning, namely supervised and unsupervised learning. Within these categories, there are multiple methods that can be applied, depending on the type of problem and type of data. In this course, two methods under each type of learning will be explored.

Supervised learning:

1. Tree-based models
2. Neural networks

Unsupervised learning:

1. K-means clustering
2. Hierarchical clustering

It should be noted that the two supervised methods are not restricted to supervised learning.

In order to select an appropriate method, you'll have to consider the data. First, you need to determine whether it is a supervised or unsupervised learning problem. If it is a supervised learning problem, it can be a classification or regression task depending on the data type of the response variable.

2. Data types

Any given piece of data can take on a value. However, the type of data will determine the kind of values it could potentially have. For this course, there are two types of data you need to be aware of: continuous and discrete data.

Explore further:

Data is often stored in tables. If it is a small data set with a few hundred rows, a simple Microsoft Excel spreadsheet may suffice. However, when there are millions of rows, a database is required.

There are two kinds of databases: SQL and noSQL databases. If you are interested, read more about SQL and noSQL databases and how to choose between them.

2.1 Continuous

Continuous data can take on any value in a continuum or set of values (e.g. any value between zero and one). Such observations are often referred to as "quantitative" measurements, where "quanta" reflects the property that such a measurement concerns a recorded amount along some dimension, such as kg or cm^3. Table 1 shows some more examples of continuous data.

Table 1: Examples of continuous data.

Example	Value
Sales figures	R5,356.46
Profits or expenses	−R3,005,741.51
Geographical coordinates	−33.9310765, 18.3919993
Average number of sales, visits, or clicks	53.324
Time taken to package a product	5.23 minutes

Can you think of examples of continuous or quantitative measurements? What data is continuous in your business domain?

2.2 Discrete (Categorical)

Discrete or categorical data can take on one of a finite set of values, with each of the values usually corresponding to some quality or associated category. Consequently, such measurements are often referred to as qualitative data. This kind of data can be numeric or non-numeric. Examples include product types, locations, and units sold.

Some discrete data is a collection of unordered items. For example, different flavours of chocolate – e.g. mint, hazelnut, and caramel – categorise the chocolate into piles. There are many kinds of unordered discrete data that are of interest to business operations, such as those shown in Table 2.

Table 2: Examples of unordered discrete data.

Category	Example values
Products	Shampoo, soap, toothpaste
Advertising campaign running (1) or not running (0)	Data with binary entries (e.g. 0, 1, 0, 0, 1, 1)
Suburbs or towns	Rondebosch, Belville, Milnerton, Paarl

Discrete data can also be an ordered set. Using the chocolate example, the chocolate could be divided by the percentage of cocoa: 45%, 70%, and 85%. Although the percentages are numerical, they are still defined categories. This is an example of ordered discrete data, as the percentage of cocoa can be sorted in ascending or descending order. Other examples of ordered discrete data are given in Table 3.

Table 3: Examples of ordered discrete data.

Category	Example values
T-shirt sizes	Small, medium, large
Months of the year	January, February, March
Education	High school, undergraduate, postgraduate

A common problem with discrete data is that it may not be standardised. For example, consider a retail company that has several stores across South Africa. Each store comes up with its own words for categorising the stock it holds. Some stores group shoes into "sandals", "sneakers", and "ankle boots", while others split the stock into "flip-flops", "closed shoes", and "boots". For the purposes of analysing boot sales, all the variations of boot entries would have to be found, including misspelt or capitalised versions of the word.

Note:

When data is non-numeric, it is often converted into a numeric representation. For example, you may want to know whether a customer upgraded their cell phone account after a call from a salesperson. The answer may be yes or no, but it will be represented or recorded in the data set as 0 for no and 1 for yes.

3. Selecting a statistical method

The approach that is taken to answer questions pertaining to a particular data set depends on a few factors, such as the type of data you have. Is it continuous or discrete? The presence or absence of a response variable is another indicator, determining whether it is a supervised or unsupervised learning problem. The purpose of the test also affects the statistical model used.

Figure 1 outlines how to choose among the models presented in this course, based on some criteria. The sections following that will help you think about how to select a method.

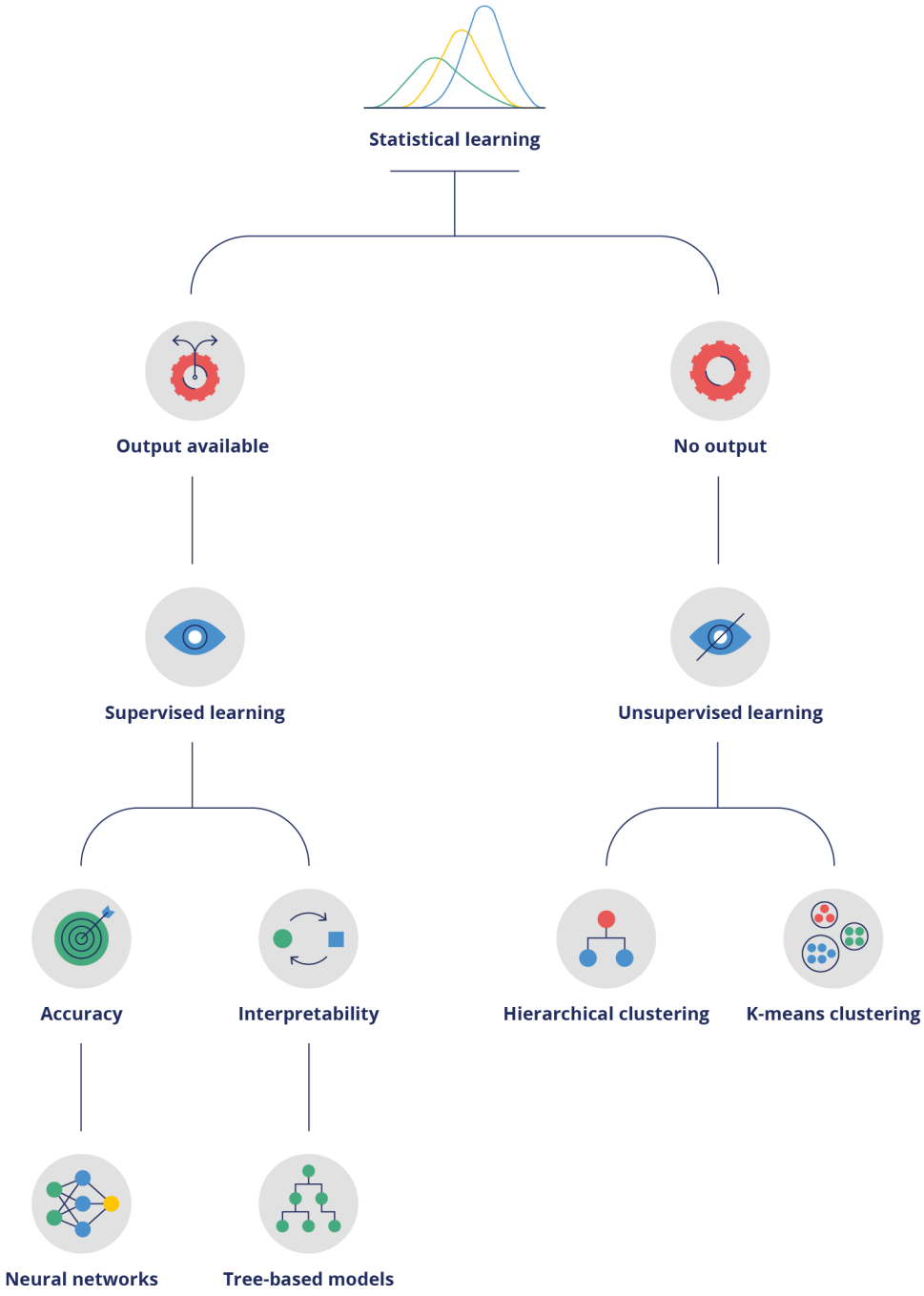

Figure 1: Selecting a statistical method.

3.1 Supervised learning

If the output is known, it is a supervised learning problem. Next, you need to decide which type of task is required (regression or classification), and which model is appropriate based on whether the particular application requires more accuracy or interpretability.

3.1.1 Regression or classification

Within supervised learning, there are two types of tasks, depending on the nature of the response variable. If the response variable is continuous, the task is a regression task, and if it is discrete, the task is a classification task. Most statistical models can be formulated to perform either of the two tasks, depending on the situation. Note that the features in a data set need not all be continuous or discrete; it is only the response variable that affects the task.

3.1.2 Accuracy vs interpretability

Balancing accuracy and interpretability is one of the first decisions you have to make when choosing a particular methodology. Depending on your industry or the purpose of the model, either accuracy or the ability to interpret how it works may be favoured, though this might not always be a clear-cut distinction.

For example, consider the banking and insurance industries, where predictive models can be very helpful. In banking, predictive models can be used for predicting loan defaulters, detecting fraudulent credit card transactions, and algorithmic trading. In insurance, applications include detecting fraudulent claims, performing risk analyses (how likely is it that the customer will claim), and customising products for customers (providing the right amount of insurance for their current situation). However, these industries are subject to stringent regulation and record-keeping practices (Hall, 2016), which means predictive models need to be very interpretable – i.e. it needs to be clear how the model got to its prediction.

In other industries, accuracy might be more important. For example, accuracy trumps interpretability in applications such as spam filters. The need to interpret how the spam filter works is not necessary. It is more important for the spam filter to have good accuracy (Brownlee, 2014).

Kuhn and Johnson (2013:4) warn against favouring interpretability over accuracy when it is not absolutely required. Predictive applications in the medical field should favour accuracy, as the consequences of a poor prediction are usually severe. For example, some patients with both HIV and tuberculosis infections have excessive immune responses against the bacteria that cause tuberculosis. Therefore, a model that predicts who may be at risk in medical situations should rather be accurate than interpretable (Kuhn & Johnson, 2013:4).

Between the two supervised learning models that will be explored in this course, tree-based models tend to be more interpretable and neural networks tend to be more accurate. Tree-based models can be visualised as decision trees, which make them easy to interpret,

even as a non-expert. Neural networks, on the other hand, are black boxes, which impedes interpretation of the results.

3.2 Unsupervised learning

The two unsupervised methods that will be compared in this course are K-means and hierarchical clustering. The first difference is that K-means clustering requires you to choose the number of clusters, while hierarchical clustering does not need upfront assumptions. However, other choices need to be made for hierarchical clustering, such as the type of linkage and the dissimilarity measure (James et al., 2013:400).

If you have a big data set, K-means clustering is preferred, because it is computationally cheaper than hierarchical clustering. However, K-means clustering favours spherical clusters, which could lead to misaligned groupings (Alto, n.d.).

4. Conclusion

The basic mental checklist when assessing which kind model to use starts with asking if it is a supervised or unsupervised problem. Next, if it is supervised, decide whether it is a regression or classification task, and consider whether accuracy or interpretability is more important. If it is an unsupervised learning problem, consider the size of the data set and whether you can choose a good number of clusters.

In this module, you learnt about the types of data and how to select an appropriate method. Continue to the activity to test your knowledge.

5. Bibliography

Alto, V. n.d. *Unsupervised learning: K-means vs hierarchical clustering*. Available: https://towardsdatascience.com/unsupervised-learning-k-means-vs-hierarchical-clustering-5fe2da7c9554 [2019, September 8].

Brownlee, J. 2014. *Model prediction accuracy versus interpretation in machine learning*. Available: https://machinelearningmastery.com/model-prediction-versus-interpretation-in-machine-learning [2019, September 5].

Hall, P. 2016. *Predictive modeling: striking a balance between accuracy and interpretability*. Available: https://www.oreilly.com/ideas/predictive-modeling-striking-a-balance-between-accuracy-and-interpretability [2019, September 5].

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. *An introduction to statistical learning: with applications in R.* New York: Springer.

Kuhn, M. & Johnson, K. 2013. *Applied predictive modeling*. New York: Springer.