MODULE 1 UNIT 2

Introduction to statistical learning

UNIVERSITY OF CAPE TOWN

IN COLLABORATION WITH getsmarter™

Table of contents

Learning outcome:

LO3: Identify supervised and unsupervised statistical learning problems.

1. Introduction

Have you ever wondered if the flow of traffic could be optimised? As cities grow and become more congested, the need for transportation solutions increases. A local South African company, WhereIsMyTransport, collects data about the web of mini-bus taxi routes, including peak usage times, common stopping points, and fares (WhereIsMyTransport, n.d.). This data could be used to predict when taxi drivers need to be in certain areas, or to implement better stopping points. It could also be used to extract the most important features to optimise traffic.

Earlier in this module, you learnt that one of the aims of data science is to find patterns in data. Finding the underlying pattern allows you to make predictions or find features that are useful for constructing predictions. To achieve this, you can use statistical learning – a collection of statistical techniques and mathematical learning paradigms that allow you to extract such patterns.

2. Statistical learning

Statistical learning combines data and statistical models to predict future outcomes, group data into clusters, or find the features that have the greatest influence on a particular outcome.

There are three considerations to keep in mind when you are thinking about using statistical learning:

1. **Data:** Is there data available? Is there enough good data available to answer the questions at hand? Good data is well labelled, clean, and complete, and therefore does not have any missing values or invalid entries. This is the most important component to consider before attempting any statistical learning, as data is an integral part of the process (Chatterjee, 2017).

2. **Patterns:** Is there a pattern that can be extracted from the data with a model? Is it intuitive that a pattern exists? Is there a relationship between the information you have and what you would like to predict? For example, it is intuitive that income, previous loan repayment, age, job security, and disposable income will impact the likelihood that an applicant will default on a loan. Sometimes it is not intuitive that a pattern exists, but it is usually apparent.

3. **Appropriateness:** Is statistical learning an appropriate approach? Statistical learning approximates patterns that cannot be directly observed (Norris, 2018), so if the pattern can be directly observed, there is no need for statistical learning.

Different fields of study use various names for inputs (the values used to make the prediction) and outputs (the value being predicted). In statistical learning, inputs are commonly referred to as predictors or features and outputs are called responses (Hastie, Tibshirani & Friedman 2017:9). Other options are presented in Table 1. You may come across these different terms in various articles or textbooks, so take note of them to avoid getting confused.

Table 1: Alternative names for inputs and outputs.

	Input	**Output**
Classic statistics	Independent variable	Dependent variable
Statistical learning	Predictor or feature	Response

Broadly speaking, there are three learning paradigms within the sphere of statistical learning: supervised, semi-supervised, and unsupervised learning. This course only covers supervised and unsupervised learning.

Definition:

In supervised learning problems, the data includes the predictor variables as well as the output variable. Supervised learning finds the function that most accurately predicts the outcome variable based on the given inputs.

In unsupervised learning problems, the objective is not to predict any particular variable, but rather to find patterns in a collection of observed variables. For example, unsupervised learning can categorise customers into different customer segments based on their preferences.

3. Supervised learning

A project by the African Soil Information Service performed wet and dry chemical tests to measure soil nutrient values all over Africa. Wet tests are more accurate than dry tests, but they are also more expensive. To overcome this problem, they trained a model so that wet-test information could be inferred from the dry-test results. New soil samples can be tested with the more affordable dry tests, and information that would be provided by a wet test can be predicted using the model. In this case, the features are the dry-test results and the responses are the wet-test results (QED, n.d.).

The majority of data science problems are solved using supervised learning techniques such as the soil nutrient example (Brownlee, 2016). It is assumed that there is a functional relationship between the input and output – formally known as the target function. While this function is unknown, it can be approximated using data, a learning algorithm, and a selection of formulas, as shown in Figure 1. These formulas include models such as neural networks and tree-based models (Abu-Mostafa, Magdon-Ismail & Lin, 2012:3).

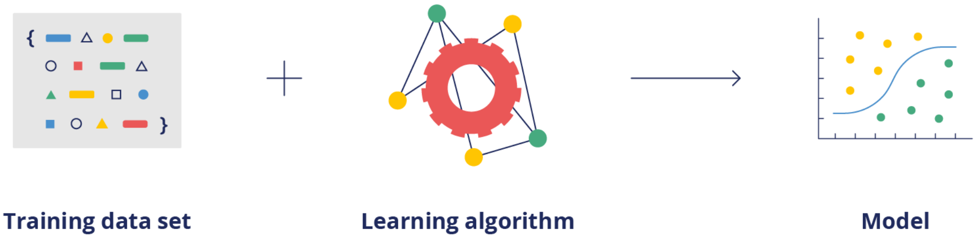

Figure 1: The components of supervised learning. (Adapted from: Abu-Mostafa, Magdon-Ismail & Lin, 2012:4)

Factors to consider when training models include the size and relevance of the data set. In terms of size, a rule of thumb is that the larger the training data set, the higher the accuracy of the predictions.

Explore further:

If you are interested in exploring why more data usually means higher accuracy, watch Professor Yaser Abu-Mostafa's lecture on the feasibility of learning, in which he examines the elements of the Hoeffding's inequality. Watch from 8:04 to 24:30.

If the data used for training a model is not relevant, it will probably give poor predictions. For example, suppose you want to train a model to predict the weather in Cape Town. You have data about wind speed, precipitation, cloud cover, and humidity for the summer months. After training the model, it appears to rarely predict rain, even though Cape Town is notorious for rainy winters. The problem is that the model was not trained on data for the winter months, which means the provided data was not relevant.

4. Unsupervised learning

Suppose a restaurant recommendation website would like to tailor its recommendations for the different customers who use its service. Using a mobile app, they offer discounts at featured restaurants and track who frequents different types of establishments. The data collected includes age, price of the meal purchased, restaurant name, location of the restaurant, and type of cuisine. A clustering algorithm – a typical example of unsupervised learning where there are no output variables – finds eight groups, as shown in Figure 2.

Upon further inspection, the most frequently occurring type of restaurant is identified for each cluster. It appears that young to middle-aged customers prefer niche restaurants, such as raw vegan and poke-bowl restaurants, and tend to be willing to pay more as they get older. Older customers prefer steak and traditional "home-cooked" meals with mid-range cost. In older customers, there are also two groups: one where they prefer expensive, French cuisine, and another where they prefer food from affordable fish-and-chips establishments. It is not clear from this result whether older people are inclined to alternate between expensive and affordable restaurants, or if there are two different groups of people.

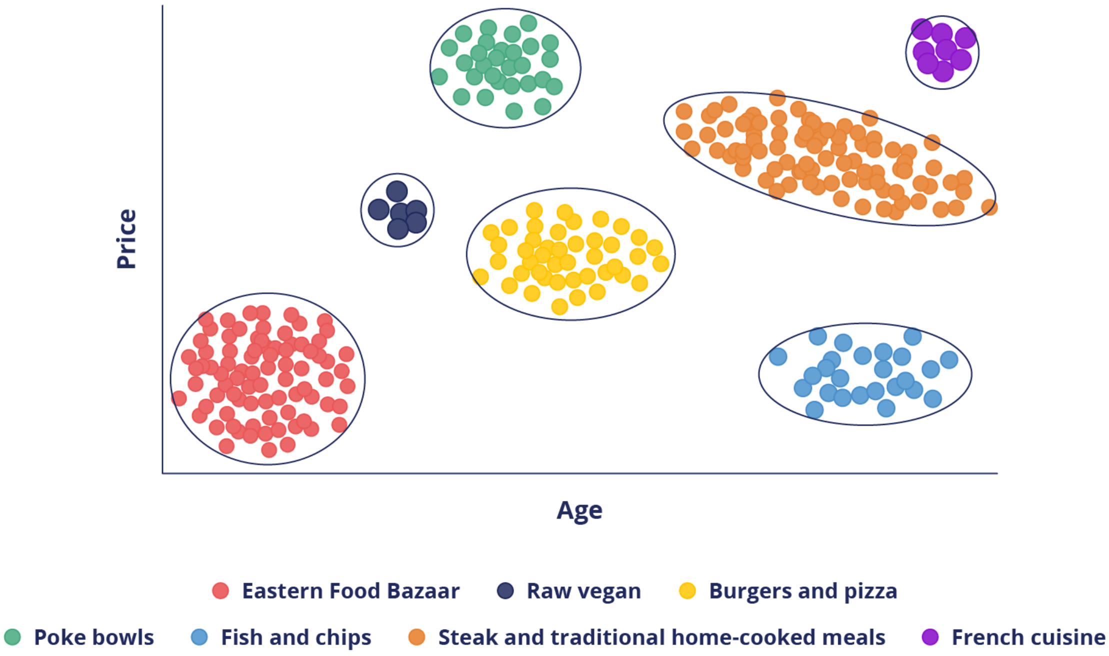

Figure 2: Customer segmentation for types of restaurant by age and price.

When looking at the data set, a single observation would be a coordinate in the feature space, whose elements include all the entries in the row (i.e. all the measurements of each observed predictor). In the restaurant example, the elements include age, price of the meal purchased, restaurant name, location of the restaurant, and type of cuisine, as shown in Table 2.

Table 2: An observation of a customer's purchase.

Person ID	Age	Price of meal	Restaurant name	Restaurant coordinates	Type of cuisine
1926	32	R100	Beef Queen	−31.899473, 26.880517	Burgers

The aim of clustering is to find observations that are similar and group them into clusters. In this way, you can find relationships in the data that are not guided by an output variable. A challenge with clustering methods is that observations may not be easily segmented into distinct groups. So, why use a clustering model if you could just plot the features and visually pick out the groups? In practice, more than two or three features are required for effective clustering. While plotting more than three features is impossible, a clustering model can use a multi-dimensional space to find relationships.

Clustering and other forms of unsupervised learning can also be used to determine features that are important for making accurate predictions. For example, suppose a rental agency used a clustering model to cluster its tenant data, and the results show that the most important predictors for which cluster a new tenant would be placed in are age and pet ownership. According to the clusters, young tenants with pets are more likely to pay their rent on time, so the agency decides to give young applicants with pets a higher preference when processing applications.

5. Conclusion

Statistical learning uses statistics to find patterns that you intuitively think may exist but cannot directly observe. Depending on the data available and the kind of question you have, you may use supervised or unsupervised learning. The key difference is the presence or absence of a response variable.

Think back to the traffic data at the beginning of this lesson. Would predicting the number of people who will use public transport on a weekday using historical sales of bus and train tickets be a supervised or unsupervised learning problem? The number of sales related to the day and time is known, so this would be a supervised problem. Determining which features would help alleviate traffic is an unsupervised problem. By clustering the data, the most important features related to reducing traffic would emerge.

Continue to the next unit to explore data types and the selection of a statistical method.

6. Bibliography

Abu-Mostafa, Y. S., Magdon-Ismail, M. & Lin, H. 2012. *Learning from data: a short course*. United States: AMLBook.

Brownlee, J. 2016. *Supervised and unsupervised machine learning algorithms*. Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms [2019, September 6].

Chatterjee, S. 2017. *Good data and machine learning*. Available: https://towardsdatascience.com/data-correlation-can-make-or-break-your-machine-learning-project-82ee11039cc9 [2019, September 08].

Hakutizwi, B. 2018. *Farming innovations changing the South African agricultural landscape*. Available: https://www.bizcommunity.com/Article/196/358/180373.html [2019, September 15].

Hastie, T., Tibshirani, R. & Friedman, J. 2017. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Corrected 12th printing, January 2017. New York, NY: Springer.

Norris, S. 2018. *How to define a machine learning problem like a detective*. Available: https://opendatascience.com/how-to-define-a-machine-learning-problem-like-a-detective [2019, September 4].

QED. n.d. *AfSIS soil chemistry data – tutorial*. Available: https://github.com/qedsoftware/afsis-soil-chem-tutorial [2019, September 16].

WhereIsMyTransport. n.d. *Collecting public transport data in Gauteng: one of Africa's largest urban regions*. Available: https://www.whereismytransport.com/case-studies/public-transport-data-gauteng [2019, September 15].

Wolinsky, H. 2015. *How big data is revolutionizing farming*. Available: https://www.chicagobusiness.com/article/20150430/ISSUE01/150429825/how-monsanto-s-640-labs-is-harvesting-big-data-to-boost-farm-productivity [2019, September 15].