MODULE 1 UNIT 2
Notes Interactive Infographic 1 Transcript

IN COLLABORATION WITH getsmarter™

The transformation of data

This interactive infographic will examine some practical examples of the transformation of regression and classification data. It will explore this process through all the steps, including importing and cleaning the data, selecting the most significant variables, and fitting the chosen model onto the data to enable decision-making.

Regression

In this example, a regression model is fitted onto a data set to illustrate how raw data is imported, cleaned, processed, and visualised to inform a decision. The problem that the machine learning model should solve is to determine the relationship between wages and the other variables included in the data set. The available data set includes individuals' wages, levels of experience, and heights. Regression will be the most appropriate model to fit onto this data set, since the goal is to show the relationship between numerical variables.

1. Import data

With both regression and classification, the first step is to import and view the data set in the form of a data table. In this example, there are four entries listed in a table, each entry showing the wages, levels of experience, and heights of the respective individuals.

As the data is imported and viewed, it is important to clean the data and identify any issues with it. The issues seen in this data set include a duplicate entry, and a missing entry regarding the height of an individual.

2. Clean data

After the duplicate entry and missing data have been removed, the data is clean and ready to use.

3. Variable selection

Explanatory variables, or independent variables, are identified from the imported data. This means that the variables that are not affected by any of the other variables are identified and removed from the data. By selecting the explanatory variables to be included in the model, the model will be more accurate when fitted onto the data set.

Using correlation analysis, which determines the correlation between two variables, is a good way to do this. In this image, a heat map is used to show how the different features are related to one another. If the correlation between variables is strong, it is shown as either blue for a positive correlation, or red for a negative correlation.

In the key shown at the bottom of the image, the strength between each combination of variables is shown in shades of the representative colour. If the correlation is a strong positive correlation, the corresponding colour will be shown as dark blue, and will become lighter as this positive correlation decreases towards 0. If the correlation is a strong

negative correlation, the corresponding colour will be shown as dark red, and will become lighter as this negative correlation increases towards 0.

Using the heat map makes the process of identifying the least valuable variables to be excluded easy. In this example, the most insignificant variable is the height of the person, since it is not correlated with the wages of the individual.

4. Fit the regression model

A regression line is then fitted onto the scatterplot of the data set to explain the data, and to calculate a prediction. In this example, the regression line can then be used to determine the linear relationship between an individual's wages and level of experience. It is clear that there is a direct correlation between wages and the level of experience, whereby the wages increase as the individual becomes more experienced.

Classification

This example explores the fitting of a classification model onto a data set, showing the different steps in the process, including importing and cleaning the data, and fitting the classification model onto the data to produce a visualisation that can be used to inform a decision. In this example, the problem that the machine learning model needs to solve is to determine the correlation between employment status and the other variables in the data set. The available data set includes the individuals' years of tertiary education completed, employment status, and height. Classification will be the most appropriate model to fit onto this data set, since the goal is to classify individuals into two categories, employed or not employed, which is a typical classification problem.

1. Import data

With both regression and classification, the first step is to import and view the data set in the form of a data table. In this example, there are four entries listed in a table, with each entry showing the years of tertiary education completed by the individual, whether the person is employed, and their respective heights.

As the data is imported and viewed, it is important to clean the data and identify any issues with it. The issues seen in this data set include a duplicate entry, and a missing entry regarding the height of an individual.

2. Clean data

After the duplicate entry and missing data have been removed, the data is clean and ready to use.

3. Variable selection

Explanatory variables, or independent variables, are selected from the imported data. By selecting the explanatory variables to be included in the model, the model will be more accurate when fitted onto the data set.

Box plots are useful tools to identify the important variables to be included in the model before it is fitted onto a classification data set. In this example, there are two box plots: one that shows the distribution of tertiary education for employed and unemployed individuals, and another that shows the distribution of height for employed and unemployed individuals.

In the first box plot, the medians are roughly equal, which means that there is no significant difference between the heights of employed individuals and the heights of unemployed individuals. However, in the second box plot, the medians are different, indicating that there is a significant difference in level of education between employed individuals and unemployed individuals.

Therefore, only level of education will be used to determine employment status.

4. Fit the classification model

A classification model is then fitted onto the data set to predict the probability of an individual with a certain level of tertiary education to be employed. In this example, it is clear that as the level of education increases, so does the probability of employment.

Conclusion

It is difficult to make predictions when viewing data in its raw format. By condensing the raw data into graphs, it becomes more accessible and useful for decision-making. In this interactive infographic, you have seen how data can be transformed from seemingly random variables into useful visualisations by following a few steps. As you progress through the next section in the Unit 2 notes, the creation of plots in R will be explained in more detail, after which you will have the opportunity to practise these steps in R.