DataVizA Tutorial: Predictive Analytics: Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 8

Problem Setup

This exercise is based on real data that can be found here. However for educational reasons some of the context has been made up.

Consider the case where you are a manufacturer of wine. You produce a number of different wines which are sold in three different markets; the Australia, Europe, and Japan. The customers in each of these regions have different wine preferences; wines favoured by Australians are not favoured by Europeans and Japanese, wines favoured by Europeans are not favoured by Australians and Japanese and wines favoured by Japanese are not favoured by Australians and Europeans.

Last year, the National Wine Institute commisioned a large scale survey to determine whether the wine is most favoured by Australian, European or Japanese customers. As well as this, data on the following chemical attributes of each wine were also collected:

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

The units of measurement of all chemical features are standardised. The data are available in the file *ExistingWines.rds* which can be found on Moodle.

This year's wines have already been produced and you need to determine which region each wine should be sold in. For this year's wines, it is relatively inexpensive to collect data on chemical attributes of the wine. These are available in the file *NewWines.rds* which is available on Moodle. However, the National Wine Institute is unable to conduct a similar survey for this year to determine the best market for each new wine.

1. Is this problem a case of classification or regression? Why or why not?

This is an example of a classification problem. The variable we would like to predict is BestMarket. This is a categorical variable.

2. Using a real example in a business setting, how could the dependent variable be different so that your answer to the previous question changes.

Regression would be used if the dependent variable was numerical. For instance if the objective was to predict the total sales on wine given the chemical characteristics of the wine, then regression can be used.

3. Suppose that a there is a wine for which the following prediction is made. The probability wine is prefered by Australians is 30%, the probability wine is prefered by Europeans is 34% and probability that wine is prefered by Japanese is 36%. Without any further information, which market would you sell this wine at?

With no further information we would predict that this wine is of a variety prefered by Japanese customers. It should be sold in Japan.

4. What are some costs associated with misclassifying a wine?

By not matching a wine to the right market there could be a drop in sales as well as revenue forgone by failing to sell the wine in the correct market. There could also be damage to the brand reputation of the winemaker.

5. Suppose Japanese strongly dislike the types of wines favoured by Australians. Also assume that the Japanese market is much larger than the Australian market. As such, the winemaker considers selling wines favoured by Australians to Japanese customers to be the most costly risk. How might this change the answer to Question 3.

Although it is not the highest, the probability that the wine is of the type prefered by Australians is still quite high. If this wine is sold in Japan there is a high risk of damaging brand reputation. As a result the company may decide to classify the wine as "Japan" only when the probability is higer than a large threshold, e.g. 0.6

6. Suppose a classification algorithm is used and the following results are obtained

True best market	Predicted Australia	Predicted Europe	Predicted Japan
Australia	38	1	4
Europe	1	46	4
Japan	2	4	25

Compute the misclassification rate

There are 1+4+1+4+2+4=16 misclassified observations. The total number of observations are 1+4+1+4+2+4+38+46+25=125. Therefore the misclassifcation rate is 16/125=0.128, or 12.8%

Concepts

Define, describe or explain the following terms

1. Training sample

The training sample is the data used to derive a rule for classifying a variable.

2. Test sample

Once a classification rule is trained its accuracy can be evaluated on a test sample. The observations in the test sample are NOT used to derive the rule for classification.

3. Leave one out cross validation

A single observation is chosen to be the test sample, all other observations are used in the training sample. This entire process is replicated so that every observation is in the test sample exactly once.

4. Sensitivity

In a two class problem one class is assigned to be positive. The sensitivity is the proportion of observations that are truly positive that are classified as positive.

5. Specificity

The specificity is the proportion of observations that are truly negative that are classified as negative.