Logistic Regression

Peter Menzies

11/01/2021

Classification vs Clustering

```
CLASSIFICATION              |      CLUSTERING

Supervised                  |     Unsupervised

Predifined classes/groups   |   Identifies similarities between objects
                                        to form "clusters"/groups

eg. * Spam/not-spam emails      eg. * Netflix recommendations
    * Fraud detection       |       * To gain information about customers

* Binary * Multi-class,
* Multi-label * Imbalanced  |       * Hard  * Soft

                                  * Connectivity  * Centroid
                                  * Distribution  * Density
```

Logistic Regression

Logistic regression is a model used for Binary Classification; that is the probability of a certain event occuring, for example passing or failing a test. By inputing one or more predictors, which can be continous or binary, the model uses linear regression to produce the "logarithm of the odds" (log-odds) of an event occuring (eg. dependant value = 1 = "pass the test"). The logistic function is then used to provide the probability. For the model to perform classification, a threshold value is required. A defining characteristic of logistic regression is that increasing a predictor's value, will multiplicatively scale the odds of an outcome at a constant rate.

Extensions of logistic regression models include; *multinomial regression*, used when the dependant variable (outcome) has more than two values; *ordinal logistic regression*, used when the dependant variable has multiple values which are ordered; *mixed logit*, which allows for correlations between the values of the dependant variable; and *conditional random fields*, which uses sets of interdependant variables.

Advantages of a logistic regression model include it being relatively easy to implement and interpret, and it not requiring high computation power. The model also gives information on the importance/weight of a predictor and it's direction of association. Models can be updated to reflect new data more easily than models such as *decision trees* and *support vector machines*. Unlike some other classifcation models, logistic regression models produce a well-calibrated probability in addition to classification.

Key disadvantages are that these models are prone to over-fitting in data sets with multiple predictors or limited training sets; they cannot solve non-linear problems; and they cannot model complex relationships as effectively as other models such as *Neural Networks*.

Also excel doesn't support logistic regression apparently.