

Problem set 9

Exercise 1. Iris classification

Use the Edgar Anderson's Iris Data in R to build a model for classification of flowers. Type “?iris” in R for more information on this dataset. Use random about 110 cases to train your model and use the rest cases to predict the model accuracy. Perform classification in R, Python, Orange, RapidMiner and compare your results.

Exercise 2. Red Wine analysis

a) Data reduction

Using the dataset winequality-red.csv perform data reduction and compare the data reduction algorithms in R, Python, Orange, and RapidMiner.

b) Cluster analysis

Use winequality-red.csv dataset to conduct cluster analysis in R, Python, Orange. Compare k-Means clustering and hierarchical clustering. Are there any differences in clustering the wine? (Some examples in R: <http://www.rpubs.com/gabrielmartos/ClusterAnalysis>).

Describe and plot your findings.

c) Classification

Use the same winequality-red.csv dataset to classify the wine according variable “quality”. Choose your train and test set. Describe the accuracy of your classification model.

Exercise 3. Google Trends – model for geographical clustering

Chose five different search terms in Google Trends (<https://trends.google.com/trends/>) or use Google Correlate. Set the map settings for world or USA for your analysis. Find the geographical clusters corresponding to your chosen search terms in R, Python, and Orange. Describe some interesting similarities in geographical regions.