DataVizA Tutorial: k Nearest Neighbour Classification

Department of Econometrics and Business Statistics, Monash University

Tutorial 9

Wine Data

These questions are based on the problem from last weeks tutorial

1. Carry out kNN classification using all data in *ExistingWines.rds* in the training set and predict the data in *NewWines.rds*. Let $k = 1$
2. What are the predictions for the first ten wines in the *NewWines.rds*
3. Repeat the analyis with $k = 5$
4. What are the predicted probabilities for the first ten wines in the *NewWines.rds* when k=5
5. Split the data in *ExistingWines.rds* into a training sample (of roughly 70%) and a test sample (of roughly 30%). Hint `runif(125)<0.7` will create a vector of length 125 where each element is either TRUE with probability 70% and false with probability 30%. You may also want to use the `ifelse` function
6. Is $k = 1$ a better choice than $k = 5$ according to the misclassification rate?