COMP30120 Tutorial

Dimension Reduction and Feature Selection

Q1

(a) In Weka, use filter-based feature selection with Information Gain to identify the 3 most discriminating and 3 least discriminating features in the *Wine* data set in the ARFF file provided. Then assess the accuracy of a 1-Nearest Neighbour classifier with:

- (i) only the 3 most discriminating features included.
- (ii) only the 3 least discriminating features included.

(b) In Weka, apply wrapper-based feature selection to the *Wine* data set using a 1-Nearest Neighbour classifier and the following search strategies: (i) forward sequential search, (ii) backward elimination.

Q2

Use Weka to apply PCA feature extraction to the *Diabetes* data in the ARFF file provided. Note that the search method should be set to *Ranker*.

Q3

(a) Explain why the feature subset selection problem with a *k*-Nearest Neighbour classifier is an exponential search problem.

(b) Describe in outline a Genetic Algorithm solution to this search problem.

(c) Describe crossover and mutation techniques for feature subset selection.

(d) Why is overfitting a potential risk in wrapper feature subset selection?