

COMP47750/COMP47990 Tutorial

Dimension Reduction

Python code in notebook “11 DimRed Tutorial”

1. In **scikit learn**, apply *filter-based feature selection* with Information Gain to identify the 3 most discriminating and 3 least discriminating features in the *Wine* dataset in the CSV file provided.

Based on these results, assess the 10-fold cross-validation classification accuracy of a 1-Nearest Neighbour classifier with:

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

2. Using **mlxtend** identify informative feature subsets by applying *wrapper-based feature selection* to the *Wine* dataset using a 3-Nearest Neighbour classifier and the following search strategies:

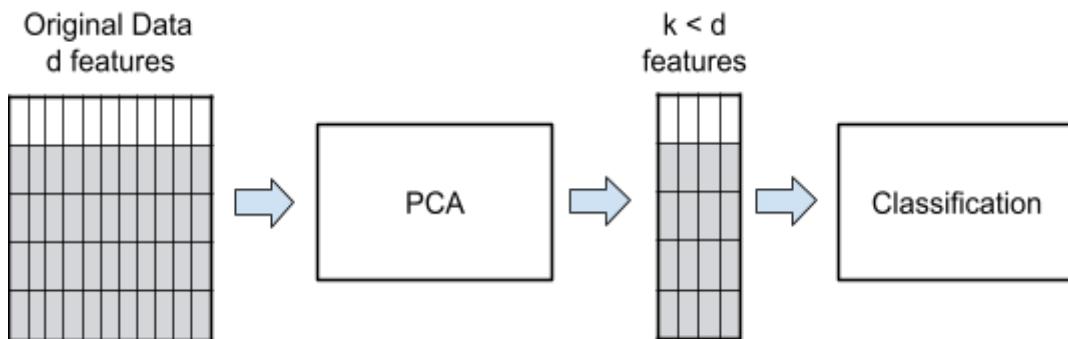
- (i) forward sequential search
- (ii) backward elimination search

Which common features were selected by both search strategies?

3. In situations where there is a lot of data available, it may not be necessary to use all the data for Wrapper-based feature selection. Try this out with the dataset in **segmentation-all.csv** (2310 data points).

- (i) Run forward-sequential-search to select 5 features and test the accuracy of the select feature subset using cross validation (code in notebook “11 DimRed Tutorial”).
- (ii) Repeat but using only 1000 data points (`X_scaled[:1000]`).
- (iii) Is the same feature subset selected? Is the accuracy the same?

4. The purpose of this exercise is to explore the impact of using PCA as a preprocessing step in classification. (*most of the code is in the notebook*)



We use the wine dataset and a k -NN classifier.

- (i) What is the base-line accuracy using the scaled features?
- (ii) Plot the explained variance of the first 10 PCs.
- (iii) Based on this plot select a number of PCs to compress the data.
- (iv) What is the accuracy of a classifier using the compressed representation?
- (v) What is the accuracy if we use just one PC?