Queen Mary
University of London

Data Mining

Assignment4

Spyridon Roumpis

181004877

Exercise 0: Now starting from this 4-attribute subset, find the best 3, 2, 1 attribute subset, filling in the table below. Which sized subset, and which set of attributes yields the best accuracy?

Subset Size	Attributes Selected	Accuracy	Attribs Removed
5	All: W, P, Hol, Vac, Health	85.9%	None
4	W, P, HOL,HEALTH	89.4737 %	VAC
3	W, P ,HOL	91.2281 %	HEATH
2	W, P	85.9649 %	HOL
1	P	80.7018 %	W

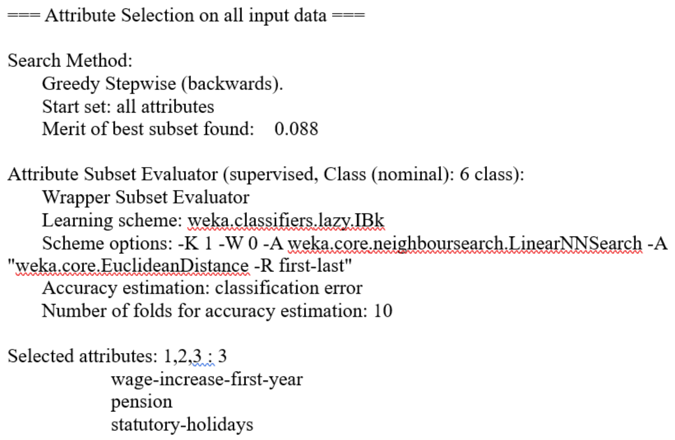
From the above table it can be seen that the best accuracy-91.2281% is achieved when the subset size is 3 and the selected attributes are wage-increase-first-year, pension and statutory-holidays.

Exercise 1: How many feature combinations did you try? How many combinations of features are there in total? Give an example of a combination of features that you did NOT try when doing backward selection.

For the first experiment-finding out the best 4 attribute subset we tried 5 different feature combinations, then for the best 3 attribute subset there were 4 different feature combinations, continuing with the best 2 attribute subset we tried 3 different feature combinations and finally for the best 1 attribute subset there were 2 feature combinations. In total we tried 5+4+3+2 = 14 different feature combinations.

In total there are 2^M-1 = 2^5-1=63 feature combinations. One combination we did not try is when selecting the best 2 attribute subset the W,P,HEALTH.

Exercise 2: How many and which attributes are selected? Do they match the results from *Section 2*?



```
=== Attribute Selection on all input data ===

Search Method:
      Greedy Stepwise (backwards).
      Start set: all attributes
      Merit of best subset found:    0.088

Attribute Subset Evaluator (supervised, Class (nominal): 6 class):
      Wrapper Subset Evaluator
      Learning scheme: weka.classifiers.lazy.IBk
      Scheme options: -K 1 -W 0 -A weka.core.neighboursearch.LinearNNSearch -A
"weka.core.EuclideanDistance -R first-last"
      Accuracy estimation: classification error
      Number of folds for accuracy estimation: 10

Selected attributes: 1,2,3 : 3
                    wage-increase-first-year
                    pension
                    statutory-holidays
```

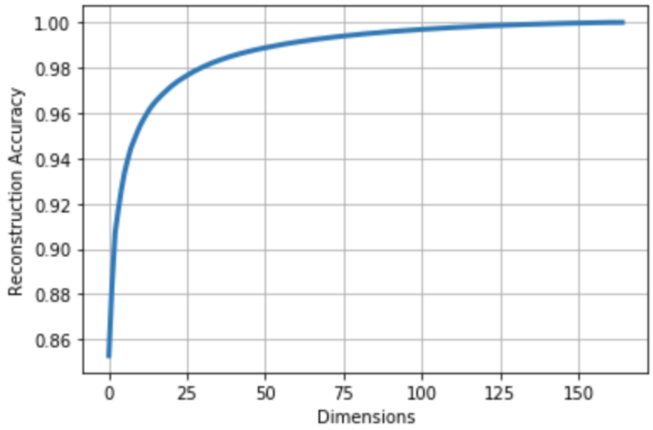
It can be seen, that the selected attributes are 3, the wage-increase-first-year, the pension and the statutory-holidays. When comparing these three with the three from the greedy feature selection from the previous section we see that the results match.

Exercise 3: Which attributes does it pick (and hence which ones are discarded?)

```
Instances:    150
Attributes:   11
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
              Copy of sepallength
              Copy of sepalwidth
              Copy of Copy of sepallength
              Copy of Copy of sepalwidth
              Copy of Copy of Copy of sepallength
              Copy of Copy of Copy of sepalwidth
Evaluation mode:    evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
      Greedy Stepwise (forwards).
      Start set: no attributes
      Merit of best subset found:    0.043

Attribute Subset Evaluator (supervised, Class (nominal): 5 class):
      Wrapper Subset Evaluator
      Learning scheme: weka.classifiers.bayes.NaiveBayes
      Scheme options:
      Accuracy estimation: classification error
      Number of folds for accuracy estimation: 5

Selected attributes: 3,4 : 2
                    petallength
                    petalwidth
```

It can be seen that the only attributes that have been selected are 2, the petallength and the petalwidth. The classifier discarded the two main other attributes seplallenght ans sepalwidth as long as the copies of these 2 attributes we created in the previous step, when we created the filters, and the fifth attribute which is the class (label) itself.

Exercise 4: Use the data used to produce the above plot to find out what number of PCs is required to explain 99% of the data variance (achieve 99% reconstruction accuracy). What # is this and does it match the value from Q8? Provide a short discussion.



It can be seen from the above graph that to obtain a 99% reconstruction accuracy the number of PCs (Principal Components) that is required is around 55. From question 8 after testing different values of the number of PCs which gave us a reconstruction error <1% is 158. The reason for this difference is the use of eigenvectors and eigenvalues, as we know they reduce the noise in the data.

Exercise 5: Which number of PCA dimensions gets the maximum face recognition accuracy? Is it better or worse than the accuracy when classifying the raw images? Why? (What factors contribute to this?) Provide a brief discussion.

Experimenting on the different values of PCA the maximum face recognition accuracy – 0.707 was achieved when the number of principal components (nPCA) was 43. This accuracy is not the same as the accuracy for the raw images classification and that is because of the 'The Curse of Dimensionality'. In a high dimensional setting, we need to learn more parameters than in a low dimensional one, so the risk of overfitting increases. This risk is especially dangerous if we have many attributes that are weakly relevant, or some very relevant and many irrelevant. With Dimensionality reduction we aim to transform a high dimensional dataset into a more convenient low dimensional dataset.