Neeraj Vashistha (190573735)

Assignment 4

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? [1 mark]

Subset Size	Attributes Selected	Accuracy	Attribs Removed
5	All: W, P, Hol, Vac, Health	85.9%	None
4	W, Hol, Vac, Health	85.9%	P
3	W, Hol, Vac	85.9%	P, Health
2	Hol, Vac	80.7%	P, Health, W
1	Hol	75.4%	P, Health, W, Vac

P is removed initially as P has the highest number of missing values (53%), then we have health, 35% missing values and so on. When we remove P and Health, we see that the accuracy has not decreased.

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. [1 mark]

5 Combinations of features we tried in backward feature selection process. There could be 2^5 - 1 (31) total combinations of features in total. During the backward selection process we did not try 'P,W,Hol' as P, W were removed earlier and if we just use 'P,W,Hol' we get an accuracy of 91%.

Exercise 2: How many and which attributes are selected? Do they match the results from Section 2? [1 mark]

In Greedy Backward selection, 3 attributes are selected, 'W, P, Hol' (wage-increase-first-year, pension, statutory-holidays). While in case of Information gain ranking method, attributes are ranked in fashion, W, Hol, Health, Vac, P, where W is highest, and P is lowest ranked. Thus, we see W and Hol are given importance in both the algorithms while P is seen most important in Greedy Backward Selection and least in Information gain ranking method.

Exercise 3: Which attribute does it pick (and hence which ones are discarded?) [1 mark]

From the below attributes:

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

Selected attributes: 3,4 : 2
petallength
petalwidth

Thus discarded attributes are sepallength and sepalwidth and copy of these.

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. [1 mark]

The reconstruction accuracy is nearly 99% when the number of dimensions are around 100, this can be observed from the graph below. In Q8 where it was asked what is the best nPCA to achieve <1% reconstruction error. If we incrementally increase the value and reach 100 we see that features like glasses and facial expression are visible and reconstruction error is 0.056415763470590005 and we are successfully be able to reconstruct the images using 100 dimension with as the size of original image (5406.72 (KB)) is reduced to 132.0 (KB).

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. [1 mark]

We get the maximum face recognition accuracy of 67% with PCA dimension around 130. If we decrease the nPCA (dimensions) to 100 the accuracy drops to 65% but if we increase the nPCA to 250 the accuracy does not increase, and it remains constant at 67%. With the raw images i.e. nPCA = 4096 the accuracy is 67% only, which means it is only required 130 dimensions(features) to attain a fair amount of accuracy.