
Homework 8
Patricia Hoffman, PhD.

You are only required to answer one of these questions. Please spend the extra time working on your Final Projects.

1) Cluster the glass data set with the Expectation-Maximization algorithm. Try another clustering algorithm on the glass data set. Which method worked better for that data set?

2) Choose a data set of your choice. Use two different clustering methods to classify the data. Which method worked the best.

3) Challenge Question: Try various unsupervised methods to classify the Synthetic Control Chart Time Series Data Set. Recall that the methods used so far do not take advantage of the fact that this is time series data. You are free to create a method or find one in any research paper. One paper that discusses time series data written by R. J. Alcock and Y. Manolopoulous can be found here
http://machinelearning2010fall.pbworks.com/w/file/32772288/TimeSeriesData10.1.1.79.1572.pdf
The data set is here
http://archive.ics.uci.edu/ml/datasets/Synthetic+Control+Chart+Time+Series

Which method worked the best? Why do you think that is the case?

4) Challenge Question: Try various supervised methods to classify the Synthetic Control Chart Time Series Data Set. Try using r-part to classify these data. Try another supervised method. Which method worked best?

5) Challenge Question: Write a wrapper function around the r function for K-means which will implement the Bisecting K-means algorithm discussed in Chapter 8 of the text. Use bisection K-means to cluster the Iris data into three clusters and then into 6 clusters.

6) In the class, the Iris data has been classified using many different techniques. Make a chart with two columns. The first column should contain the name of the technique and the second column should contain the number Iris flowers which were correctly classified with that technique. Did any unsupervised techniques work better than any of the supervised techniques? Which technique worked the best?

7) Classify the sonar data using the Expectation-Maximization, K Means, Divisive and Agglomerative Hierarchical Clustering Techniques. What is the percent correctly classified with each of these methods?

8) Choose one of the techniques discussed in class. Discuss what types of data the technique can be used on. Discuss the pros and cons of the technique. What types of data does it work well on? What types of data does it have trouble with and why?

9) The text, "An Introduction to Information Retrieval" by Christopher D. Manning is on the web at http://nlp.stanford.edu/IR-book/html/htmledition/cluster-cardinality-in-k-means-1.html The Akaike Information Criterion (AIC), Equation 197 of the section titled "Cluster cardinality in K-means" can be used to determine the best K in the K-means algorithm for a given set of data.

$$\textbf{AIC:} \qquad K = \arg\min_{K} [\mathrm{RSS}_{min}(K) + 2MK]$$

where $RSS_{min}(K)$ is the minimal Residual Sum of Squares (RSS) of all clusterings with K clusters. Observe that the $RSS_{min}(K)$ is a monotonically decreasing function in K. For K means, $RSS_{min}(K)$ measures the distortion to which cluster members deviate from the centroid of the cluster. 2MK is the model complexity in which K is the number of clusters and M is the dimension of the input data (the number of features).

Use the Iris data set. Let K run from 1 to 12. For each K generate 10 different sets of initial starting points and run K means using each of these sets. According to the AIC, which K is the best for the Iris Data?

10) Appendix B of the text describes Principal component Analysis (PCA) and also Singular Value Decomposition (SVD). Both of these techniques can be used for dimensionality reduction. Apply both PCA and SVD to the Iris Data. Recall the class example svm1.R which classified the Iris data with a Support Vector Machine and only miss classified one flower. This example used all four features. With the use of PCA or SVD, can you obtain the same results

using SVM on the Iris data using only 3 features? Figure B2 of the text plots the first two principle components of the Iris data. Use R to generate this plot. Make a similar plot using SVD.

11) Appendix B of the text describes Principal component Analysis (PCA) and also Singular Value Decomposition (SVD). Pick a classifying technique for the sonar data that has been used in the class. See if you can reduce the number of features (using PCA or SVD) and obtain the same results with the reduced feature size.