Assignment 5 (given by Zoran Obradovic)

No. of Problems: Three.
Out: March 03, 2006
Due: March 15 in class.

Problem 1:
Rank chapters 7 to 13 with respect to your preference for reading+presentation assignment. In addition, you can suggest a machine learning topic of your interest that is not covered by the textbook (in such a case, write a brief outline of the proposed presentation with reference(s)). Also, rank dates of 3/22, 3/29, 4/05, 4/12 based on when you would prefer to give a lecture on the assigned topic (lecture is limited to 60 minutes and should be PowerPoint based). When assigning reading+presentation tasks I will take into consideration your topic and time preference where conflicts will be resolved based on time flexibility).

Problem 2:
Do Problem 6.1 on page 198.

Problem 3 (five tasks):
Download the Adult Database from UCI Machine Learning Repository and clean missing values.

- Apply Naive Bayes algorithm to learn the Adult problem using all features.
- Perform feature selection using feedforward search technique. Compare the effects of using Euclidean distance vs. Naive Bayes algorithm as a selection criteria. Report rank of selected features based on both approaches. How many attributes seems to be the best choice? To explore feature selection variability for what seems to be the best number of features repeat the experiment using 5-fold cross-validation (compare accuracy of Naive Bayes algorithm based on features selected using Euclidean distance selection criteria vs. Naive Bayes selection).
- Perform feature reduction using Principal Components Algorithm (PCA). Report retained variance when reducing to various dimensions. Reduce data to the same dimension identified in the feature selection experiment as the best choice. Develop Naive Bayes algorithm using this data representation and compare the result to feature selection based data reduction.
- Cluster Adult database using k-means and also using hierarchical clustering approach. Discuss the outcome.
- Develop cluster-specific Naive Bayes predictors and report how these compare to using a single predictor developed on the entire data.