CSE343/543: Machine Learning

Assignment 3 November 16, 2015, 11:59 PM

Any instance of cheating will be considered as academic dishonesty and appropriate penalty will be applied.

1. Apply regression on the Year Prediction MSD dataset to predict the year. Try multiple regression (i.e. the regression problem using the multi-dimensional features) and report the prediction error on the dataset. Also determine which is the best feature and report the results of linear regression with this feature. Also show the scatter plot for this feature. Use initial 50% data for training and remaining data for predicting the year. You can use any inbuilt function for regression.
   Link: (http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD#).

2. Q 13.3 from the textbook (Machine Learning Tom M. Mitchell). Chapter 13, page 388.

3. Implement the **k-means algorithm** on the Wine database (http://archive.ics.uci.edu/ml/datasets/Wine). (**No toolbox can be used**)
   a. Use 50% data from each for training respectively. Use Euclidean and Chebyshev distance as your distance measure. Report which distance measurement performs better?
   b. Assign each final cluster a name by choosing the most frequently occurring class label of the examples in the cluster.
   c. Compute and report the training error and draw the confusion matrix.
   d. Use the remaining dataset for testing and compute the testing error (total as well as class-wise).
   e. Repeat the above four questions with kernel k-means algorithm. Apply the polynomial kernel k-means algorithm (degree >= 2) to find clusters (k = 3). Compare the results with k-means (without kernel) in terms of accuracy and time complexity.

4. Using the concept of genetic algorithm in game playing solve the checkerboard problem using memetic algorithm. You can use a toolbox only for hill climbing (if required) - the remaining has to be your own code.