CS-584 – Assignment 5 (optional)

Expectation Maximization

Due by: April 30, 2016

Assignment Specifications

In this assignment you will implement the Expectation Maximization algorithm in one of the following applications: clustering assuming a Gaussian mixture model, or factor analysis assuming Gaussian distribution. You have to use two or more external data sets. Links to data sets are available on the course web-page (e.g. the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/). It is essential that you evaluate the performance of each algorithm you implement and the effects of varying different parameters on the performance of the learning algorithm. Use cross validation to test performance. The grade for this assignment will be based in part on the performance of your implementation and on the thoroughness of your evaluation. Make sure to explain the results you obtain and do not unnecessarily repeat similar results. The code you write should be modular and well documented. The implementation needs to be in Python.

Clustering

1. Select two datasets. Each dataset should contain examples from multiple classes. For training purposes assume that the class label of each example is unknown (if it is known, ignore it).

2. Implement the K-means algorithm and apply it to the data you selected. Evaluate performance by measuring the sum of Euclidean distance of each example from its class center. Test the performance of the algorithm as a function of the parameter k. Project the data onto 2D and plot the data and the cluster centers you obtained. The data in each cluster should be plotted using a different mark or color.

3. Implement the EM algorithm assuming a Gaussian mixture. Apply the algorithm to your datasets and report the parameters you obtain. Evaluate performance by measuring the sum of Mahalanobis distance of each example from its class center. Test performance as a function of the number of clusters. Project the data onto 2D and plot the data and the cluster centers you obtained. The data in each cluster should be plotted using a different mark or color.

4. Suggest and test a method for automatically determining the number of clusters.

5. Using a dataset with known class labels compare the labeling error of the K-means and EM algorithms. Measure the error by assigning a class label to each example. Assume that the number of clusters is known.

Factor Analysis

1. Generate examples for two distinct 2D Gaussian clusters and plot the data. Map the examples you generated to a higher dimension using a non-linear map (e.g. polynomial) and a linear map (i.e. using a wighted sum) to form a non-linear map set and a linear map set respectively.

2. Perform PCA dimensionality reduction (2D) to each of the sets you generated, and plot the results you obtain.

3. Apply the EM factor analysis to each of the sets you generated and plot the factors (2D) you obtain. Evaluate the results you obtain and compare them to the PCA results.

4. Select two datasets. Each dataset should contain examples from multiple classes. For training purposes assume that the class label of each example is unknown (if it is known, ignore it).

5. Apply the PCA and EM factor analysis algorithms to the datasets you selected and analyze the results you obtain.

6. Suggest and test a method for automatically determining the number of factors.

Submission Instructions

Follow the submission instructions of assignment 1. All submissions must be made before the end of the semester.