MASSACHVSETTS INSTITVTE OF TECHNOLOGY
Department of Electrical Engineering and Computer Science
6.036—Introduction to Machine Learning
Spring 2014

Project 3: Pun With Words   Issued: Tues., 4/15 Due: Fri., 4/25 at 9am

Project Submission: Please submit two files—a *single* PDF file containing all your answers, code, and graphs, and a *second* .zip file containing all the code you wrote for this project, to the Stellar web site by 9am, April 25th.

Introduction

Your task is to build mixture model for collaborative filtering. In this project you will be given a fraction of the data matrix containing users and movie ratings from the netflix database. We have processed a subset of Netflix's data such that you get access to all the ratings for a subset of movies. The goal of this project will be to use the hidden structure in the different types of users that exist using the EM algorithm, then with the knowledge and this hidden structure we hope to be able to complete a partially observed rating matrix as an end goal.

Notation

We will use X to denote the data matrix. It will be an $n \times d$ matrix, meaning that there will be n rows and d columns. The rows of the data matrix indicate users and the columns indicate movies. A single entry $x_j^{(i)}$ of the matrix will indicate the rating person i gave to movie j and the rating will be in the set from 1 to 5, i.e. $x_j^{(i)} \in \{1, ..., 5\}$.

1 **Part 1**

For this part of the project you will explore the connections of the results that clustering gives versus the clusters that EM algorithm gives.

(a) Use the toy data set and the k-means code we provided to plot different clusters for cluster sizes $k = [5, 1015]$. Notice that each data point is fully assigned to a single cluster. Meaning, that each point can only have one underlying hidden label for this model.

(b) Recall the mixture model that we was presented in class $P(x^{(t)}|\theta) = \sum_{j=1}^{K} p(j|\theta)p(x^{(t)}|j, \theta)$. A data point may be generated by first choosing a cluster and then choosing a data point x according to that cluster. Conceptually, explain how once you have cluster assignments given by k-means, how the data generation process for your model for k-means would differ from one from a mixture model. [Hint: k-means divides the plane with Voronoi diagrams but a mixture model allows weighted averages across cluster types].

(c) Consider a mixture model that uses a Gaussian as the conditional distribution given the hidden label i.e. $p(x^{(t)}|j, \theta) = N(x^{(t)}|\mu^{(j)}, \sigma_j^2)$. We will try to understand the EM algorithm on this model. Recall that the EM algorithm aims to learn these parameters by sequential minimization (similar to how we presented clustering and boosting). In the E-step we will have the model fixed (i.e. hold the $\mu^{(j)}$ and the σ_j^2, similar to holding the clusters fix) and compute the soft count assignments for each data point i.e. the posterior probability $p(j|x^{(t)}) = p(j|t)$. The M-step will receive these soft-counts and treat them as fixed. Now that we have these soft counts, we will compute the maximum-likelihood estimates for our model but with the hard-counts replaced by the soft count.