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 **Clustering to discover the types of users**

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

Recall that rows from the data matrix indicate users. In this section we will cluster the users to discover their underlying type.

(b)

(c)