COMS 4721: Machine Learning for Data Science

Columbia University, Spring 2015

Homework 4: Due April 14, 2015

Submit the written portion of your homework as a single PDF file through Courseworks (less than 5MB). In addition to your PDF write-up, submit all code written by you in their original extensions through Courseworks. Do not submit in .rar, .zip, .tar, .doc, or other file types. Your grade will be based on the contents of one PDF file and original source code. Everything resulting from the problems on this homework other than the raw code should be put in the PDF file.

Show all work for full credit. Late homeworks will not be accepted – i.e., homework submitted to Courseworks after midnight on the due date.

Problem 1 (K-means) – 25 points

Implement the K-means algorithm discussed in class. Generate 500 observations from a mixture of three Gaussians on $\mathbb{R}^2$ with $\pi = [0.2, 0.5, 0.3]$,

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mu_2 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mu_3 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

1. For $K = 2, 3, 4, 5$, plot the value of the K-means objective function per iteration for 20 iterations (the algorithm may converge before that).
2. For $K = 3, 5$, plot the 500 data points and indicate the cluster of each for the final iteration by marking it with a color or a symbol.

Problem 2 (Matrix factorization) –75 points

In this problem, you will implement the MAP inference algorithm for the matrix completion problem discussed in class. As a reminder, for $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$, we have

$$u_i \sim N(0, \lambda^{-1} I), \quad i = 1, \ldots, N_1, \qquad v_j \sim N(0, \lambda^{-1} I), \quad j = 1, \ldots, N_2.$$

We are given an $N_1 \times N_2$ matrix M with missing values. Given the set $\Omega = \{(i, j) : M_{ij} \text{ is measured}\}$, for each $(i, j) \in \Omega$ we have $M_{ij} \sim N(u_i^T v_j, \sigma^2)$.

Run your code on the user-movie ratings dataset provided on Courseworks and the course website. For your algorithm, set $\sigma^2 = 0.25$, $d = 20$ and $\lambda = 10$. Train the model on the larger training set for 100 iterations. For each user-movie pair in the test set, predict the rating by mapping the relevant dot product to the closest integer from 1 to 5. Since the equations are in the slides, there's no need to re-derive it.

1. Plot the RMSE of your predictions on this test set as a function of training iteration. On a separate plot show the log joint likelihood as a function of iteration.

2. Pick three movies and for each movie find the 5 closest movies according to Euclidean distance using their respective locations v_i. List the query movie, the five nearest movies and their distances. A mapping from index to movie is given with the data.

3. Perform K-means on the $u_1, \ldots, u_{N_1}$ learned by your algorithm. Set $K = 30$. The centroids can be interpreted as personality types (as far as movies are concerned). Pick 5 centroids and characterize wthe cluster by showing the 10 closest movies to each centroid.