CS4786/5786: Machine Learning for Data Science, Spring 2016
4/4/2016: Assignment 2: Stability of Spectral Clustering and EM algorithm for Mixture Models

---

Instructions Due at 11:59pm Friday April 15th on CMS. Submit what you have at least once by an hour before that deadline, even if you haven't quite added all the finishing touches — CMS allows resubmissions up to, but not after, the deadline. If there is an emergency such that you need an extension, contact the professors.

You may work in groups of one up to four[1]. Each group of two or more people must create a group on CMS well before the deadline (there is both an invitation step and an accept process; make sure both sides of the handshake occur), and submits 1 submission per group. **We will not be automatically transferring the groups from A1, so that you can change groups if you want, but this means you should form groups anew for A2.** You may choose different groups for different assignments. Please ensure that each member of the group can individually defend or explain your group's submission equally well.

You will submit both a writeup and some datafiles you create. The writeup can be handwritten or typeset, but please make sure it is easily readable either way.

Keep an eye on the course webpage for any announcements or updates.

Academic integrity policy We distinguish between "merely" violating the rules for a given assignment and violating *academic integrity*. To violate the latter is to commit *fraud* by claiming credit for someone else's work. For this assignment, an example of the former would be getting an answer from person X who is not in your CMS-declared group but stating in your homework that X was the source of that particular answer. You would cross the line into fraud if you did not mention X. The worst-case outcome for the former is a grade penalty; the worst-case scenario in the latter is academic-integrity hearing procedures.

The way to avoid violating academic integrity is to always document any portions of work you submit that are due to or influenced by other sources, even if those sources weren't permitted by the rules.[2]

[1]The choice of the number "four" is intended to reflect the idea of allowing collaboration, but requiring that all group members be able to fit "all together at the whiteboard", and thus all be participating equally at all times. (Admittedly, it will be a tight squeeze around a laptop, but please try.)

[2]We make an exception for sources that can be taken for granted in the instructional setting, namely, the course materials. To minimize documentation effort, we also do not expect you to credit the course staff for ideas you get from them, although it's nice to do so anyway.

*Q1 (***Clustering Sensitivity).** The goal of this assignment is for you to explore the sensitivity of spectral clustering algorithm (normalized-cut version), by showing that small perturbations of the initial data can lead to a quite different clustering, even for binary clusterings.

Biggest hint I can give you for this problem is: draw the graph using pen and paper. Think simple! In grading, we care about your explanations at least as much, and perhaps more, than the datasets you provide.

In this assignment, K, the number of clusters per clustering, is fixed at $K = 2$, and n, the number of data points each initial dataset should contain, is fixed at $n = 30$. When you are asked to provide a vector c of cluster assignments, in such vectors, the t^{th} entry c_t is 1 if the t^{th} datapoint is in the first cluster, 0 otherwise.[3]

Q 1 **Spectral clustering:**

- Provide an initial 30×30 adjacency matrix $A^{\text{spectral},I}$ of an undirected graph (symmetric matrix with $0-1$ entries). Perform spectral clustering with the normalized Laplacian matrix on the data points and provide the vector $c^{\text{spectral},I}$ of cluster assignments you obtained. $c^{\text{spectral},I}$ **should have an equal number of** 1**'s and** 0**'s.**
- Add anywhere between 1 to 3 edges to $A^{\text{spectral},I}$ to create the new adjacency matrix $A^{\text{spectral},II}$. Run spectral clustering on this modified dataset to get new cluster assignment vector $c^{\text{spectral},II}$.
- **Goal:** $c^{\text{spectral},II}$ **and** $c^{\text{spectral},I}$ **must vary by over 30%. That is,**

$$\min_{C=c^{\text{spectral},II}, C=\mathbf{1}-c^{\text{spectral},II}} \frac{1}{30}\sum_{t=1}^{30} \mathbb{1}_{\{c_t^{\text{spectral},I} \neq C_t\}} \geq 0.3$$

Deliverables: Submit a **writeup** explaining the way you generated the data points and the corresponding modifications, and why you expected the new datasets to result in significantly different clusterings. **Kindly include a depiction** of both the initial and the modified **graphs**. Color code the nodes according to the spectral cluster assignments for the corresponding graphs. The picture with the explanation will carry most of the grade.

Also submit the adjacency matrix and the the modified adjacency matrix produced by adding the extra 1 to 3 edges; and the cluster assignments you obtained by running the algorithms over the initial and modified datasets.

Specifically, submit your datasets as csv files obeying the following requirements. `AspectralI.csv` and `AspectralI.csv` must each be 30 lines long, each line consisting of 30 comma-separated values indicating the initial and modified adjacency matrix. Also submit `cspectralI.csv` and `cspectralII.csv`, each 30 lines, where each line contains one value that is either 0 or 1, indicating the cluster assignment of the corresponding original point.

[3] It's up to you which cluster is the "first" one, so in this sense the cluster labels are arbitrary; we just need to know which points are in different clusters and which points are in the same cluster.

Q 2 **EM for Mixture of Poisson distributions:**

Story: You own a candy cane store. Each day customers walk into your store and buy candies and based on how many candies customers bought you want to model customers into K groups. Knowing about mixture models, you decide to model the problem using mixture of K poisson distributions. Here is the generative story:

For $t = 1$ to n

Draw group c_t for t^{th} customer from mixture distribution π over the K groups. (ie. $c_t \sim \pi$)

Draw $x_t \sim \text{Poisson}(x_t; \lambda_{c_t})$

End

That is, first each customer is assigned a group from 1 to K by randomly sampling from mixture distribution π. Next, the number of candies bought by customer in the group c_t is drawn as a poisson distribution with parameter λ_{c_t}. Poisson distribution is a distribution over natural numbers given by:

$$P(X_t = x_t; \lambda) = \frac{\lambda^{x_t} e^{-\lambda}}{x_t!}$$

Your goal for this problem is to write down the updates in the E and M step while running EM algorithm to compute the MLE for this mixture model given n data points $x_1, \ldots, x_n$ denoting the number of candies the customers bought from your store. Note that the parameters for the model are π and $\lambda_1, \ldots, \lambda_K$.

(a) **Write down the E-step update for Q_t's. That is write down what $Q_t^{(i)}(c_t = k)$ is for any given iteration i (in terms of parameters from previous iteration).**

(b) **Write down the M-step update for π**

(c) **Write down the M-step update for $\lambda_1, \ldots, \lambda_K$**

Hint: We already did few examples of mixture models in class and we saw that E-step and update for π in M-step are straightforward and we derived them generically for mixture models. So bulk of this question is to write update for λ_k's.