CS4786/5786: Machine Learning for Data Science, Fall 2016
4/4/2016: Assignment 4: EM algorithm for Mixture Models

Instructions Due at 11:59pm October 25th 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 professor. You have a slip day of at most one day for the assignment. The assignment is to be done individually. 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 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.[1]

[1]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.

Q 1 **EM for Mixture of Poisson distributions:**

Story: The case of the K typists You have a collection of documents (each a page long) typed by K different typists. Each typist has his/her own skill level and have different number of expected typos they make per page. While you know that there are K different typists, you dont know who typed each page, you don't know what each typists skill level is and you dont even know how many pages each typist has typed. Is is well known that if one looks at the distribution of number of mistakes per page made while typing for a person, it follows the so called Poisson distribution. The Poisson distribution is a distribution over natural numbers given by:

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

So for instance, if there was only one typist, if you count number of mistakes X_t on page t, it follows the above described distribution. However there are K typists and you dont know who typed which page. Knowing about mixture models, you decide to model the problem using mixture of K poisson distributions. Here is the generative model:

For $t = 1$ to n

Draw typist identity c_t for t^{th} page from mixture distribution π over the K possible typists. (ie. $c_t \sim \pi$)

Draw $x_t \sim \text{Poisson}(x_t; \lambda_{c_t})$, the number of typos on that page you counted.

End

Your goal for this problem is to write down the updates in the E and M step while running EM algorithm for this mixture model given n data points $x_1, \ldots, x_n$ denoting the number of typos on each of the n pages. Note that the parameters for the model are π and $\lambda_1, \ldots, \lambda_K$.

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

(b) M-step for MLE update for π in any mixture model is the same, it is $\pi_k = \sum_{t=1}^{n} Q_t^{(i)}(k)/n$. **Write down the M-step for MLE update of $\lambda_1, \ldots, \lambda_K$.** . (If you only write final answer or do not follow the steps below you will get a 0) Hint/steps:

i. First write down the for any given $k \in \{1, \ldots, K\}$, $\lambda_k^{(i)}$ as a solution to a maximization problem. (this is just writing down the M-step as was done in lecture notes or in class slides). (Eg. $\lambda_k = \underset{\lambda>0}{\text{argmax}} \ldots$

ii. Next to get the solution to the above optimization problem, set the partial derivative w.r.t. each of the λ_k's to 0 and work out what the values of λ_k's should be.

(c) Above we did the MLE estimate. Now say you had prior knowledge that all the λ_k's are drawn from an exponential distribution with parameter α (known to you). That is, each λ_k has prior density function: $p(x) = \alpha e^{-\alpha x}$, the exponential distribution.
In this case, write down what the M-step for MAP update for $\lambda_1, \ldots, \lambda_K$ are.
Hint: Use the steps you used for the previous sub-problem, but account for the prior.