CS 189 Introduction to Machine Learning
Fall 2019 Jennifer Listgarten & Stella Yu **DIS9**

1 Imputation of Missing Data using EM

This question is about adapting EM to a discrete problem of missing data.

Recall that in the case of a mixture of Gaussians we did soft imputation of the individual cluster assignments in the E-step and then estimation of $\boldsymbol{\theta}$, the set of parameters defining the individual Gaussians and the prior for Z, the random variable defining the cluster assignment, in the M-step. We iterates this process until convergence. EM, however, can be generalized to many other settings involving hidden variables and parameter estimation. In the mixture of Gaussian case, our hidden variables were the cluster assignments. In the following problem, we will explore how to apply EM to the setting where our hidden variables are missing data instead.

Suppose we have $\mathbf{Y} = [Y_1, Y_2, Y_3, Y_4, Y_5]$ where Y_i are random variables which are jointly distributed multinomially with probabilities $(\frac{1}{8}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{3}{8} + \frac{\theta}{4})$. Recall that for a multinomial distribution is a generalization of the binomial distribution and with a probability mass function parameterized by event probabilities $p_1, \ldots, p_k$. The PMF is $p(y_1, \ldots, y_k; p_1, \ldots, p_k) = \frac{n!}{y_1! \ldots y_5!} p_1^{y_1} \cdots p_k^{y_k}$.

In this problem, we observe data coming from an experiment that had 8 observations:

$$\mathbf{y} = [y_1, y_2, y_3, y_4, y_5] = [?, ?, 2, 2, 3]$$

where y_1 and y_2 are missing from our observations. Because there were 8 observations taken, we know that $y_{sum} = y_1 + y_2 = 1$ but we don't know which one is 1 and which one is 0.

To be clear, we know what the distribution is, but we don't know the parameter θ in the distribution and we don't have values for the first two categories.

1. **What is the log likelihood function, i.e. $p(\mathbf{y}|\theta)$?**

2. Recall that the EM algorithm iterates between soft imputation for our unobserved variables (E-step) and performing parameter estimation via maximization (M-step). In the E-step for the mixture of Gaussian case, we computed $p_{\boldsymbol{\theta}}(Z_i = k|X = x_i)$ where Z_i is a Bernouilli random variable determining the cluster assignment for x_i. Here our unobserved/hidden variables are Y_1 and Y_2, so we would like to compute $q_k^{t+1} = p_{\boldsymbol{\theta}}(Y_1 = k, Y_2 = \bar{k}|y_{sum})$ for $k = 0, 1$ since we know $y_1 + y_2 = 1$. Here $\bar{k}$ just means the opposite so $\bar{k} = 1$ if $k = 0$ and $\bar{k} = 0$ if $k = 1$. **Derive the q_0^{t+1}, q_1^{t+1} given the current $\boldsymbol{\theta}^t$.**

3. Recall that in the M-step we update our estimate for the parameters θ by maximizing the expected complete log-likelihood: $\theta^{t+1} = \operatorname{argmax}_\theta \mathbb{E}_q \left[\log(p_{\boldsymbol{\theta}}(\mathbf{Y}|y_{sum}, y_3, y_4, y_5))\right]$. **Write the**

expression for the complete log-likelihood and the closed form expression for the expected complete log-likelihood in terms of $q_0^{t+1}, q_1^{t+1}, y_3, y_4, y_5$, and θ?

4. **Maximize the expression for the expected complete log-likelihood to obtain an expression for θ^{t+1}.**

5. Using q_0^{t+1}, q_1^{t+1} computed in the E-step, **obtain a reasonable estimate for y_1 and y_2 and justify your answer.**

6. Let's consider how we might approach the problem using the MLE directly. One way to do this would be to marginalize out Y_1 and Y_2 of the log-likelihood by summing over all possible pairs: $(Y_1 = 0, Y_2 = 1)$ and $(Y_1 = 1, Y_2 = 0)$. **Write the expression for the MLE minimization for θ. Explain how you would compute an estimate for θ, but no need to compute it. How would you go from there to imputing Y_1 and Y_2?**

2 Coin tossing with unknown coins

This question is about adapting EM and the spirit of k-means to a discrete problem of tossing coins.

We have a bag that contains two kinds of coins that look identical. The first kind has probability of heads p_a and the other kind has probability of heads p_b, but we don't know these. We also don't know how many of each kind of coin are in the bag; so the probability, α_a, of drawing a coin of the first type is also unknown (and since $\alpha_a + \alpha_b = 1$, we do not need to separately estimate α_b, the probability of drawing a coin of the second type).

What we have is n pieces of data: for each data point, someone reached into the bag, pulled out a random coin, tossed it ℓ times and then reported the number h_i which was the number of times it came up heads. The coin was then put back into the bag. This was repeated n times. The resulting n head-counts $(h_1, h_2, h_3, \dots, h_n)$ constitute our data.

Our goal is to estimate p_a, p_b, α_a from this data in some reasonable way.

For this problem, the binomial distribution can be good to have handy:

$$P(H = h) = \binom{\ell}{h} p^h (1-p)^{\ell-h}$$

for the probability of seeing exactly h heads having tossed a coin ℓ times with each toss independently having probability p of turning up heads. Also recall that the mean and variance of a binomial distribution are given respectively by ℓp and $\ell p(1-p)$.

1. **How would you adapt the main ideas in the k-means algorithm to construct an analogous approach to estimating $\widehat{p}_a, \widehat{p}_b, \widehat{\alpha}_a$ from this data set? Give an explicit algorithm, although it is fine if it is written just in English.**

2. Suppose that the true $p_a = 0.4$ and the true $p_b = 0.6$ and $\alpha_a = 0.5$, and $\ell = 5$. For $n \to \infty$, **will your "k-means" based estimates (those from the preceding question) for $\widehat{p}_a$ and $\widehat{p}_b$ yield the correct parameter estimates (namely, $\widehat{p}_a = 0.4$ and $\widehat{p}_b = 0.6$)? Why or why not?**

Hint: Draw a sketch of the typical histograms of the number of heads of each coin on the same axes.

3. How would you adapt the EM for Gaussian Mixture Models that you have seen to construct an EM algorithm for estimating $\widehat{p}_a, \widehat{p}_b, \widehat{\alpha}_a$ from this data set?

   You don't have to solve for the parameters in closed form, but (i) **write down the E-step update equations (i.e. write down the distributions that should be computed for the E-step — not in general, but specifically for this problem) and (ii) the objective function that gets maximized for the M-step and also what you are maximizing with respect to (again, not just the general form, but specific to this problem).** If you introduce any notation, be sure to **explain what everything means. Explain in words what the E- and M-steps are doing on an intuitive level.**