STAT 154: Homework 3

Release date: **Thursday, February 21**

Due by: **11 PM, Wednesday, Mar 6**

This homework follows regular submission format, i.e., it is to be submitted by each student in the class *individually*–no teams!

1 A closer look at EM (25 points)

In this question, we consider a simple mixture model and work our way through a derivation of the EM updates.[1] *While the question looks long, please be patient in reading it–the description itself will help you understand EM better and also different parts of the problem only appear long because of the detailed explanation.*

We work with the following simple two mixture model:

$$\begin{aligned} Z &\sim \text{Bernoulli}(1-w)+1 \\ X|Z=1 &\sim \mathcal{N}(\mu_1,1), \quad \text{and} \\ X|Z=2 &\sim \mathcal{N}(\mu_2,1), \end{aligned} \tag{1}$$

where Z denotes the label of the Gaussian from which X is drawn. Given a set of observations only for X (i.e., the labels are unobserved), our goal is to infer the maximum-likelihood parameters for μ_1, μ_2 and w. Note that to simplify your calculations, we have fixed the variance parameter and assumed it to be known.

(a) (3 points) Let $\theta = (\mu_1, \mu_2, w)$ denote the parameters of the model. **Write down the expressions of the joint likelihood $p(X = x, Z = 1; \theta)$ and $p(X = x, Z = 2; \theta)$. What is the marginal likelihood $p(X = x; \theta)$ and the log-likelihood $\ell(X = x; \theta)$? Given n i.i.d. samples $\{x_1, \ldots, x_n\}$, write the expression for the log-likelihood $\ell(X_1 = x_1, \ldots, X_n = x_n; \theta)$.**

(b) (4 points) To simplify notation, from now on, we use the notation

$$\ell(x;\theta) = \ell(X = x;\theta), \quad \text{and} \quad p(x,k;\theta) = p(X = x, Z = k;\theta).$$

Let q denote a distribution on the (hidden) labels $\{Z_i\}_{i=1}^n$ given by

$$q(Z_1 = z_1, \ldots, Z_n = z_n) = \prod_{i=1}^{n} q_i(Z_i = z_i). \tag{2}$$

[1] The question borrows ideas from one of the homeworks of machine learning class CS 189, Spring 2018.

Note that since $Z \in \{1, 2\}$, q has n parameters, namely $\{q_i(Z_i = 1), i = 1, \ldots, n\}$. **Show that for a given point x_i, we have**

$$\ell(x_i; \theta) \geq \mathcal{F}_i(\theta; q_i) := \underbrace{\sum_{k=1}^{2} q_i(k) \log p(x_i, k; \theta)}_{\mathcal{L}(x_i; \theta, q_i)} + \underbrace{\sum_{k=1}^{2} q_i(k) \log \left(\frac{1}{q_i(k)} \right)}_{H(q_i)}, \tag{3}$$

where $H(q_i)$ denotes the Shannon-entropy of the distribution q_i. Thus **conclude that we obtain the following lower bound on the log-likelihood:**

$$\ell(\{x_i\}_{i=1}^n; \theta) \geq \mathcal{F}(\theta; q) := \sum_{i=1}^{n} \mathcal{F}_i(\theta; q_i). \tag{4}$$

Hint: Jensen's inequality, the concave-$\cap$ nature of the log, and reviewing lecture notes might be useful.

(c) (2 points) The EM algorithm can be considered a coordinate-ascent[2] algorithm on the lower bound $\mathcal{F}(\theta; q)$ derived in the previous part, where we ascend with respect to θ and q in an alternating fashion. More precisely, one iteration of the EM algorithm is made up of 2-steps:

$$q^{t+1} = \arg\max_{q} \mathcal{F}(\theta^t; q) \qquad \text{(E-step)}$$
$$\theta^{t+1} \in \arg\max_{\theta} \mathcal{F}(\theta; q^{t+1}). \qquad \text{(M-step)}$$

Given an estimate θ^t, the previous part tells us that $\ell(\{x_i\}_{i=1}^n; \theta^t) \geq \mathcal{F}(\theta^t; q)$. **Verify that equality holds in this bound if we plug in $q(Z_1 = z_1, \ldots, Z_n = z_n) = \prod_{i=1}^n p(Z = z_i | X = x_i; \theta^t)$ and hence we can conclude that**

$$q^{t+1}(Z_1 = z_1, \ldots, Z_n = z_n) = \prod_{i=1}^{n} p(Z = z_i | X = x_i; \theta^t). \tag{5}$$

is a valid maximizer for the problem $\max_q \mathcal{F}(\theta^t; q)$ and hence a valid E-step update.

(d) (2 points) **Derive the expressions for $p(Z = 1 | X = x_i; \theta^t)$ and $p(Z = 2 | X = x_i; \theta^t)$ to complete the E-step computations where $\theta^t = (\mu_1^t, \mu_2^t, w^t)$.**

(e) (3 points) We now discuss the M-step. Using the definitions from equations (3) and (4), we have that

$$\mathcal{F}(\theta; q^{t+1}) = \sum_{i=1}^{n} (\mathcal{L}(x_i; \theta, q_i^{t+1}) + H(q_i)) = H(q^{t+1}) + \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}_i; \theta, q_i^{t+1}),$$

[2] A coordinate-ascent algorithm is just one that fixes some coordinates and maximizes the function with respect to the others as a way of taking iterative improvement steps.

where we have used the fact that entropy in this case is given by $H(q^{t+1}) = \sum_{i=1}^{n} H(q_i^{t+1})$. Notice that although (as computed in previous part), q^{t+1} depends on θ^t, the M-step only involves maximizing $\mathcal{F}(\theta; q^{t+1})$ with respect to just the parameter θ while keeping the parameter q^{t+1} fixed. Now, noting that the entropy term $H(q^{t+1})$ does not depend on the parameter θ, we conclude that the M-step simplifies to solving for

$$\arg\max_{\theta} \underbrace{\sum_{i=1}^{n} \mathcal{L}(\mathbf{x}_i; \theta, q_i^{t+1})}_{=:\mathcal{L}(\theta; q^{t+1})}.$$

We use the simplified notation

$$q_i^{t+1} := q_i^{t+1}(Z_i = 1) \quad \text{and} \quad 1 - q_i^{t+1} := q_i^{t+1}(Z_i = 2)$$

and recall that $\theta = (\mu_1, \mu_2, w)$.

Show that the expression for $\mathcal{L}(\theta; q^{t+1})$ for the 2-mixture case is given by

$$\mathcal{L}((\mu_1, \mu_2, w); q^{t+1}) = C + \sum_{i=1}^{n} \left[q_i^{t+1} \left(\log w - \frac{(x_i - \mu_1)^2}{2} \right) + (1 - q_i^{t+1}) \left(\log(1 - w) - \frac{(x_i - \mu_2)^2}{2} \right) \right],$$

where C is a constant that does not depend on θ or q^{t+1}.

(f) (6 points) Using the expression of $\mathcal{L}$ from the previous part, **derive the expressions for the gradients of $\mathcal{L}(\theta; q^{t+1})$ with respect to μ_1, μ_2, w. By setting these gradients to zero, show that the M-step updates are given by**

$$\mu_1^{t+1} = \frac{\sum_{i=1}^{n} q_i^{t+1} x_i}{\sum_{i=1}^{n} q_i^{t+1}}, \quad \mu_2^{t+1} = \frac{\sum_{i=1}^{n} (1 - q_i^{t+1}) x_i}{\sum_{i=1}^{n} (1 - q_i^{t+1})}, \quad \text{and} \quad w^{t+1} = \frac{\sum_{i=1}^{n} q_i^{t+1}}{n}.$$

This complete the derivation of EM updates in the simpler model introduced in the beginning of the problem.

(g) (5 points) We now see EM in action and compare it with K-means. (No code is required in the submission.) Generate 1000 samples from the following mixture of two Gaussians in two dimensions:

$$\begin{aligned} Z &= \text{Bernoulli}(0.5) + 1 \\ X|Z = 1 &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \\ X|Z = 2 &\sim \mathcal{N}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \right) \end{aligned} \tag{6}$$

where I_2 denotes the identity matrix in two dimensions. **Generate both (Z, X) for the data and scatter plot the X values with color based on Z. Run K-means on X with $K = 2$ and report the cluster centers and scatter plot the data**

with estimated labels. Run EM on X to fit two clusters too (use Mclust package with $G = 2$) and report the mean parameters and scatter plot the data with estimated labels. Note that we need 3 plots for this part. Justify qualitatively why the cluster centers obtained by K-means and EM and the estimated label are different.