STA414/2104: Practice Problems 5

February, 2018

These practice problems are not for credit. Students may complete independently, in groups, or however they like. Treat these as representative of what might be on the tests.

Questions involving coding may be completed in any language. If you would like help regarding your language of choice from the course team, use R or Python. There will not be any *code*-related questions on the tests, but you will be asked about comuptational algorithms.

1. (a) Implement the K-Means algorithm discussed in lecture 7

   (b) Fit this model to the Iris data. Use the code given in Problems 4 as a guide- although you may wish to use Python instead of R, it's up to you.

   (c) Produce a plot of the predicted group memberships, and compare this to what the language-standard kmeans function produces (either in R or Python). Use the plots shown in the KNN code example to guide you.

   (d) Produce a *confusion matrix*, a 3×3 matrix comparing the predicted vs actual classifications.

   (e) Compare this to the results using the Gaussian Mixture Model in the code example in class. Which gives better results?

2. Prove the following statements that we made about Gaussian Mixture Models in the lecture. Chapter 9 of Bishop (2006) can walk you through some of these, as can the Bishop lecture slides that were included in this week's materials for reference

   (a) Derive the Maximum Likelihood Estimates for $\pi_k, \boldsymbol{\mu}_k$ and Σ_k,

      (i) By maximizing the marginal log-likelihood like we did before we introduced the EM algorithm

      (ii) By maximizing the expected complete-data log-likelihood in the M-Step of the EM algorithm

      Don't forget to incorporate the constraint $\sum_{k=1}^{K} \pi_k = 1$.

   (b) Show that

   $$E_{\mathbf{z}|\mathbf{x}}(z_{nk}) = \hat{z}_{nk}$$

   Use Bayes' theorem to derive the actual distribution of $z_{nk}|\mathbf{x}_n$, and use this to argue the statement- it's much cleaner to do it this way then to try to sum things directly.

3. (a) Extend the code example from class to work with K groups, and any input dataset.

   (b) Fit GMMs with $K = 2, 3, 4, \ldots$ to the Iris dataset. Look at the predicted classifications; discuss how you might choose the "best" value of K, if you didn't know the true classifications

   (c) Fit GMMs with $K = 2, 3, 4, \ldots$ to the Old Faithful dataset, obtained in R using `data(faithful)`. This is a 2-d dataset, so you can plot it and look at the classifications by colouring the points. What happens when you increase K beyond the "correct" answer of $K = 2$?

4. In this question, you will prove that each iteration of the EM algorithm is guaranteed to increase (or at least, not decrease) the log-likelihood. Let $Y = (X, Z)$ be the complete data, consisting of observed data X and missing or latent data Z. We wish to maximize the marginal likelihood for X:

$$\ell_x(\theta) = \log f(X|\theta)$$

with respect to θ.

(a) Show that $\ell_x(\theta)$ can be written

$$\ell_x(\theta) = \log f(X, Z|\theta) - \log f(Z|X, \theta)$$

(b) Take conditional expectations on both sides with respect to $Z|X, \theta^t$, to obtain

$$\ell_x(\theta) = Q(\theta, \theta^t) - H(\theta, \theta^t)$$

where $Q(\theta, \theta^t)$ was defined in lecture (the expected complete-data log likelihood given the observed data and current parameter estimates) and

$$H(\theta, \theta^t) = E_{z|x,\theta^t}(\log f(Z|X, \theta))$$

This notation can be tricky- note that both Q and H are functions of both θ and θ^t- the latter appears in the expectation operator, and can be easy to miss.

(c) Now we write

$$\ell_x(\theta^{t+1}) - \ell_x(\theta^t) = (Q(\theta^{t+1}, \theta^t) - Q(\theta^t, \theta^t)) - (H(\theta^{t+1}, \theta^t) - H(\theta^t, \theta^t))$$

We will prove our statement by arguing that $\ell_x(\theta^{t+1}) - \ell_x(\theta^t) \geq 0$.

(i) Explain why the first term is non-negative,

$$Q(\theta^{t+1}, \theta^t) - Q(\theta^t, \theta^t) \geq 0$$

(ii) Recall *Jensen's Inequality*: for g a concave function (for example, $g(x) = \log(x)$, hint hint),

$$E(g(X)) \leq g(E(X))$$

Use this fact to show that

$$H(\theta^{t+1}, \theta^t) - H(\theta^t, \theta^t) \leq 0$$

thus proving the statement.

5. This example of EM is taken from Bishop (2006), section 9.3, page 445. In this question we will derive and implement an EM algorithm for a mixture of Bernoulli distributions. Suppose we observe a collection of length-D binary vectors $\mathbf{x}_n = (x_{n1}, \ldots, x_{nD})$, where each $x_{in} \sim Bern(\mu_i)$ independently; that is, each element of $\mathbf{x}_n$ is a Bernoulli random variable with a different probability parameter. The distribution of $\mathbf{x}_n$ for any $n = 1 \ldots N$ is then

$$p(\mathbf{x}_n|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

We wish to model these vectors as having come from one of $K > 1$ component distributions. Like before, define latent binary variables $\mathbf{z}_n = (z_{n1}, \ldots, z_{nK})$ indicating whether $\mathbf{x}_n$ belongs to each of groups $1 \ldots K$. The marginal distribution of these latent variables is given by

$$p(\mathbf{z}_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{nk}}$$

where the π_k are the prior group probabilities, $\sum_{k=1}^{K} \pi_k = 1$. Finally, like in class, write the conditional distribution of $\mathbf{x}_n$ given $\mathbf{z}_n$ as

$$p(\mathbf{x}_n|\mathbf{z}_n) = \prod_{k=1}^{K} p(\mathbf{x}_n|\boldsymbol{\mu}_k)^{z_{nk}}$$

where each $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kD})$ is the parameter vector for the k^{th} component distribution.

(a) Show that the marginal distribution of $\mathbf{x}_n$ is

$$p(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)$$

(b) Show that the marginal (incomplete-data) log-likelihood is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\pi}|\mathbf{X}) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)\right)$$

(c) Show that the complete-data log-likelihood is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\pi}|\mathbf{X}, \mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left(\log \pi_k + \sum_{i=1}^{D}\big(x_{ni}\log \mu_{ki} + (1-x_{ni})\log(1-\mu_{ki})\big)\right)$$

(d) Because this is linear in the missing data, we know the E-Step will be tractable. Evaluate $\hat{z}^t = E(z_{nk}|\boldsymbol{\mu}^t, \boldsymbol{\pi}^t, \mathbf{X})$ using Bayes' Theorem, and sub it in to the complete-data log likelihood, to find $Q(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\mu}^t, \boldsymbol{\pi}^t)$

(e) Maximize the $Q(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\mu}^t, \boldsymbol{\pi}^t)$ function with respect to $\boldsymbol{\mu}, \boldsymbol{\pi}$, to find that the M-Step updates are

$$\pi_k^{t+1} = \frac{1}{N}\sum_{n=1}^{N} \hat{z}_{nk}^t$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \hat{z}_{nk}^t \mathbf{x}_n}{\sum_{n=1}^{N} \hat{z}_{nk}^t}$$

You will implement this EM algorithm in a later assignment for clustering MNIST (as Bishop shows in the section where this problem is from).