CMSC 35400 / STAT 37710

Spring 2020

Homework 7

You must clearly indicate where your solutions to individual subproblems are in gradescope. If you force the graders to do this for you, points will be deducted from your total.

1. **(Mixture models and EM algorithm)** Consider a 1-dimensional Gaussian Mixture Model with 2 clusters and parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)$. Here (w_1, w_2) are the mixing weights, and $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)$ are the centers and variances of the clusters. We are given a dataset $D = \{x_1, x_2, x_3\} \subset \mathbb{R}$. In this problem, you will apply the EM-algorithm to find the parameters of the Gaussian mixture model.

   a) Write down the complete log-likelihood that is being optimized, for this problem.

   Assume that the dataset D consists of the following three points, $x_1 = 1, x_2 = 10, x_3 = 20$. At some step in the EM-algorithm, we compute the expectation step which results in the following matrix:

$$R = \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}$$

   where r_{ic} denotes the probability of x_i belonging to cluster c. In the next questions, leave all results unsimplified, i.e. in fractional form.

   b) Given the above R for the expectation step, write the result of the maximization step for the mixing weights w_1, w_2. You can use the equations for maximum likelihood updates without proof.

   c) Do the same for μ_1, μ_2. Given the above R for the expectation step, write the result of the maximization step for the centers μ_1, μ_2 . You can use the equations for maximum likelihood updates without proof.

   d) Do the same for σ_1^2, σ_2^2. Given the above R for the expectation step, write the result of the maximization step for the variance values σ_1^2, σ_2^2. You can use the equations for maximum likelihood updates without proof.

2. **(A different perspective on EM algorithm)** In this question you will show that EM can be seen as an iterative algorithm which maximizes a lower bound on the log-likelihood. We will treat any general model $P(X, Z)$ with observed variables X and latent variables Z. For simplicity, we will assume that Z is discrete and takes values in $\{1, 2, \ldots, m\}$. If we observe X, the goal is to maximize the log-likelihood

$$\ell(\theta) = \log P(x; \theta) = \log \sum_{z=1}^{m} P(x, z; \theta)$$

with respect to the parameter vector θ. $Q(Z)$ denotes any distribution over the latent variables.

a) Show that if $Q(z) > 0$ when $P(x, z) > 0$, then it holds that

$$\ell(\theta) \geq \mathbb{E}_Q[\log P(X, Z)] - \sum_{z=1}^{m} Q(z) \log Q(z)$$

Hence, we have a bound on the log-likelihood parametrized by a distribution $Q(Z)$ over the latent variables. (*Hint: Consider using Jensen's inequality* $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$ *for convex function* ϕ).

b) Show that for a fixed θ, the lower bound is maximized for $Q^*(Z) = P(Z \mid X; \theta)$. Moreover, show that the bound is exact (holds with equality) for this specific distribution $Q^*(z)$.

c) Show that if we optimize with respect to Q and θ in an alternating manner, this corresponds to the EM procedure. Discuss what this implies for the monotonicity and convergence properties of EM.

3. (Learning Bayesian network) Consider learning a Bayesian network of four variables X, Y, Z, W given a data set sampled from the joint distribution. The empirical pairwise mutual information has been computed as

$$\hat{I}(X;Y) = 0.32, \qquad \hat{I}(X;Z) = 0.38, \qquad \hat{I}(X;W) = 0.27,$$
$$\hat{I}(Y;Z) = 0.39, \qquad \hat{I}(Y;W) = 0.27, \qquad \hat{I}(Z;W) = 0.39.$$

Answer the following questions and briefly justify each answer.

a) Draw a Bayesian network that maximizes the likelihood of the observed data.

b) Draw a tree-shaped Bayesian network that maximizes the likelihood of the observed data.