Exercises
Learning and Intelligent Systems
SS 2017

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Gunnar Rätch and Prof. Dr. Thomas Hofmann
Web: `http://las.inf.ethz.ch/teaching/lis-s17/`
Email questions to:
Gideon Dresdner, dgideon@student.ethz.ch

Series 6, June 8th, 2017
(Latent variable models and EM)

Please turn in solutions until Thursday, June 15th. It is not mandatory to submit solutions and sample solutions will be published in two weeks. If you choose to submit your solution, please send an e-mail from your ethz.ch address with subject Exercise6 containing a PDF (LATEX or scan) to `dgideon@student.ethz.ch` until submission time.

Problem 1 (EM for Naïve Bayes):

Assume that you want to train a naïve Bayes model on data with missing class labels. Specifically, there are k *binary* variables $X_1, \dots X_k$ corresponding to the features, and a variable Y taking on values in $\{1, 2, \dots, m\}$ denoting the class. Let us denote the set of model parameters as $P(X_i = 1|Y = y) = \theta_{i|y}$ and $P(Y = y) = \theta_y$.

You are given n data points $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \{0, 1\}^k$ and $y_i \in \{1, 2, \dots, m, \times\}$. The value $\times$ means that the label of the data point is missing.

- Write down the log-likelihood $\ell(\theta)$ of the data as a function of the parameters θ.
- Recall that the E-step of the EM algorithm computes the posterior over the unknown variables when we *fix* the parameters θ. Compute these probabilities $\gamma_j(\mathbf{x}_i) = P(Y = j \mid \mathbf{x}_i; \theta)$ for j s.t. $y_i = \times$.
- Once we have the quantities $\gamma_j(\cdot)$, we can compute the M-step update, which is computed as the maximizer θ^* of $\sum_{i=1}^n \sum_{j=1}^m \gamma_j(\mathbf{x}_i) \log P(\mathbf{x}_i, y_i = j; \theta)$. Show how to compute θ^*. Note that there are constraints on θ^* to make sure that the distributions are valid (non-negative and sum up to 1).

Problem 2 (EM for a 1D Laplacian Mixture Model):

In this problem you will derive the EM algorithm for a *one-dimensional* Laplacian mixture model. You are given n observations $x_1, \dots, x_n \in \mathbb{R}$ and we want to fit a mixture of m Laplacians, which has the following density

$$f(x) = \sum_{j=1}^{m} \pi_j f_L(x; \mu_j, \beta_j), \tag{1}$$

where $f_L(x; \mu_j, \beta_j) = \frac{1}{2\beta_j} e^{-\frac{1}{\beta_j}|x-\mu_j|}$, and the mixture weights π_j are a convex combination, i.e. $\pi_j \geq 0$ and $\sum_{j=1}^m \pi_j = 1$. For simplicity assume that the scale parameters $\beta_j > 0$ are known beforehand and thus *fixed*.

- Introduce latent variables so that we can apply the EM procedure.
- Analogously to the previous question, write down the steps of the EM procedure for this model. If some updates cannot be written analytically, give an approach on how to compute them.

  (*Hint: Recall a property of functions that makes them easy to optimize.*)

Problem 3 (A different perspective on EM [1]):

In this question you will show that EM can be seen as iteratively maximizing 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 the sake of simplicity, we will assume that Z is discrete and takes on values in $\{1, 2, \ldots, m\}$. If we are observe $X = \mathbf{x}$, the goal is to maximize the log-likelihood

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

with respect to the parameter vector θ. In what follows we will denote by $Q(Z)$ *any* distribution over the latent variables.

- Show that if $Q(z) > 0$ when $P(\mathbf{x}, z) > 0$, then it holds that *(Hint: Consider using Jensen's inequality)*

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

  Hence, we have a bound on the log-likelihood parametrized by a distribution $Q(Z)$ over the latent variables.

- 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 $Q^*(Z)$.

  (Hint: Do not forget to add Lagrange multipliers to make sure that Q^* is a valid distribution.)

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

[1] This is an advanced question.