Exercises
Introduction to Machine Learning
SS 2018

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `http://las.inf.ethz.ch/teaching/lis-s16/`
Email questions to:
Anastasia Makarova, anmakaro@inf.ethz.ch

Series 6, May 14th, 2018
(EM Algorithm and Semi-Supervised Learning)

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).

Solution 1:

The log-likelihood is equal to

$$
\begin{aligned}
\ell(\theta) &= \log P(\mathcal{D}) \\
&= \sum_{\substack{i=1 \\ y_i=\times}}^{n} \log P(\mathbf{x}_i; \theta) + \sum_{\substack{i=1 \\ y_i \neq \times}}^{n} \log P(\mathbf{x}_i, y_i; \theta) \\
&= \sum_{\substack{i=1 \\ y_i=\times}}^{n} \log \sum_{j=1}^{m} P(\mathbf{x}_i, Y = j; \theta) + \sum_{\substack{i=1 \\ y_i \neq \times}}^{n} \log P(\mathbf{x}_i, y_i; \theta) \\
&= \sum_{\substack{i=1 \\ y_i=\times}}^{n} \log \sum_{j=1}^{m} P(\mathbf{x}_i \mid Y = j; \theta) P(Y = j; \theta) + \sum_{\substack{i=1 \\ y_i \neq \times}}^{n} \log P(\mathbf{x}_i, y_i; \theta) \\
&= \sum_{\substack{i=1 \\ y_i=\times}}^{n} \log \sum_{j=1}^{m} \theta_j \prod_{l=1}^{k} \theta_{l|j}^{x_{i,l}} (1 - \theta_{l|j})^{1-x_{i,l}} + \sum_{\substack{i=1 \\ y_i \neq \times}}^{n} \log \theta_{y_i} \prod_{l=1}^{k} \theta_{l|y_i}^{x_{i,l}} (1 - \theta_{l|y_i})^{1-x_{i,l}}.
\end{aligned}
$$

To compute the requested posterior probabilities, note that by Bayes' rule

$$\begin{aligned}\gamma_j(\mathbf{x}_i) &= P(y_i = j \mid \mathbf{x}_i; \theta) \\ &= \frac{P(\mathbf{x}_i \mid y_i = j; \theta) P(y_i = j; \theta)}{P(\mathbf{x}_i; \theta)} \\ &= \frac{1}{Z} P(\mathbf{x}_i \mid y_i = j; \theta) P(y_i = j; \theta) \\ &= \frac{1}{Z} \theta_j \prod_{l=1}^{k} \theta_{l|j}^{x_{i,l}} (1 - \theta_{l|j})^{1 - x_{i,l}}.\end{aligned}$$

We then have to compute the normalizer Z so that $\sum_{j=1}^{m} \gamma_j(\mathbf{x}_i) = 1$. Note that for those data points $\mathbf{x}_i$ for which we are given the labels y_i we set the $\gamma_j(\mathbf{x}_i)$ to be a deterministic distribution, i.e. $\gamma_j(x_i) = [j = y_i]$.

To compute the M-step update we have to optimize the following quantity

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_j(\mathbf{x}_i) \log P(\mathbf{x}_i, y_i = j; \theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_j(\mathbf{x}_i) \big[\log \theta_j + \sum_{l=1}^{k} \log \theta_{l|j}^{x_{i,l}} (1 - \theta_{l|j})^{1 - x_{i,l}} \big]$$

with respect to the parameters θ. We form the Lagrangian by adding a multiplier λ to make sure that $\sum_{j=1}^{m} \theta_j = 1$:

$$\mathcal{L}(\theta, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_j(\mathbf{x}_i) \big[\log \theta_j + \sum_{l=1}^{k} \log \theta_{l|j}^{x_{i,l}} (1 - \theta_{l|j})^{1 - x_{i,l}} \big] + \lambda (\sum_{j=1}^{m} \theta_j - 1).$$

By setting the derivatives to zero we obtain:

$$\frac{\partial}{\partial_{\theta_{l|j}}} \mathcal{L}(\theta, \lambda) = \sum_{\substack{i=1 \\ x_{i,l}=1}}^{n} \gamma_j(\mathbf{x}_i) / \theta_{l|j} + \sum_{\substack{i=1 \\ x_{i,l}=0}}^{n} \gamma_j(\mathbf{x}_i) / (\theta_{l|j} - 1) = 0 \implies \theta_{l|j} = \frac{\sum_{i=1}^{n} [x_{i,l} = 1] \gamma_j(\mathbf{x}_i)}{\sum_{i=1}^{n} \gamma_j(\mathbf{x}_i)}$$

$$\frac{\partial}{\partial_{\theta_j}} \mathcal{L}(\theta, \lambda) = \sum_{i=1}^{n} \gamma_j(\mathbf{x}_i) / \theta_j + \lambda = 0 \implies \theta_j = -\frac{\sum_{i=1}^{n} \gamma_j(\mathbf{x}_i)}{\lambda}$$

From the constraint $\sum_{j=1}^{m} \theta_j = 1$ we find the correct multiplier to be $\lambda = -\sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_j(\mathbf{x}_i) = -n$.

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 $x_1, \ldots, x_n \in \mathbb{R}$ and we want to fit a mixture of m Laplacians, which has the following density

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

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.*)

Solution 2:

For each data point x_i, we introduce a latent variable $Y_i \in \{1, 2, \ldots, m\}$ denoting the component that point belongs to. For the E-step, we compute the posterior over the classes similarly to the previous problem, i.e.

$$\gamma_j(x_i) = P(y_i = j \mid x_i) \propto P(x_i \mid y_i = j)P(y_i = j) = \pi_j f_L(x_i; \mu_j, \beta_j).$$

Again, we have to normalize, so that the final posterior is equal to

$$\gamma_j(x_i) = \frac{\pi_j f_L(x_i; \mu_j, \beta_j)}{\sum_{l=1}^{m} \pi_l f_L(x_i; \mu_l, \beta_l)}.$$

In the M-step, we optimize

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_j(x_i) \log P(x_i, y_i = j) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_j(x_i) \log \pi_j f_L(x_i; \mu_j, \beta_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_j(x_i)(\log \pi_j - \frac{1}{\beta_j}|x_i - \mu_j|) + \text{const.} \end{aligned} \tag{1}$$

We add a Lagrange multiplier λ to make sure that $\sum_{j=1}^{m} \pi_j = 1$ and obtain the Lagrangian

$$\mathcal{L}(\pi, \mu, \lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_j(x_i)(\log \pi_j - \frac{1}{\beta_j}|x_i - \mu_j|) + \lambda(\sum_{j=1}^{m} \pi_j - 1).$$

Exactly as in the previous problem, by setting the gradient with respect to π_j to zero, we obtain

$$\frac{\partial}{\partial_{\pi_j}} \mathcal{L}(\pi, \mu, \lambda) = \sum_{i=1}^{n} \gamma_j(x_i)/\pi_j + \lambda = 0 \implies \pi_j = \frac{\sum_{i=1}^{n} \gamma_j(x_i)}{-\lambda}.$$

The multiplier is again equal to $\lambda = -n$ and we arrive at the same equation as in the last example. If we want to maximize (1) with respect to the variables μ_j, we have to solve m *separate* optimization problems, one for each μ_j. These m problems have the following form

$$\underset{\mu_j}{\text{maximize}} - \sum_{i=1}^{n} \frac{\gamma_j(x_i)}{\beta_j}|x_i - \mu_j|.$$

These are one-dimensional convex optimization problems (the negative of the objective is easily seen to be convex). While one can try solving this via an iterative process like subgradient descent, a direct approach is also possible if we observe that the function is piecewise linear and the breakpoints are $x_1, x_2, \ldots, x_n$. Hence, the optimum must be attained at one of these n points and we can simply set μ_j to the point x_i with the largest objective value.

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.

Solution 3:

For the first part, note that

$$\begin{aligned}
\ell(\theta) &= \log P(\mathbf{x}; \theta) \\
&= \log \sum_{z=1}^{m} P(\mathbf{x}, z; \theta) \\
&= \log \sum_{z=1}^{m} \frac{P(\mathbf{x}, z; \theta)}{Q(z)} Q(z) \\
&= \log \mathbb{E}_{Z \sim Q}[\frac{P(\mathbf{x}, z; \theta)}{Q(z)}] \\
&\geq \mathbb{E}_{Z \sim Q}[\log \frac{P(\mathbf{x}, z; \theta)}{Q(z)}] \\
&= \mathbb{E}_{Z \sim Q}[\log P(\mathbf{x}, z; \theta)] - \sum_{z=1}^{m} Q(z) \log Q(z),
\end{aligned}$$

where for the inequality we have used Jensen's inequality. Now, assume that we want to maximize the above with respect to Q, and let us add a multiplier λ to make sure that Q sums up to 1. Then, we have the following Lagrangian

$$\mathcal{L}(Q, \lambda) = \sum_{z=1}^{m} Q(z) \log P(\mathbf{x}, z; \theta) - \sum_{z=1}^{m} Q(z) \log Q(z) + \lambda(\sum_{z=1}^{m} Q(z) - 1).$$

[1] This is an advanced question.

By setting the derivative of the Lagrangian with respect to $Q(z)$ to zero, we have

$$\frac{\partial}{\partial_{Q(z)}}\mathcal{L}(Q,\lambda) = \log P(\mathbf{x},z;\theta) - 1 - \log Q(z) + \lambda = 0 \implies Q(z) = e^{\lambda-1}P(\mathbf{x},z;\theta).$$

Hence, we have that $Q(z) \propto P(\mathbf{x},z;\theta)$ and this is exactly the posterior $P(Z \mid \mathbf{x};\theta)$, which we had to show. It is also easy to see that the bound is tight, as

$$\mathbb{E}_{Z\sim Q}[\log \frac{P(\mathbf{x},z;\theta)}{Q(z)}] = \sum_{z=1}^{m} Q(z) \log \frac{P(\mathbf{x},z;\theta)}{Q(z)} = \sum_{z=1}^{m} P(z \mid \mathbf{x};\theta) \log \frac{P(z \mid \mathbf{x};\theta)P(\mathbf{x};\theta)}{P(z \mid \mathbf{x};\theta)} = \log P(\mathbf{x};\theta).$$

Then we can easily see the EM algorithm as optimizing the lower bound with respect to $Q(\cdot)$ and θ in an alternating manner. Specifically, if we optimize with respect to Q we have shown that the optimal Q is the posterior, and this is exactly the E-step. Optimizing with respect to θ for fixed Q is clearly equivalent to the M-step. As the lower bound is monotonically increased at every step the EM algorithm has to converge.