Exercises
Introduction to Machine Learning
SS 2019

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s19`
Email questions to:
omineeva@student.ethz.ch, aytunc.sahin@inf.ethz.ch

Series 7, May 20th, 2019 (Mixture Models, EM Algorithm)

Problem 1 (Mixture Models and Expectation-Maximization Algorithm):

Consider a one-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 (μ_1, σ_1^2), (μ_2, σ_2^2), are the centers and variances of the clusters. We are given a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} \subset \mathbb{R}$, and apply the EM-algorithm to find the parameters of the Gaussian mixture model.

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

$$\begin{aligned}\ln f(\mathcal{D} \mid (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)) = &\ln \Big\{ w_1 \mathcal{N}(\mathbf{x}_1; \mu_1, \sigma_1) + w_2 \mathcal{N}(\mathbf{x}_1; \mu_2, \sigma_2) \Big\} \\ &+ \ln \Big\{ w_1 \mathcal{N}(\mathbf{x}_2; \mu_1, \sigma_1) + w_2 \mathcal{N}(\mathbf{x}_2; \mu_2, \sigma_2) \Big\} \\ &+ \ln \Big\{ w_1 \mathcal{N}(\mathbf{x}_3; \mu_1, \sigma_1) + w_2 \mathcal{N}(\mathbf{x}_3; \mu_2, \sigma_2) \Big\}\end{aligned}$$

Assume that the dataset $\mathcal{D}$ consists of the following three points, $\mathbf{x}_1 = 1, \mathbf{x}_2 = 10, \mathbf{x}_3 = 20$. 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 $\mathbf{x}_i$ belonging to cluster c.
In the next questions, leave all results unsimplified, i.e. in fractional form.

2. 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.

$$w_1' = \frac{1}{3}(1 + 0.4 + 0) = \frac{1.4}{3}$$

$$w_2' = \frac{1}{3}(0 + 0.6 + 1) = \frac{1.6}{3}$$

3. 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.

   In general,

$$\mu_k' = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(x_n) x_n$$

where $N_k = \sum_{n=1}^{N} \gamma_k(x_n)$.

For this example,

$$\mu_1' = \frac{1}{1.4}(1 \cdot 1 + 0.4 \cdot 10 + 0 \cdot 20) = \frac{5}{1.4}$$

$$\mu_2' = \frac{1}{1.6}(0 \cdot 1 + 0.6 \cdot 10 + 1 \cdot 20) = \frac{26}{1.6}$$

4. 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.

   In general,

$$(\sigma_k^2)' = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \mu_k')(\mathbf{x}_n - \mu_k')^T$$

   where $N_k = \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)$.

   For this example,

$$(\sigma_1^2)' = \frac{1}{1.4}\left(1 \cdot (1 - \frac{5}{1.4})^2 + 0.4 \cdot (10 - \frac{5}{1.4})^2 + 0 \cdot (20 - \frac{5}{1.4})^2\right)$$

$$(\sigma_2^2)' = \frac{1}{1.6}\left(0 \cdot (10 - \frac{26}{1.6})^2 + 0.6 \cdot (10 - \frac{26}{1.6})^2 + 1 \cdot (20 - \frac{26}{1.6})^2\right)$$

5. The previous two questions are doing soft-EM. Calculate the maximization step of μ_1, μ_2 for hard-EM.

$$\mu_1' = \frac{1}{1}(1) = 1$$

$$\mu_2' = \frac{1}{2}(10 + 20) = 15$$

Problem 2 (Mixture Models and Maximum a Posteriori estimation):

Consider a mixture of K multivariate Bernoulli distributions with parameters $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K\}$, where $\boldsymbol{\mu}_k = \{\mu_{k1}, ..., \mu_{kd}\}$. You will use EM algorithm to compute MLE and MAP estimates.

1. What is the M step for μ_{ki} using MLE?
2. Now, suppose you want to do MAP estimation. What is the E step?
3. What is the M step for μ_{ki} using MAP? You can assume a Beta(α, β) prior.

Solution 2:

1.

We have K mixture components where each component is a vector of d independent Bernoullis. In other words,

$$p(x|\pi, \mu) = \sum_{k=1}^{K} \pi_k p(x|\mu) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{d} \mu_{ki}^{x_i}(1 - \mu_{ki})^{1-x_i}$$

Expected value of the complete data log-likelihood can be written as:

$$\mathbb{E}[\log(p(x, z|\pi, \mu))] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log \pi_k + \sum_{i=1}^{d}\left(x_{ni}\log \mu_{ki} + (1 - x_{ni})\log(1 - \mu_{ki})\right)\right)$$

where r_{nk} denotes the posterior probability from the E step. Note that the derivative of Bernoulli distribution is $\frac{x_{ni}}{\mu_{ki}} - \frac{(1-x_{ni})}{(1-\mu_{ki})}$Taking the derivative with respect to μ_{ki} and setting it to zero gives you

$$\mu_{ki} = \frac{\sum_{n=1}^{N} r_{nk}x_{ni}}{\sum_{n=1}^{N} r_{nk}}$$

2. The E Step is the same for the MLE case, namely

$$r_{nk} = \frac{\pi_k \prod_{i=1}^{d} \mu_{ki}^{x_{ni}}(1 - \mu_{ki})^{1-x_{ni}}}{\sum_{k=1}^{K} \pi_k \prod_{i=1}^{d} \mu_{ki}^{x_{ni}}(1 - \mu_{ki})^{1-x_{ni}}}$$

3.

According to Bayes' theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$
$$\log p(\boldsymbol{\theta}|\boldsymbol{X}) \propto \log p(\boldsymbol{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

Therefore, we need to add a log prior to the expected value of the complete data log-likelihood. The function we need to maximize is $\mathbb{E}[\log(p(x, z|\pi, \mu))] + \log p(\mu)$, where $p(\mu) = \prod_{k=1}^{K}\prod_{i=1}^{d} p(\mu_{ki})$ and

$$p(\mu_{ki}) = \frac{\mu_{ki}^{\alpha-1}(1 - \mu_{ki})^{\beta-1}}{\mathcal{B}(\alpha, \beta)}$$

We can write

$$\log p(\mu) = \sum_{k=1}^{K}\sum_{i=1}^{d}(\alpha - 1)\log \mu_{ki} + (\beta - 1)\log(1 - \mu_{ki}) - \log \mathcal{B}(\alpha, \beta)$$

We take derivative of the following expression with respect to μ_{ki} and set it to zero:

$$\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log \pi_k + \sum_{i=1}^{d}\left(x_{ni}\log \mu_{ki} + (1 - x_{ni})\log(1 - \mu_{ki})\right)\right) +$$
$$\sum_{k=1}^{K}\sum_{i=1}^{d}(\alpha - 1)\log \mu_{ki} + (\beta - 1)\log(1 - \mu_{ki})$$

which gives

$$\mu_{ki} = \frac{\sum_{n=1}^{N}(r_{nk}x_{ni}) + \alpha - 1}{\sum_{n=1}^{N}(r_{nk}) + \alpha + \beta - 2}$$

Problem 3 (A Different Perspective on EM):

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 the sake of 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

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

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

- Show that if $Q(z) > 0$ when $P(\mathbf{x}, z) > 0$, 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)*

- 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, 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

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

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.