Exercises
Introduction to Machine Learning
SS 2019

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

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

Solutions will be published on Monday, May 27th.

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*.

   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.

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.

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.

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

Problem 2 (Mixture Models and MAP estimation):

Consider a mixture of K multivariate Bernoulli distributions with parameters $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K\}$, where $\boldsymbol{\mu}_i = \{\mu_{i1}, ..., \mu_{iD}\}$. You will use EM algorithm to compute MLE and MAP estimates.

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

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.