Exercises
Introduction to Machine Learning
SS 2018

Series 7, May 22, 2018 (EM Convergence)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s18`
Email questions to:
Mohammad Reza Karimi, `mkarimi@ethz.ch`

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 `Exercise7` containing a PDF (LaTeX or scan) to `mkarimi@ethz.ch` until Tuesday, May 29, 2018.

*The problems marked with an asterisk * are intended for deeper understanding. One should look at these problems as an opportunity to have more insight into the theory. Note that these problems are not necessarily harder than the other ones.*

Problem 1 (EM for Censored Linear Regression):

Suppose you are trying to learn a model that can predict how long a program will take to run for different settings. In some situations, when the program is taking too long, you abort the program and just note down the time at which you aborted. These values are *lower bounds* for the actual running time of the program. We call this type of data **right-censored**. Concretely, all you know is that the running time $y_i \geq c_i$, where c_i is the censoring time. Written in another way, one can say $y_i = \min\{z_i, c_i\}$ where z_i is the true running time. Our goal is to derive an EM algorithm for fitting a linear regression model to right-censored data.

(a) Let $z_i = \mu_i + \sigma\varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 1)$. Suppose that we do not observe z_i, but we observe the fact that it is higher than some threshold. Namely, we observe the event $E = \mathbb{I}(z_i \geq c_i)$. Show that

$$\mathbb{E}[z_i \,|\, z_i \geq c_i] = \mu_i + \sigma R\left(\frac{c_i - \mu_i}{\sigma}\right)$$

and

$$\mathbb{E}[z_i^2 \,|\, z_i \geq c_i] = \mu_i^2 + \sigma^2 + \sigma(c_i + \mu_i)R\left(\frac{c_i - \mu_i}{\sigma}\right),$$

where we have defined

$$R(x) := \frac{\phi(x)}{1 - \Phi(x)}.$$

Here, $\phi(x)$ is the pdf of the standard Gaussian, and $\Phi(x)$ is its cdf.

(b) Derive the EM algorithm for fitting a linear regression model to right-censored data. Describe completely the E-step and M-step.

Problem 2 (Soft k-means, Revisited):

(a) Consider the following optimization problem:

$$\max_{\mathbf{c}\in\mathbb{R}^k} \sum_{i=1}^{k} v_i \log(c_i) \quad \text{s.t.} \quad c_i > 0,\ \sum_{i=1}^{k} c_i = 1,$$

where $\mathbf{v} \in \mathbb{R}_+^k$ is a vector of non-negative weights. Check that the M-step of soft k-means includes solving such an optimization problem.

(b) Let $\mathbf{c}^\star = \frac{1}{\sum_i v_i}\mathbf{v}$. Verify that $\mathbf{c}^\star$ is a probability vector.

(c) Show that the optimization problem is equivalent to the following problem:

$$\min_{\mathbf{c}\in\mathbb{R}^k} \mathrm{D}_{\mathrm{KL}}(\mathbf{c}^\star \| \mathbf{c}) \quad \text{s.t.} \quad c_i > 0,\ \sum_{i=1}^{k} c_i = 1,$$

(d) Using the properties of KL divergence, prove that $\mathbf{c}^\star$ is indeed the solution to the optimization problem.

Problem 3 (Yet another perspective on EM):

The EM algorithm is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables. Take a probabilistic model in which we denote all of the *observed* variables as $\mathbf{X}$ and all of the hidden variables as $\mathbf{Z}$ (here we assume $\mathbf{Z}$ is discrete, for the sake of simplicity). Let us assume that the joint distribution is $p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the set of all parameters describing this distribution (*e.g.* for a Gaussian distribution, $\boldsymbol{\theta} = (\mu, \Sigma)$). The goal is to maximize the likelihood function

$$p(\mathbf{X} \,|\, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta}).$$

(a) For an arbitrary distribution $q(\mathbf{Z})$ over the latent variables, show that the following decomposition holds:

$$\ln p(\mathbf{X} \,|\, \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{D}_{\mathrm{KL}}(q \| p_{\mathrm{post}}), \tag{1}$$

where $p_{\mathrm{post}} = p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta})$ is the posterior distribution. Also find the formulation of $\mathcal{L}(q, \boldsymbol{\theta})$.

(b) Verify that $\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X} \,|\, \boldsymbol{\theta})$, and that equality holds if and only if $q(\mathbf{Z}) = p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta})$.

(c) Suppose that the current value of the parameters is $\boldsymbol{\theta}_{\mathrm{curr}}$. Verify that in the E-step, the lower bound $\mathcal{L}(q, \boldsymbol{\theta}_{\mathrm{curr}})$ is maximized with respect to the distribution $q(\mathbf{Z})$, while keeping $\boldsymbol{\theta}_{\mathrm{curr}}$ fixed. Since the left-hand-side of (1) does not depend on $q(\mathbf{Z})$, maximizing $\mathcal{L}(q, \boldsymbol{\theta}_{\mathrm{curr}})$ will result in minimizing the KL divergence between q and p_{post}, which happens at $q^\star = p_{\mathrm{post}}$.

(d) Verify that in the M-step, the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ while keeping $q(\mathbf{Z})$ fixed, resulting in a new value of parameters $\boldsymbol{\theta}_{\mathrm{new}}$. This step will result in an increase in left-hand-side of (1) (if it is not already in a local maximum).

(e) Substitute $q(\mathbf{Z}) = p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}_{\mathrm{curr}})$ in (1), and observe that

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q[\text{complete-data log likelihood}] - H(q).$$

In other words, in the M-step we are maximizing the expectation of the complete-data log likelihood[1], since the entropy term is independent of $\boldsymbol{\theta}$. Compare this result with the EM for Gaussian mixture models.

(f) Show that the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$, where $q(\mathbf{Z}) = q^\star(\mathbf{Z}) = p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}_{\mathrm{curr}})$, has the same gradient w.r.t. $\boldsymbol{\theta}$ as the log likelihood function $p(\mathbf{X} \,|\, \boldsymbol{\theta})$ at the point $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{curr}}$. This shows that the lower bound becomes tangent to the log likelihood function at the end of E-step.

(g) Have you found an argument to prove the convergence of EM algorithm?

[1] $p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$

Problem 4 (On Statistical Distances*):

In some situations in statistics and machine learning, the objective that we are going to optimize is some distribution (*e.g.* in the E-step of EM algorithm). This motivates us to understand a bit more about the *space* of probability distributions.

For simplicity, let $\mathcal{P}$ be the set of all probability distributions over the set $\{1, \ldots, n\}$, *i.e.*

$$\mathcal{P} = \{(p_1, \ldots, p_n) \,|\, \textstyle\sum p_i = 1, p_i \geq 0\}.$$

Usually $\mathcal{P}$ is called the probability simplex, or simply the n-simplex. One can equip $\mathcal{P}$ with a metric, inducing a geometry on $\mathcal{P}$. Recall that a metric is a function $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$ satisfying the following criteria:

- (Non-negativity) $d(p, q) \geq 0$ for all $p, q \in \mathcal{P}$ and equality holds iff $p = q$,
- (Symmetry) $d(p, q) = d(q, p)$,
- (Triangle inequality) $d(p, q) + d(q, r) \geq d(p, r)$.

A metric on the probability simplex is also called a **statistical distance**. Here we mention a few distances and some of their properties:

(a) **Total Variation Distance.** For $p, q \in \mathcal{P}$, we define their TV distance as

$$\mathrm{D}_{\mathrm{TV}}(p, q) := \frac{1}{2}\|p - q\|_1 = \frac{1}{2}\sum_{i=1}^{n} |p_i - q_i|.$$

Prove that TV distance is indeed a metric, and equals to the largest possible difference between the probabilities that the two probability distributions p and q can assign to the same event, *i.e.*

$$\mathrm{D}_{\mathrm{TV}}(p, q) = \max_{E \subseteq \{1, \ldots, n\}} |p(E) - q(E)|.$$

(b) **Kullback-Leibler Divergence.** For $p, q \in \mathcal{P}$, we define their KL divergence as

$$\mathrm{D}_{\mathrm{KL}}(p\|q) = -\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i}.$$

(b.1) Prove that KL divergence satisfies the first property of a metric: it is non-negative, and it is zero if and only if the distributions are equal.

(b.2) Give an example that $\mathrm{D}_{\mathrm{KL}}(p\|q) \neq \mathrm{D}_{\mathrm{KL}}(q\|p)$.

(b.3) Give a counter-example for the triangle inequality for KL divergence.

(b.4) Prove the Pinsker's Inequality:

$$\mathrm{D}_{\mathrm{TV}}(p, q) \leq \sqrt{\tfrac{1}{2}\mathrm{D}_{\mathrm{KL}}(p\|q)}.$$

(b.5) Although KL divergence fails to be a metric on $\mathcal{P}$, it satisfies some convergence properties. As an example, prove the following theorem: Let $p^{(1)}, p^{(2)}, \ldots$ be a sequence of probability distributions in $\mathcal{P}$, such that

$$\lim_{n \to \infty} \mathrm{D}_{\mathrm{KL}}(p^{(n)}\|q) = 0,$$

i.e. the sequence is "converging" to q with respect to KL divergence. Prove that this sequence is actually converging to q in Euclidean sense, *i.e.*

$$\lim_{n \to \infty} \|p^{(n)} - q\|_2 = 0.$$

(b.6) Let X and Y be two random variables with distributions p_X and p_Y and joint distribution $p_{X,Y}$. If X and Y were independent, the we had $p_{X,Y} = p_X p_Y$. Otherwise, if one tries to give a "measure of independence" of X and Y, one idea is to consider

$$\mathrm{D}_{\mathrm{KL}}(p_{X,Y} \| p_X p_Y).$$

This value is called the **mutual information** between X and Y, denoted by $I(X, Y)$. Prove that

$$I(X, Y) = H(X) - H(X \mid Y),$$

where $H(X)$ is the entropy[2] of X and $H(X \mid Y)$ is the conditional entropy of X given Y. In Bayesian point-of-view, the mutual information shows how much information does knowledge about Y reveal about X.

[2] Entropy of a random variable X is defined as $H(X) := \mathbb{E}_X[-\log X] = -\sum_x p_X(x) \log p_X(x)$, and is a measure of "uncertainty" of X. For example if X has the uniform distribution, it has the highest entropy. If the base of $\log$ is 2, entropy is measured with the unit "bits", suggesting the idea that one needs $H(X)$ bits to encode the outcome of X with zeros and ones. Convince yourself that this definition makes sense.