CSCI 567, Fall'18
Haipeng Luo

Homework #4 Written Assignments

Due: 11:59 pm, November 4, 2018

Instructions

Submission: Assignment submission will be via `courses.uscden.net`. By the submission date, there will be a folder set up in which you can submit your files. Please be sure to follow all instructions outlined here.

You can submit multiple times, but only *the last submission* counts. As a results, if you finish some problems, you might want to submit them first, and update later when you finish the rest. You are encouraged to do so. This way, if you forget to finish the homework on time or something happens (remember Murphy's Law), you still get credit for whatever you have turned in.

Problem sets must be typewritten or neatly handwritten when submitted. In both cases, your submission must be a single PDF. It is strongly recommended that you typeset with LaTeX. There are many free integrated LaTeX editors that are convenient to use (e.g Overleaf, ShareLaTeX). Choose the one(s) you like the most. This tutorial Getting to Grips with LaTeX is a good start if you do not know how to use LaTeX yet.

Please also follow the rules below:

- The file should be named as `firstname_lastname_USCID.pdf` e.g., `Don_Quijote_de_la_Mancha_8675309045.pdf`).
- Do not have any spaces in your file name when uploading it.
- Please include your name and USC ID in the header of the report as well.

Collaboration: You may discuss with your classmates. However, you need to write your own solutions and submit separately. Also in your report, you need to list with whom you have discussed for each problem. Please consult the syllabus for what is and is not acceptable collaboration. Review the rules on academic conduct in the syllabus: a single instance of plagiarism can adversely affect you significantly more than you could stand to gain.

Notes on notation:

- Unless stated otherwise, scalars are denoted by small letter in normal font, vectors are denoted by small letters in bold font and matrices are denoted by capital letters in bold font.
- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 Optimization over the simplex

In this exercise you will prove two optimization results over the simplex that we used multiple times in the lectures. These results will also help you solve the other problems in this homework.

The $K-1$ dimensional simplex is simply the set of all distributions over K elements, denoted by $\Delta = \{\mathbf{q} \in \mathbb{R}^K \mid q_k \geq 0,\ \forall k \text{ and } \sum_{k=1}^K q_k = 1\}$.

1.1 Let $a_1, \ldots, a_K$ be K positive numbers. Prove that the solution of the following optimization problem

$$\arg\max_{\mathbf{q}\in\Delta} \sum_{k=1}^K a_k \ln q_k$$

is $\mathbf{q}^*$ such that $q_k^* = \frac{a_k}{\sum_{k'} a_{k'}}$ (that is, $q_k^* \propto a_k$). Hint: the Lagrangian of this problem is

$$L(\mathbf{q}, \lambda, \lambda_1, \ldots, \lambda_K) = \sum_{k=1}^K a_k \ln q_k + \lambda \left(\sum_{k=1}^K q_k - 1 \right) + \sum_{k=1}^K \lambda_k q_k$$

for Lagrangian multipliers $\lambda \neq 0$ and $\lambda_1, \ldots, \lambda_k \geq 0$. Now apply KKT conditions to find $\mathbf{q}^*$.

1.2 Let $b_1, \ldots, b_K$ be K real numbers and H be the entropy function. Prove that the solution of the following optimization problem

$$\arg\max_{\mathbf{q}\in\Delta} \mathbf{b}^T\mathbf{q} + H(\mathbf{q}) = \arg\max_{\mathbf{q}\in\Delta} \sum_{k=1}^K (q_k b_k - q_k \ln q_k)$$

is $\mathbf{q}^*$ such that $q_k^* \propto e^{b_k}$. Hint: follow the exact same steps as in the previous problem, that is, write down the Lagrangian and then apply KKT conditions.

1.3 In the lecture we derived EM through a lower bound of the log-likelihood function. Specifically, on Slide 45 of Lec 8, we find the tightest lower bound by solving

$$\arg\max_{\mathbf{q}_n\in\Delta} \mathbb{E}_{z_n\sim\mathbf{q}_n} \left[\ln p(\mathbf{x}_n, z_n\ ; \theta^{(t)}) \right] + H(\mathbf{q}_n).$$

Use the result from Problem 1.2 to find the solution (you already know what it is from the class).

Problem 2 Gaussian Mixture Model

In the lecture we applied EM to learn Gaussian Mixture Models (GMMs) and showed the M-Step without a proof on Slide 51 of Lec 8 . In this problem you will prove this for the simpler 1D case. Specifically consider a 1D GMM that has the following density function for x:

$$p(x) = \sum_{k=1}^{K} \omega_k \mathcal{N}(x \mid \mu_k, \sigma_k) = \sum_{k=1}^{K} \frac{\omega_k}{\sqrt{2\pi}\sigma_k} \exp\left(\frac{-(x-\mu_k)^2}{2\sigma_k^2}\right)$$

where:

- K: the number of Gaussians components
- μ_k and σ_k: mean and standard deviation of the k-th component
- ω_k: mixture weights - they represent how much each component contributes to the final distribution. It satisfies two properties:
$$\forall k, \omega_k > 0 \text{ and } \sum_k \omega_k = 1$$

Prove that the MLE of the expected complete log-likelihood (with γ_{nk} being the posterior of latent variables computed from the previous E-Step)

$$\sum_n \sum_k \gamma_{nk} \ln \omega_k + \sum_n \sum_k \gamma_{nk} \ln \mathcal{N}(x_n \mid \mu_k, \sigma_k)$$

is the following

$$\omega_k = \frac{\sum_n \gamma_{nk}}{N}, \qquad \mu_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} x_n, \qquad \sigma_k^2 = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} (x_n - \mu_k)^2.$$

Hint: you can make use of the result from Problem 1.1.

Problem 3 MLE and EM

3.1 Let $X \in \mathbb{R}$ be a random variable. We assume that it is uniformly-distributed on some unknown interval $(0, \theta]$, where $\theta > 0$. In particular,

$$P(X = x\,; \theta) = \begin{cases} \frac{1}{\theta} & \text{, if } x \in (0, \theta], \\ 0 & \text{, otherwise,} \end{cases} \tag{1}$$

$$= \frac{1}{\theta} \mathbf{1}[0 < x \leq \theta], \tag{2}$$

where $\mathbf{1}$ is an indicator function that outputs 1 when the condition is true, and 0 otherwise.

Suppose $x_1, x_2, \ldots, x_N$ are drawn i.i.d. from this distribution. Write down the likelihood of the observations and then find the maximum likelihood estimator (MLE).

3.2 Now suppose X is distributed according to a **mixture** of two uniform distributions: one on some unknown interval $(0, \theta_1]$ and the other on $(0, \theta_2]$, for some $\theta_1, \theta_2 > 0$. In particular,

$$\begin{aligned} P(X = x) &= P(X = x, z = 1) + P(X = x, z = 2) \\ &= P(z = 1)P(X = x \mid z = 1) + P(z = 2)P(X = x \mid z = 2) \\ &= \omega_1 U(X = x\,; \theta_1) + \omega_2 U(X = x\,; \theta_2) \end{aligned}$$

where U is the uniform distribution defined as in Eq. (1) or Eq. (2), and ω_1, ω_2 are mixture weights such that

$$\omega_1 \geq 0, \omega_2 \geq 0, \text{ and } \omega_1 + \omega_2 = 1.$$

Suppose $x_1, x_2, \ldots, x_N$ are drawn i.i.d. from this mixture of uniform distributions. MLE does not admit a closed-form for this problem, and we will use the EM algorithm to approximately find the MLE.

- First, the E-Step fixes a set of parameters $\theta_1, \theta_2, \omega_1, \omega_2$ and computes for each n the posterior distribution $\gamma_{nk} = P(z_n = k \mid x_n\,; \theta_1, \theta_2, \omega_1, \omega_2)$ of the latent variable z_n, where $k \in \{1, 2\}$ indicates which mixture component x_n belongs to. Write down the explicit form of this posterior distribution. Then write down the expected complete log-likelihood using γ_{nk} (as a function of the four parameters $\theta_1, \theta_2, \omega_1, \omega_2$).
- Next, derive the M-Step by maximizing the expected complete log-likelihood you derived from the last problem over the four parameters. Hint: you will need to use the fact $0 \ln 0 = 0$.
- Finally, based on the previous two problems, describe how EM behaves for this problem. Specifically, describe 1) how you would initialize θ_1 and θ_2 (think about what constraints you need to impose); 2) how EM works at each round; 3) how many rounds it takes for θ_1 and θ_2 to converge; 4) what the four parameters converge to.

Problem 4 Naive Bayes

Recall the naive Bayes model we have seen in class. Given a random variable $X \in R^D$ and a dependent class variable $Y \in [C]$, the joint distribution of features X and class Y is defined as

$$P(X = \mathbf{x}, Y = c) = P(Y = c)P(X = \mathbf{x}|Y = c) = P(Y = c)\prod_{d=1}^{D} P(X_d = x_d|Y = c)$$

In this problem, we consider a naive Bayes model where each feature x_d of each class c is modeled as a (different) Gaussian. That is,

$$P(X_d = x_d \mid Y = c\,; \mu_{cd}, \sigma_{cd}) = \frac{1}{\sqrt{2\pi}\sigma_{cd}} \exp\left(-\frac{(x_d - \mu_{cd})^2}{2\sigma_{cd}^2}\right)$$

where μ_{cd} and σ_{cd} are the mean and the standard deviation, respectively. Moreover, we model Y as a multinomial distribution with parameter $\boldsymbol{\theta}$ (a distribution over C elements). That is,

$$P(Y = c\,; \boldsymbol{\theta}) = \theta_c \quad \forall c \in [C].$$

4.1 What are the parameters to be learned in this model?

4.2 Given the dataset $\{(\mathbf{x}_n \in R^D, y_n \in [C])\}_{n=1}^{N}$, assumed to be drawn i.i.d. from this model, write down explicitly the expression for the joint log-likelihood.

4.3 Based on the joint log-likelihood derived from the previous problem, find the MLE for this model.