CSCI 567, Spring'18
Michael Shindler

Problem Set #4

Due: 11:59 pm, April 3, 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 directions outlined here.

You can submit multiple files, but only the last one submitted counts. That means if you finish some problems and want to submit something now, and then a later file when you finish, that's fine. If I were taking the class, that's what I'd do: that way, if I forget to finish the homework or something happens (remember Murphy's Law), I still get credit for what I finished and turned in. Remember, there are no grace days on problem sets, just on programming assignments!

Problem sets must be typewritten or neatly handwritten when submitted; if the grader cannot read your handwriting on a problem, they may elect to grade it as a zero. If it is handwritten, your submission must still be submitted as a PDF.

It is strongly recommended that you typeset with LaTeX and use that to generate a PDF file.

- The file should be named as `firstname_lastname_USCID.pdf` (e.g., Jenny_Tutone_8675309.pdf).
- Do not have any spaces in your file name when uploading it.
- Please include your name and USCID in the header of the report as well.

There are many free integrated LaTeX editors that are convenient to use. Choose the one(s) you like the most. This http://www.andy-roberts.net/writing/latex seems like a good tutorial.

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.
- The bias term is subsumed in the input vector, so the input vector is actually $x = [x', 1]^T$, unless mentioned otherwise.
- $\|.\|$ means L2-norm unless specified otherwise i.e. $\|.\| = \|.\|_2$

Problem 1 EM (20 points)

1.1 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 $P(x_1, \ldots, x_N)$. What is the maximum likelihood (ML) estimate of θ? Give a sentence or two explaining why your equation corresponds to the maximum likelihood estimate.

1.2 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 $(0, \theta_2]$, where $0 < \theta_1 \leq \theta_2$. In particular,

$$P(X = x) = \omega_1 U(X = x|\theta_1) + \omega_2 U(X = x|\theta_2), \tag{3}$$

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

$$\omega_1 \geq 0, \omega_2 \geq 0, \text{and } \omega_1 + \omega_2 = 1. \tag{4}$$

Suppose $x_1, x_2, \ldots, x_\mathcal{N}$ are drawn i.i.d. from this mixture of uniform distributions. We will use an EM algorithm to derive the ML estimates of the parameters in this model. Answer the following three questions.

- First, what is the form of $P(k|x_n, \theta_1, \theta_2, \omega_1, \omega_2)$, where $k \in \{1, 2\}$ indicates the corresponding mixture component? You answer should be explicit. You may include the indicator function as in Eq. (2) and the U function as in Eq. (3).
- Second, what is the form of the expected complete-data log-likelihood of the observations $\{x_1, \ldots, x_\mathcal{N}\}$, given that the parameters from the last EM iteration are $\{\theta_1^{\text{OLD}}, \theta_2^{\text{OLD}}, \omega_1^{\text{OLD}} > 0, \omega_2^{\text{OLD}} > 0\}$, where $\theta_2^{\text{OLD}} \geq \max\{x_1, \ldots, x_\mathcal{N}\} \geq \theta_1^{\text{OLD}} \geq \min\{x_1, \ldots, x_\mathcal{N}\}$? You answer should be explicit. You may include the indicator function as in Eq. (2) and the U function as in Eq. (3).
- Third, what are the forms of the M-step updates for θ_1 and θ_2?
  You may use $P_{\text{OLD}}(k|x_n) = P(k|x_n, \theta_1^{\text{OLD}}, \theta_2^{\text{OLD}}, \omega_1^{\text{OLD}}, \omega_2^{\text{OLD}})$, where $k \in \{1, 2\}$, to simplify your derivation/answer if needed. You can view $\log 0 = -\infty$ and $0 \times \log 0 = 0$ in this question.

Problem 2 The connection between GMM and K-means (20 points)

Consider a Gaussian mixture model (GMM) in which all components have (diagonal) covariance $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ and the K-means algorithm introduced in lectures.

2.1 In the case where both the GMM and the K-means algorithm have K components and the parameters π_k are pre-defined to be nonzero $\forall k \in [K]$, show that in the limit $\sigma \to 0$, **maximizing** the following expected complete-data log likelihood w.r.t. $\{\boldsymbol{\mu}_k\}_{k=1}^K$ for the GMM model

$$\sum_n^N \sum_k^K \gamma(z_{nk}) \log p(\mathbf{x}_n, z_n = k) = \sum_n^N \sum_k^K \gamma(z_{nk}) [\log \pi_k + \log \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \sigma^2 \mathbf{I})],$$

$$\text{where} \quad \gamma(z_{nk}) = \frac{\pi_k \exp(-||\mathbf{x}_n - \boldsymbol{\mu}_k||^2 / 2\sigma^2)}{\sum_j \pi_j \exp(-||\mathbf{x}_n - \boldsymbol{\mu}_j||^2 / 2\sigma^2)}$$

is equivalent (up to a scaling or constant factor) to **minimizing** the distortion measure J w.r.t. $\{\boldsymbol{\mu}_k\}_{k=1}^K$ for the K-means algorithm given by

$$J = \sum_k^K \sum_n^N r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||_2^2,$$

$$\text{where} \quad r_{nk} = \begin{cases} 1, \text{ if } k = \arg\min_{k'} ||\mathbf{x}_n - \boldsymbol{\mu}_{k'}||_2^2, \\ 0, \text{ otherwise.} \end{cases}$$

Hint: Start by showing that $\gamma(z_{nk}) \to r_{nk}$ as $\sigma \to 0$. Note that for this question the only set of parameters to learn for the GMM are $\{\boldsymbol{\mu}_k\}_{k=1}^K$. Any term independent of $\{\boldsymbol{\mu}_k\}_{k=1}^K$ can be treated as a constant.

Problem 3 Gaussian Mixture Model Parameters (20 points)

In the lecture you learned about Gaussian Mixture Model (GMM) and that it has the following density function for x:

$$p(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \sigma_k) \tag{5}$$

where:

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

If we have a z_n for every x_n to denote the distribution the specific x_n comes from, and a binary variable $\gamma_{nk} \in \{0, 1\}$ to indicate whether $z_n = k$, the complete log-likelihood is for data set $X = x_1, x_2, ..., x_N$ is given by:

$$L = \sum_n lnp(x_n, z_n) = \sum_k \sum_n \gamma_{nk} log w_k + \sum_k \sum_n \gamma_{nk} log \mathcal{N}(x_n|\mu_k, \sigma_k) \tag{6}$$

Show that the maximum likelihood estimation of the parameters is:

$$w_k = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}} \tag{7}$$

$$\mu_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} x_n \tag{8}$$

$$\sigma_k = \frac{1}{\sum_n \gamma_{nk}} \sum_n \gamma_{nk} (x_n - \mu_k)^2 \tag{9}$$

What to submit: Your derivations of equations 7, 8, 9 parameters that maximize the likelihood.
Hint: When deriving w_k, keep in mind that there are certain properties it needs to satisfy

Problem 4 Mixture density models (20 points)

Consider a density model given by a mixture distribution

$$P(x) = \sum_{k=1}^{K} \pi_k P(\boldsymbol{x}|k), \text{ where } \pi_k \geq 0 \; \forall k \in [K] \text{ and } \sum_{k=1}^{K} \pi_k = 1, \tag{10}$$

and suppose that we partition the vector $\boldsymbol{x}$ into two parts so that $\boldsymbol{x} = [\boldsymbol{x}_a^T, \boldsymbol{x}_b^T]^T$. Show that the conditional density $P(\boldsymbol{x}_b|\boldsymbol{x}_a)$ is itself a mixture distribution. That is,

$$P(\boldsymbol{x}_b|\boldsymbol{x}_a) = \sum_{k=1}^{K} \lambda_k P(\boldsymbol{x}_b|\boldsymbol{x}_a, k), \text{ where } \lambda_k \geq 0 \; \forall k \in [K] \text{ and } \sum_{k=1}^{K} \lambda_k = 1 \tag{11}$$

Find an expression for the mixing coefficients λ_k of the component densities in terms of π_k and $P(\boldsymbol{x}_a|k)$. Do not forget to verify if your answer obeys the constraint on λ_k mentioned above.

*Hint:*a) λ_k is a function of $\boldsymbol{x}_a$ instead of a constant. b) You may consider Bayes rule for derivation.

*What to submit:*No more than ten lines of derivation that leads to an expression for the mixing coefficients λ_k.