CSCI 567, Summer 2022
Prof. Victor Adamchik

Theory Assignment 4

Due: 11:59 pm, June 25, 2022

Instructions

Submission: Assignment submission will be via courses.uscden.net. By the submission date, there will be a folder named 'Theory Assignment 1' set up in which you can submit your files. Please be sure to follow all directions outlined here.

You can submit multiple times, but only *the last submission* counts. That means if you finish some problems and want to submit something first and update later when you finish, that's fine. In fact you are encouraged to do this: that 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 USCID in the header of your 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 Kmeans Clustering **(32 points)**

In Lecture 15 (slide 5), we describes the loss function of k-means as follows -

$$F(\{\gamma_{nk}\}, \{\mu_l\}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} ||x_n - \mu_k||_2^2$$

Refer to slide 10 of lecture 15 for the update rule of k-means.

1.1 Show that the loss function decreases in each iteration of K-means. **(8 points)**

1.2 Argue that k-means converges in a finite number of iterations. **(8 points)**

1.3 We update our loss function to use l_1 norm instead of l_2 norm such that

$$F(\{\gamma_{nk}\}, \{\mu_k\}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} |x_n - \mu_k|$$

How does the performance of this compare to the original loss function for the case of data consisting of outliers? **(4 points)**

1.4 Show that cluster centers are medians of the points in the cluster. **(12 points)**

Problem 2 Gaussian Mixture Model (32 points)

Let $X_1, ..., X_n \in \mathbb{R}^d$ be independent, identically distributed points sampled from a mixture of two normal (Gaussian) distributions. They are drawn independently from the probability distribution function (PDF)

$p(x) = \theta N_1(x) + (1-\theta)N_2(x)$, where $N_1(x) = \frac{1}{(\sqrt{2\pi})^d} e^{-\|x-\mu_1\|^2/2}$, and $N_2(x) = \frac{1}{(\sqrt{2\pi})^d} e^{-\|x-\mu_2\|^2/2}$

are the PDFs for the isotropic multivariate normal distributions $\mathcal{N}(\mu_1, 1)$ and $\mathcal{N}(\mu_2, 1)$, respectively. The parameter $\theta \in (0, 1)$ is called the mixture proportion. In essence, we flip a biased coin to decide whether to draw a point from the first Gaussian (with probability θ) or the second (with probability $1 - \theta$).

Each data point is generated as follows. First draw a random Z_i, which has value 1 with probability θ, and has value 2 with probability $1 - \theta$. Then, draw $X_i \sim \mathcal{N}(\mu_{Z_i}, 1)$. Our learning algorithm gets X_i as an input, but does not know Z_i .

Our goal is to find the maximum likelihood estimates of the three unknown distribution parameters $\theta \in (0, 1), \mu_1 \in R^d$, and $\mu_2 \in R^d$ from the sample points $X_1, ..., X_n$. Unlike MLE for one Gaussian, it is not possible to give explicit analytic formulas for these estimates. Instead, we develop a variant of k-means clustering which (often) converges to the correct maximum likelihood estimates of $\theta, \mu_1, and \mu_2$. This variant doesn't assign each point entirely to one cluster; rather, each point is assigned an estimated posterior probability of coming from normal distribution 1.

2.1 Let $\tau_i = P(Z_i = 1|X_i)$. That is, τ_i is the posterior probability that point X_i has $Z_i = 1$. Use Bayes' Theorem to express τ_i in terms of $X_i, \theta, \mu_1, \mu_2$, and the Gaussian PDFs $\mathcal{N}_1(x)$ and $\mathcal{N}_2(x)$. To help you with question 2.3, also write down a similar formula for $1 - \tau_i$, which is the posterior probability that $Z_i = 2$.

(6 points)

2.2 Write down the log-likelihood function, $l(\theta, \mu_1, \mu_2; X_1, ..., X_n) = \ln p(X_1, ..., X_n)$, as a summation. Note: it doesn't simplify much.

(5 points)

2.3 Express $\frac{\partial l}{\partial \theta}$ in terms of θ and $\tau_i, i \in \{1, ..., n\}$ and simplify as much as possible. There should be no normal PDFs explicitly in your solution, though the τ_i's may implicitly use them. Hint: Recall that $(\ln f(x)) = \frac{f(x)}{f(x)})$.

(5 points)

2.4 Express $\nabla_{\mu_1} l$ in terms of μ_1 and $\tau_i, X_i, i \in \{1, ..., n\}$. Do the same for $\nabla_{\mu_2} l$ (but in terms of μ_2 rather than μ_1). Again, there should be no normal PDFs explicitly in your solution, though the τ_i's may implicitly use them. Hint: It will help (and get you part marks) to first write $\nabla_{\mu_1} N_1(x)$ as a function of $N_1(x)$, x, and μ_1.

(7 points)

2.5 We conclude: if we know μ_1, μ_2, and θ, we can compute the posteriors τ_i. On the other hand, if we know the τ_i's, we can estimate μ_1, μ_2, and θ by using the derivatives in 2.3 and 2.4 to find the maximum likelihood estimates. This leads to the following k-means-like algorithm.

- Initialize $\tau_1, \tau_2, ..., \tau_n$ to arbitrary values in the range $[0, 1]$.
- Repeat the following two steps.
  1. Update the Gaussian cluster parameters: for fixed values of $\tau_1, \tau_2, ..., \tau_n$, update μ_1, μ_2, and θ.
  2. Update the posterior probabilities: for fixed values of μ_1, μ_2, and θ, update $\tau_1, \tau_2, ..., \tau_n$.

In part 2.1, you wrote the update rule for step 2. Using your results from parts 2.3 and 2.4, write down the explicit update formulas for step 1.

(9 points)