CSCI 567, Spring 2019
Prof. Victor Adamchik

Theory Assignment 4

Due: 11:59 pm, April 19th, 2019

Instructions

Submission: Assignment submission will be via `courses.uscden.net`. By the submission date, there will be a folder named 'Theory Assignment 4' 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 Optimization over the simplex (12 points)

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. Solve the following optimization problem **(6 points)**

$$\begin{aligned} \arg\max_{\mathbf{q}} & \sum_{k=1}^K a_k \ln q_k \\ \text{s.t. } & q_k \geq 0 \\ & \sum_{k=1}^K q_k = 1 \end{aligned}$$

Ans: The stationary condition states that for each k, **(2 points)**

$$\frac{a_k}{q_k^*} + \lambda + \lambda_k = 0 \quad (\text{with } \lambda_k \geq 0)$$

$\frac{a_k}{q_k^*} + \lambda - \lambda_k = 0$ is also acceptable if we constrain $\lambda_k \leq 0$.
and thus

$$q_k^* = -\frac{a_k}{\lambda + \lambda_k} \neq 0.$$

The complementary slackness condition implies that $\lambda_k q_k^* = 0$ and thus $\lambda_k = 0$. **(2 points)**
Finally, feasibility implies

$$\sum_{k=1}^K q_k^* = -\sum_{k=1}^K \frac{a_k}{\lambda} = 1.$$

Solving for λ and plugging it back gives the solution $q_k^* = \frac{a_k}{\sum_{k'} a_{k'}}$. **(2 points)**

1.2 Let $b_1, \ldots, b_K$ be K real numbers. Solve the following optimization problem **(6 points)**

$$\begin{aligned} \arg\max_{\mathbf{q} \in \Delta} & \sum_{k=1}^K (q_k b_k - q_k \ln q_k) \\ \text{s.t. } & q_k \geq 0 \\ & \sum_{k=1}^K q_k = 1 \end{aligned}$$

Ans: The Lagrangian of this problem is

$$L(\mathbf{q}, \lambda, \lambda_1, \ldots, \lambda_K) = \sum_{k=1}^K (q_k b_k - q_k \ln q_k) + \lambda \left(\sum_{k=1}^K q_k - 1 \right) + \sum_{k=1}^K \lambda_k q_k, \quad (\text{with } \lambda_k \geq 0)$$

$\sum_{k=1}^K (q_k b_k - q_k \ln q_k) + \lambda \left(\sum_{k=1}^K q_k - 1 \right) - \sum_{k=1}^K \lambda_k q_k$ is also acceptable if we constrain $\lambda_k \leq 0$.
The stationary condition implies that for each k **(2 points)**

$$b_k - 1 - \ln q_k + \lambda + \lambda_k = 0,$$

and thus

$$q_k = \exp\left(b_k - 1 + \lambda + \lambda_k\right) \propto e^{b_k + \lambda_k} \neq 0.$$

Complementary slackness then implies $\lambda_k = 0$ and thus $q_k \propto e^{b_k}$. **(2 points)**

Therefore $q_k = \frac{e^{b_k}}{\sum_{k'} e^{b_k}}$. **(2 points)**

Problem 2 Gaussian Mixture Model and EM (10 points)

2.1 In the lecture we applied EM to learn Gaussian Mixture Models (GMMs) and showed the M-Step without a proof. In this problem you will prove this for a simpler case. Consider a GMM that has the following density function for $\mathbf{x}$:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k) = \sum_{k=1}^{K} \frac{\omega_k}{(\sqrt{2\pi})^D |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)$$

where K is the number of Gaussian components, D is the length of $\mathbf{x}_n$. μ_k is the mean of k-th Gaussian component, Σ_k is the covariance matrix of the k-th Gaussian component and ω_k is the mixture weight for the k-th Gaussian component. For simplicity, we assume prior knowledge of $\Sigma_k = \sigma_k^2 I$.

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

$$\begin{aligned} \arg\max_{\omega_k, \boldsymbol{\mu}_k, \Sigma_k} & \sum_n \sum_k \gamma_{nk} \ln \omega_k + \sum_n \sum_k \gamma_{nk} \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) \\ \text{s.t. } & \omega_k \geq 0 \\ & \sum_{k=1}^{K} \omega_k = 1 \end{aligned}$$

Hint: you can make use of the result from Problem 1.1. **(8 points)**

Ans:

To find $\omega_1, \ldots, \omega_K$, we simply solve

$$\begin{aligned} \arg\max_{\boldsymbol{\omega}} & \sum_n \sum_k \gamma_{nk} \ln \omega_k. \\ \text{s.t. } & \omega_k \geq 0 \\ & \sum_{k=1}^{K} \omega_k = 1 \end{aligned} \qquad \textbf{(1 points)}$$

According to Problem 1.1 with $a_k = \sum_n \gamma_{nk}$, the solution is

$$\omega_k^* = \frac{\sum_n \gamma_{nk}}{\sum_k \sum_n \gamma_{nk}} = \frac{\sum_n \gamma_{nk}}{\sum_n 1} = \frac{\sum_n \gamma_{nk}}{N}. \qquad \textbf{(2 points)}$$

To find μ_k and σ_k, we solve for each k

$$\begin{aligned} \arg\max_{\boldsymbol{\mu}_k, \Sigma_k} \sum_n \gamma_{nk} \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) &= \arg\max_{\boldsymbol{\mu}_k, \Sigma_k} \sum_n \gamma_{nk} \ln \left[\frac{1}{(\sqrt{2\pi}\sigma_k)^D} \exp\left(-\frac{1}{2\sigma_k^2} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2\right)\right] \\ &= \arg\max_{\boldsymbol{\mu}_k, \Sigma_k} \sum_n \gamma_{nk} \left(-D \ln \sigma_k - \frac{\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2}{2\sigma_k^2}\right), \end{aligned}$$

where D is the length of $\mathbf{x}_n$. **(1 points)**

First we set the derivative w.r.t. $\boldsymbol{\mu}_k$ to 0:

$$\frac{1}{\sigma_k^2} \sum_n \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0,$$

which gives

$$\boldsymbol{\mu}_k^* = \frac{\sum_n \gamma_{nk} \mathbf{x}_n}{\sum_n \gamma_{nk}} \qquad \textbf{(2 points)}$$

Next we set the derivative w.r.t. σ_k to 0:

$$\sum_n \gamma_{nk} \left(-\frac{D}{\sigma_k} + \frac{\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^3} \right) = 0.$$

Solving for σ_k gives

$$(\sigma_k^*)^2 = \frac{\sum_n \gamma_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k^*\|^2}{D \sum_n \gamma_{nk}}. \qquad \textbf{(2 points)}$$

2.2 In the lecture we derived EM through a lower bound of the log-likelihood function. Specifically, we find the tightest lower bound by solving

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

Use the result from Problem 1.2 to find the solution (you already know what it is from the class). **(2 points)**

Ans: This is exactly in the same form of the problem in 1.2 with $b_k = \ln p(\mathbf{x}_n, z_n = k \,; \theta^{(t)})$. So the solution is

$$q_{nk}^* \propto p(\mathbf{x}_n, z_n = k \,; \theta^{(t)}),$$

or in other words,

$$q_{nk}^* = \frac{p(\mathbf{x}_n, z_n = k \,; \theta^{(t)})}{\sum_{k=1}^K p(\mathbf{x}_n, z_n = k \,; \theta^{(t)})} = \frac{p(\mathbf{x}_n, z_n = k \,; \theta^{(t)})}{p(\mathbf{x}_n \,; \theta^{(t)})} = p(z_n = k \mid \mathbf{x}_n \,; \theta^{(t)}). \qquad \textbf{(2 points)}$$

Problem 3 Hidden Markov Model (13 points)

Recall a hidden Markov model is parameterized by

- initial state distribution $P(Z_1 = s) = \pi_s$
- transition distribution $P(Z_{t+1} = s' \mid Z_t = s) = a_{s,s'}$
- emission distribution $P(X_t = o \mid Z_t = s) = b_{s,o}$

3.1 Suppose we observe a sequence of outcomes $x_1, \ldots, x_T$ and would like to predict the next state Z_{T+1}, that is, we want to figure out for each possible state s,

$$P(Z_{T+1} = s \mid X_{1:T} = x_{1:T}).$$

Write down how one can compute this probability using the the forward message: **(4 points)**

$$\alpha_s(T) = P(Z_T = s, X_{1:T} = x_{1:T}).$$

Ans:

$$\begin{aligned} P(Z_{T+1} = s \mid X_{1:T} = x_{1:T}) &= \frac{P(Z_{T+1} = s, X_{1:T} = x_{1:T})}{P(X_{1:T} = x_{1:T})} \\ &= \frac{\sum_{s'} P(Z_{T+1} = s, Z_T = s', X_{1:T} = x_{1:T})}{\sum_{s''} P(Z_T = s'', X_{1:T} = x_{1:T})} \quad \text{(marginalizing } \textbf{(1 points)}\text{)} \\ &= \frac{\sum_{s'} P(Z_T = s', X_{1:T} = x_{1:T}) P(Z_{T+1} = s \mid Z_T = s', X_{1:T} = x_{1:T})}{\sum_{s''} P(Z_T = s'', X_{1:T} = x_{1:T})} \\ &= \frac{\sum_{s'} P(Z_T = s', X_{1:T} = x_{1:T}) P(Z_{T+1} = s \mid Z_T = s')}{\sum_{s''} P(Z_T = s'', X_{1:T} = x_{1:T})} \\ &\qquad \text{(Markov property } \textbf{(1 points)}\text{)} \\ &= \frac{\sum_{s'} \alpha_{s'}(T) a_{s',s}}{\sum_{s''} \alpha_{s''}(T)} \quad \textbf{(2 points)} \end{aligned}$$

3.2 We are given a HMM with the following probabilities:

Transition probabilities:

Current	Next: A	B	End
Start	0.7	0.3	0
A	0.2	0.7	0.1
B	0.7	0.2	0.1

Emission probabilities:

State:	Word: ''	*'fight'*	*'on'*
Start	1	0	0
A	0	0.4	0.6
B	0	0.7	0.3

We assume that the process stops at state 'End'.

(a) Suppose the process starts from state 'Start' at $t = 0$, and we observe $x_{1:2} =$ *fight on* , write down the forward messages $\alpha_s(2)$ and determine the most likely state at $t = 3$ by comparing the probability for each state. **(4 points)**

Ans:

$t = 1$:

$$\alpha_A(1) = 0.7 * 0.4 = 0.28$$
$$\alpha_B(1) = 0.3 * 0.7 = 0.21$$

$t = 2$:

$$\alpha_A(2) = 0.6 * [0.7 * 0.21 + 0.2 * 0.28] = 0.1218$$
$$\alpha_B(2) = 0.3 * [0.2 * 0.21 + 0.7 * 0.28] = 0.0714 \qquad \textbf{(2 points)}$$

$$\begin{aligned} P(Z_3 = A | X_{1:2} = x_{1:2}) &= \frac{\sum_{s'} \alpha_{s'}(2) a_{s',A}}{\sum_{s''} \alpha_{s''}(2)} \\ &= \frac{\alpha_A(2) a_{A,A} + \alpha_B(2) a_{B,A}}{\alpha_A(2) + \alpha_B(2)} \\ &\approx 0.3848 \\ P(Z_3 = B | X_{1:2} = x_{1:2}) &\approx 0.5152 \qquad \textbf{(1 points)} \\ P(Z_3 = End | X_{1:2} = x_{1:2}) &= 0.1 \end{aligned}$$

Therefore, the most likely state at $t = 3$ given the observed sequence is B. **(1 points)**

(b) Suppose the process starts from state 'Start' at $t = 0$, and we observe the whole output sequence as *fight on on* , what is the most likely sequence of states that produce this? **(5 points)**

Ans:

$t = 1$:

$$\delta_A(1) = 0.7 * 0.4 = 0.28$$
$$\delta_B(1) = 0.3 * 0.7 = 0.21 \qquad \textbf{(1 points)}$$

$t = 2$:

$$\begin{aligned} \delta_A(2) &= 0.6 * \max\{0.7 * 0.21, 0.2 * 0.28\} \\ &= 0.0882 \\ \Delta_A(2) &= B \\ \delta_B(2) &= 0.3 * \max\{0.2 * 0.21, 0.7 * 0.28\} \\ &= 0.0588 \\ \Delta_B(2) &= A \qquad \textbf{(1 points)} \end{aligned}$$

$t = 3$:

$$\begin{aligned} \delta_A(3) &= 0.6 * \max\{0.7 * 0.0588, 0.2 * 0.0882\} \\ &= 0.024696 \\ \Delta_A(3) &= B \\ \delta_B(3) &= 0.3 * \max\{0.2 * 0.0588, 0.7 * 0.0882\} \\ &= 0.018522 \\ \Delta_B(3) &= A \qquad \textbf{(1 points)} \end{aligned}$$

Via backtracking, $z_3^* = A$, $z_2^* = B$, $z_1^* = A$. **(2 points)**

Problem 4 Principal Component Analysis (14 points)

4.1 In the class we showed that PCA is finding the directions with most variance. In this problem, you will show that by finding the directions with most variance, the linear dependency in features is emoved.

Suppose we have a centered data matrix $X \in \mathbb{R}^{N \times D}$ ($N > D$) with N observations and D features, i.e. rows of X correspond to observations and columns correspond to features. If some features are linearly dependent, i.e. $R \triangleq \text{rank}(X) < D$, prove that we only need R principal components to cover the entire spectrum. **(4 points)**

Ans:

For PCA, we apply eigenvalue decomposition to the covariance matrix $X^T X$ and choose the top p eigenvectors.

$$\text{rank}(X^T X) = \text{rank}(X) = R \qquad \textbf{(2 points)}$$

Since the rank of the covariance matrix is R, there are R non-zero eigenvalues.
Let $\lambda_1 \geq \cdots \geq \lambda_D$ be sorted eigenvalues, then

$$\frac{\sum_{d=1}^{R} \lambda_d}{\sum_{d=1}^{D} \lambda_d} = \frac{\sum_{d=1}^{R} \lambda_d}{\sum_{d=1}^{R} \lambda_d} = 100\% \qquad \textbf{(2 points)}$$

Alternative solution:
To cover the entire spectrum:

$$\begin{aligned} X^T X &= \sum_{d=1}^{D} \lambda_d \mathbf{v}_d \mathbf{v}_d^T \\ &= \sum_{d=1}^{R} \lambda_d \mathbf{v}_d \mathbf{v}_d^T \qquad \textbf{(2 points)} \end{aligned}$$

4.2 In this problem, you will show that PCA is in fact also minimizing reconstruction error in some sense.

Specifically, suppose we have a dataset $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ with zero mean. We define the reconstruction error in terms of L2 distance as

$$\epsilon = \sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2. \tag{1}$$

To reconstruct the dataset, we first present the data to a set of orthonormal basis $\mathbf{v}_1, \ldots, \mathbf{v}_D \in \mathbb{R}^D$ as

$$\begin{aligned} \mathbf{x}_n &= \sum_{d=1}^{D} \alpha_{n,d} \mathbf{v}_d, \\ \text{with } \alpha_{n,d} &= \mathbf{x}_n^T \mathbf{v}_d. \end{aligned}$$

The approximate reconstruction using the first M ($M < D$) elements of the basis is

$$\hat{\mathbf{x}}_n = \sum_{d=1}^{M} \alpha_{n,d} \mathbf{v}_d.$$

(a) Prove that the reconstruction error defined in Eq. 1 can be expressed as

$$\epsilon = \sum_{d=M+1}^{D} \mathbf{v}_d X^T X \mathbf{v}_d,$$

where $X \in \mathbb{N} \times \mathbb{D}$ is the data matrix. **(4 points)**

Ans:

$$\begin{aligned}
\epsilon &= \sum_{n=1}^{N} \|\mathbf{x}_n - \sum_{d=1}^{M} \alpha_{n,d}\mathbf{v}_d\|^2 \\
&= \sum_{n=1}^{N} \| \sum_{d=M+1}^{D} \alpha_{n,d}\mathbf{v}_d\|^2 \\
&= \sum_{n=1}^{N} \left[\sum_{d=M+1}^{D} \alpha_{n,d}\mathbf{v}_d\right]^T \left[\sum_{d'=M+1}^{D} \alpha_{n,d'}\mathbf{v}'_d\right] \\
&= \sum_{n=1}^{N} \sum_{d=M+1}^{D} \sum_{d'=M+1}^{D} \alpha_{n,d}\alpha_{n,d'}\mathbf{v}_d^T\mathbf{v}'_d \qquad \textbf{(2 points)} \\
&= \sum_{n=1}^{N} \sum_{d=M+1}^{D} \alpha_{n,d}^2 \\
&= \sum_{n=1}^{N} \sum_{d=M+1}^{D} \mathbf{v}_d^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{v}_d \\
&= \sum_{d=M+1}^{D} \mathbf{v}_d^T X^T X \mathbf{v}_d \qquad \textbf{(2 points)}
\end{aligned}$$

(b) Prove that minimizing the reconstruction error

$$\begin{aligned}
\min_{\mathbf{v}_{M+1},\ldots,\mathbf{v}_D} & \sum_{d=M+1}^{D} \mathbf{v}_d^T X^T X \mathbf{v}_d \\
\text{s.t. } & \|\mathbf{v}_d\| = 1 \\
& \mathbf{v}_d^T\mathbf{v}_{d'} = 0 \text{ for } d \neq d'
\end{aligned}$$

is equivalent to maximizing the variance of projections

$$\begin{aligned}
\max_{\mathbf{v}_1,\ldots,\mathbf{v}_M} & \sum_{d=1}^{M} \mathbf{v}_d^T X^T X \mathbf{v}_d \\
\text{s.t. } & \|\mathbf{v}_d\| = 1 \\
& \mathbf{v}_d^T\mathbf{v}_{d'} = 0 \text{ for } d \neq d'.
\end{aligned}$$

Hint: the transpose of an orthogonal matrix is also orthogonal, i.e. $\sum_{d=1}^{D} v_{d,i}v_{d,j} = \mathbb{I}[i = j]$, where $v_{d,i}$ is the i-th element of $\mathbf{v}_d$. **(6 points)**

Ans:

$$\sum_{d=1}^{M} \mathbf{v}_d^T X^T X \mathbf{v}_d + \sum_{d=M+1}^{D} \mathbf{v}_d^T X^T X \mathbf{v}_d = \sum_{d=1}^{D} \mathbf{v}_d^T X^T X \mathbf{v}_d \qquad \textbf{(2 points)}$$

Denote $\Sigma = X^T X$, then

$$\begin{aligned}
\sum_{d=1}^{D} \mathbf{v}_d^T X^T X \mathbf{v}_d &= \sum_{d=1}^{D} \mathbf{v}_d^T \Sigma \mathbf{v}_d \\
&= \sum_{d=1}^{D} \sum_{i=1}^{D} \sum_{j=1}^{D} v_{d,i} v_{d,j} \Sigma_{ij} \\
&= \sum_{i=1}^{D} \sum_{j=1}^{D} \Sigma_{ij} \sum_{d=1}^{D} v_{d,i} v_{d,j} \\
&= \sum_{i=1}^{D} \sum_{j=1}^{D} \Sigma_{ij} \mathbb{I}[i=j] \\
&= \sum_{i=1}^{D} \Sigma_{ii} \\
&= Tr(\Sigma) \qquad \textbf{(3 points)}
\end{aligned}$$

Therefore the sum of the two objective function is a constant, which means that the two optimization problems are equivalent. **(1 points)**