CSCI-567 Spring 2019 Midterm Exam-2 Ans: [Rubric]

Problem	1	2	3	4	5	Total
Points	26	20	26	18	10	100

Please read the following instructions carefully:

- The exam has a total of **16 pages** (including this cover and two blank pages in the end). Each problem have several questions. Once you are permitted to open your exam (and not before), you should check and make sure that you are not missing any pages.
- Duration of the exam is **3 hours**. Questions are not ordered by their difficulty. Budget your time on each question carefully.
- Select **one and only one answer** for all multiple choice questions.
- You must answer each question on the same page of the question. Answers should be **concise** and written down **legibly**. You can use the last two blank pages as scratch paper.
- This is a **closed-book/notes** exam. Consulting any resources is NOT permitted.
- Any kind of cheating will lead to **score 0** for the entire exam and be reported to SJACS.
- Raise your hand to ask a proctor if you have any question regarding the exam.

1 Multiple Choice & True/False Questions (26 points)

Each true/false question is 1 *point* and each multiple choice question is 2 *points*.

Lagrangian duality and SVM (6 points)

1. Support vector machines give a probability distribution over the possible labels given an input example. Ans: False

   (A) True

   (B) False

2. Suppose an SVM will be trained on a dataset with N data points and d features. The SVM's primal formulation requires learning $O(N)$ parameters, and the dual counterpart requires learning $O(d)$ parameters. Ans: False

   (A) True

   (B) False

3. In a soft margin SVM, which of the following tends to occur when hyperparameter $C \to \infty$. Ans: B

   (A) Underfitting

   (B) Overfitting

   (C) High train error

   (D) Low test error

4. Which of the following statements is *not* true about SVM? Ans: C

   (A) For two dimensional data points, the separating hyperplane learned by a linear SVM will be a straight line.

   (B) In theory, a Gaussian kernel SVM can model any complex separating hyperplane.

   (C) The support vectors are expected to remain the same between linear kernels and higher–order polynomial kernels.

   (D) Overfitting in an SVM is a function of number of support vectors.

Gaussian Mixture Model (6 points)

5. Which of the following statements of the Expectation-Maximization (EM) algorithm is true? Ans: D

   (A) Before running the EM algorithm, we need to choose the step size.

   (B) EM always converges to the global optimum of the likelihood.

   (C) In EM, the lower-bound for the log-likelihood function we maximize is always non-concave.

   (D) None of the above.

6. Which of the following statements of Gaussian Mixture Model (GMM) is true? Ans: B

   (A) GMM is a non-parametric method that all the training samples need to be stored.

   (B) GMM is a probabilistic model that can be used to explain how data is generated.

   (C) When learning a GMM, the labels of the samples are available.

   (D) None of the above

7. Both K-means and GMM always converge to some local optimum. Ans: A

   (A) True

   (B) False

8. K-means is the special case of EM for GMM. Ans: A

   (A) True

   (B) False

Boosting (6 points)

9. A decision stump can only lead to linear decision boundary for classification. Ans: A

   (a) True.

   (b) False.

10. The AdaBoost algorithm will eventually reach zero training error regardless of the type of weak classifier it uses, when enough iterations are performed. Ans: B

    (a) True.

    (b) False.

11. Which of the following statement is true? Ans: A or B is fine

    (A) In the Adaboost algorithm, weights of the misclassified examples may not go up.

    (B) The Adaboost algorithm is robust to overfitting.

    (C) The testing error of the classifier learned with Adaboost algorithm (combination of all the weak classifier) monotonically increases as the number of iterations in the boosting algorithm increases.

    (D) None of the above

12. Which of the following statement about the Adaboost algorithm is true? Ans: D

    (A) If a weak learner with $< 50\%$ predictive accuracy presents, the AdaBoost algorithm will not work.

    (B) AdaBoost may output a linear classifier if the base classifiers are linear.

    (C) Boosting algorithm can select the same weak classifier more than once.

    (D) All of the above

Hidden Markov Model (4 points)

13. Which of the following statements of hidden Markov model (HMM) is true? Ans: A

    (A) We can infer the backward message at time t from the backward message at time $t+1$ using the backward algorithm.

    (B) We can learn a HMM using the forward algorithm.

    (C) We can infer the likelihood of two consecutive states at a given time using the Viterbi algorithm.

    (D) None of the above.

14. Both GMM and HMM can be learned by applying the EM algorithm. Ans: A

    (A) True

    (B) False

15. Given a sequence of observations and a learned HMM, we can infer the real corresponding path of hidden states. Ans: B

    (A) True

    (B) False

PCA and dimensionality reduction (4 points)

16. Which of the following statement about kernel PCA is true? Ans: C

    (A) The first step of kernel PCA is to center the original dataset.

    (B) Kernel PCA outputs a compressed dataset that is a linear transformation of the original dataset.

    (C) Kernel PCA requires computing eigenvalues and eigenvectors of the Gram matrix.

    (D) Kernel PCA is a parametric approach.

17. We use PCA to pre-process the data: we keep the top k PCA projections as the features. Smaller k can help reduce overfitting. Ans: A

    (A) True

    (B) False

18. PCA can be used to visualize data, compress data, or de-noise data. Ans: A

    (A) True

    (B) False

2 Lagrangian Duality & SVM (20 points)

Consider a max-margin linear SVM that solves the following optimization problem.

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i[\mathbf{w}^T\mathbf{x}_i + b] \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

where the optimization parameters are $\mathbf{w}$, b, and the training set consists of points $\{(\mathbf{x}, y)\}_{i=1}^n$. $\mathbf{x}$ is a vector of real values, and $y \in \{-1, 1\}$ is the class label. In this section, you will derive the dual optimization problem. For all questions, you may write $\mathbf{w}$ and $\mathbf{x}$ as w and x out of convenience.

19. Please write the Lagrangian for the above optimization problem. Express your answer as the Lagrangian $\mathcal{L}$ in terms of $\mathbf{w}$, b, α_i, where α_i is the Lagrange multiplier for the inequality constraint at each data instance i. **(5 points)**

    Ans:

    $$\mathcal{L}(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1]$$

    - Wrong sign: -1
    - Missing sum over i: -2
    - Missing the "1" in the expression: -2

20. Using your expression for the lagrangian $\mathcal{L}$ in Question 19, please write the stationary conditions of $\mathcal{L}$ with respect to $\mathbf{w}$ and b. Please simplify your answers as much as possible. **(5 points)**

    Ans:

    $$\begin{aligned} \nabla_{\mathbf{w}}\mathcal{L} &= \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = 0 \\ \implies \mathbf{w} &= \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \\ \nabla_b \mathcal{L} &= \sum_{i=1}^n \alpha_i y_i = 0 \end{aligned}$$

    - Missing result for bias: -2
    - Missing equality: -1
    - Did not set to 0: -1
    - Sign issue: -1 (still deducted if there was a sign issue in Q19)
    - Incorrect term -2
    - Missing term -1
    - Extra term -1
    - Extra condition -1
    - Issue with sum -1
    - Totally incorrect -5

21. Now write an expression for the minimized $\mathcal{L}$ such that $\mathcal{L}$ doesn't depend on $\mathbf{w}$ or b. Simplify your answer into two summation terms (one of which is a double summation). **(5 points)**

Ans:

$$\mathcal{L}(\alpha) = \underbrace{\sum_{i=1}^{n} \alpha_i}_{\textbf{(2 points)}} \underbrace{- \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j}_{\textbf{(3 points)}}$$

- write $x_i x_j$ instead of $\mathbf{x}_i^T \mathbf{x}_j$ in vector form, or $(\sum_i \alpha_i x_i y_i)^2$ instead of $\| \sum_i \alpha_i \mathbf{x}_i y_i \|^2$, -1
- write $\mathbf{x}_i \mathbf{x}_j^T$, wrong dimension, -0.5
- double summation w/o negative sign in the front, -1
- if write $\frac{1}{2} \sum_i \alpha_i$ or $-\sum_i \alpha_i$, -1

22. Finally, write the dual optimization problem with appropriate constraints. **(5 points)**

Ans:

$$\max_{\alpha} \quad L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \qquad \textbf{(1 points)}$$

$$\text{s.t.} \quad \alpha_i \geq 0, i = 1, \ldots, n \qquad \textbf{(2 points)}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \textbf{(2 points)}$$

- If the answer of Q21 is wrong, but the same wrong answer is written as the objective function, -0.5
- If the answer of Q20 is wrong, but the same wrong answer is written in the constraints, -1
- Extra wrong inequality constraint, -1
- Extra wrong equality constraint, -1
- Wrong sign for the objective function, -0.5

3 HMM/EM (26 points)

Recall a hidden Markov model is parameterized by

- initial probability: $\pi_s = P(X_1 = s)$
- transition probability: $a_{s,s'} = P(X_{t+1} = s' \mid X_t = s)$
- emission probability: $b_{s,o} = P(O_t = o \mid X_t = s)$

where X_t denotes the state at time t and O_t denotes the observation at time t.

Some probabilities of interest are defined as follow:

- the state at time t, given an observation sequence: $\gamma_s(t) = P(X_t = s \mid O_{1:T} = o_{1:T})$
- the likelihood of two consecutive states at time t: $\xi_{s,s'}(t) = P(X_t = s, X_{t+1} = s' \mid O_{1:T} = o_{1:T})$
- forward message: $\alpha_s(t) = P(X_t = s, O_{1:t} = o_{1:t})$
- backward message: $\beta_s(t) = P(O_{t+1:T} = o_{t+1:T} \mid X_t = s)$

23. Compute $\beta_s(T)$, where time T is the time step when the last observation O_T is observed. Hint: use $\gamma_s(t)$. **(5 points)**

Ans:

Full credit:

$$\gamma_s(t) = \frac{\alpha_s(t)\beta_s(t)}{P(O_{1:T} = o_{1:T})}.$$

$$\Rightarrow \quad \gamma_s(T) = \frac{\alpha_s(T)\beta_s(T)}{P(O_{1:T} = o_{1:T})}. \qquad \textbf{(2 points)}$$

On the other hand,

$$\begin{aligned}\gamma_s(T) &= P(X_t = s \mid O_{1:T} = o_{1:T}) \\ &= \frac{P(X_T = s, O_{1:T} = o_{1:T})}{P(O_{1:T} = o_{1:T})} \\ &= \frac{\alpha_s(T)}{P(O_{1:T} = o_{1:T})},\end{aligned} \qquad \textbf{(2 points)}$$

which yields

$$\begin{aligned}&\frac{\alpha_s(T)\beta_s(T)}{P(O_{1:T} = o_{1:T})} = \frac{\alpha_s(T)}{P(O_{1:T} = o_{1:T})} \\ \Rightarrow \quad &\beta_s(T) = 1 \qquad \textbf{(1 points)}\end{aligned}$$

Partial credit 4 points.

$$\begin{aligned}\beta_s(T) &= P(O_{T+1:T} = o_{T+1:T} \mid X_t = s) && \textbf{(1 points)} \\ &= P(\emptyset \mid X_t = s) && \textbf{(2 points)} \\ &= 1 && \textbf{(1 points)}\end{aligned}$$

24. Prove that $\gamma_s(t) = \sum_{s'} \xi_{s,s'}(t)$. Hint: use the definition of $\gamma_s(t)$. **(4 points)**

Ans:

$$\begin{aligned}\gamma_s(t) &= P(X_t = s \mid O_{1:T} = o_{1:T}) \\ &= \sum_{s'} P(X_t = s, X_{t+1} = s' \mid O_{1:T} = o_{1:T}) && \textbf{(2 points)} \\ &= \sum_{s'} \xi_{s,s'}(t). && \textbf{(2 points)}\end{aligned}$$

25. Recall that for the following optimization problem

$$\begin{aligned}&\max_x f(x) \\ &\text{s.t. } g_i(x) \le 0 \\ &\quad\ h_j(x) = 0,\end{aligned}$$

the KKT conditions are

- **Stationarity** $\nabla f(x^*) = \sum_i \mu_i \nabla g_i(x^*) + \sum_j \lambda_j \nabla h_j(x^*)$
- **Primal feasibility** $g_i(x^*) \le 0,\ h_j(x^*) = 0$
- **Dual feasibility** $\mu_i \ge 0$
- **Complementary slackness** $\mu_i g_i(x^*) = 0$

where μ_i and λ_j are the Lagrange multipliers.

Suppose we are using Baum-Welch algorithm to learn a hidden Markov model from one observation sequence $o_1, \cdots, o_T$, by maximizing the complete log-likelihood:

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau)}\right) = \sum_s \gamma_s(1) \ln \pi_s + \sum_{t=1}^{T-1} \sum_{s,s'} \xi_{s,s'}(t) \ln a_{s,s'} + \sum_{t=1}^{T} \sum_s \gamma_s(t) \ln b_{s,o_t}$$

Your task is to derive the update for transition probabilities $a_{s,s'}$ in terms of $\xi_{s,s'}(t)$ and $\gamma_s(t)$.

(a) Please write the Lagrangian for the corresponding optimization problem (you do need the entire complete log-likelihood Q). **(4 points)**

Ans: The update of $a_{s,s'}$ can be found by

$$\arg\max_{a_{s,s'}} \sum_{t=1}^{T-1} \sum_{s,s'} \xi_{s,s'}(t) \ln a_{s,s'}$$

$$\text{s.t. } a_{s,s'} \geq 0 \qquad \textbf{(1 points)}$$

$$\sum_{s'} a_{s,s'} = 1 \qquad \textbf{(1 points)}$$

Therefore the Lagrangian for each pair of (s, s') is

$$\mathcal{L}(a_{s,s'}, \mu_{s'}, \lambda) = \sum_{t=1}^{T-1} \sum_{s,s'} \xi_{s,s'}(t) \ln a_{s,s'} + \sum_{s'} \mu_{s'} a_{s,s'} - \lambda \left(\sum_{s'} a_{s,s'} - 1 \right)$$

$$\text{with } \mu_{s'} \geq 0 \qquad \textbf{(2 points)}$$

Also acceptable:

$$\mathcal{L}(a_{s,s'}, \mu_{s'}, \lambda) = \sum_{s} \gamma_s(1) \ln \pi_s + \sum_{t=1}^{T-1} \sum_{s,s'} \xi_{s,s'}(t) \ln a_{s,s'} + \sum_{t=1}^{T} \sum_{s} \gamma_s(t) \ln b_{s,o_t}$$

$$+ \sum_{s'} \mu_{s'} a_{s,s'} - \lambda \left(\sum_{s'} a_{s,s'} - 1 \right)$$

$$\text{with } \mu_{s'} \geq 0$$

(b) Please write the stationary conditions. **(4 points)**

Ans:

Stationarity:

$$\sum_{t=1}^{T-1} \xi_{s,s'}(t) \frac{1}{a^*_{s,s'}} = -\mu_{s,s'} + \lambda$$

(c) Using the other two KKT conditions express the Lagrange multipliers in terms of $\xi_{s,s'}(t)$. **(5 points)**

Ans:

Complementary slackness:

$$\mu_{s,s'} a^*_{s,s'} = 0 \qquad \textbf{(2 points)}$$

Primal feasibility:

$$\sum_{s'} a^*_{s,s'} = \sum_{s'} \frac{\sum_{t=1}^{T-1} \xi_{s,s'}(t)}{\lambda} = 1 \qquad \textbf{(3 points)}$$

(d) Eliminate the Lagrange multipliers and write the final update for transition probabilities $a_{s,s'}$ in terms of $\xi_{s,s'}(t)$ and $\gamma_s(t)$. **(4 points)**

Ans:

$$\text{Stationarity} \Rightarrow a^*_{s,s'} = \frac{\sum_{t=1}^{T-1} \xi_{s,s'}(t)}{\lambda - \mu_{s,s'}} \quad \textbf{(1 points)}$$

$$\text{Complementary slackness} \Rightarrow \mu_{s,s'} = 0 \quad \textbf{(1 points)}$$

$$\Rightarrow a^*_{s,s'} = \frac{\sum_{t=1}^{T-1} \xi_{s,s'}(t)}{\lambda}$$

$$\text{Primal feasibility} \Rightarrow \lambda = \sum_{s'} \sum_{t=1}^{T-1} \xi_{s,s'}(t) \quad \textbf{(1 points)}$$

$$\Rightarrow a^*_{s,s'} = \frac{\sum_{t=1}^{T-1} \xi_{s,s'}(t)}{\sum_{s''} \sum_{t=1}^{T-1} \xi_{s,s''}(t)} = \frac{\sum_{t=1}^{T-1} \xi_{s,s'}(t)}{\sum_{t=1}^{T-1} \gamma_s(t)} \quad \textbf{(1 points)}$$

4 Boosting (18 points)

In this question we will look into the AdaBoost algorithm (shown in Alg. 1), where the base algorithm is simply searching for a classifier with the smallest weighted error from a fixed classifier set $\mathcal{H}$.

Algorithm 1: Adaboost

Given: A training set $\{(\boldsymbol{x}_n, y_n \in \{+1, -1\})\}_{n=1}^N$, and a set of classifier $\mathcal{H}$, where each $h \in \mathcal{H}$ takes a feature vector as input and outputs $+1$ or -1.

Goal: Learn $H(\boldsymbol{x}) = \text{sign}\left(\sum_{t=1}^T \beta_t h_t(\boldsymbol{x})\right)$, where $h_t \in \mathcal{H}$, $\beta_t \in \mathbb{R}$, and $\text{sign}(a) = \begin{cases} +1, & \text{if } a \geq 0, \\ -1, & \text{otherwise.} \end{cases}$

Initialization: $D_1(n) = \frac{1}{N}$, $\forall n \in [N]$.

for $t = 1, 2, \cdots, T$ **do**

  Find $h_t = \arg\min_{h \in \mathcal{H}} \sum_{n: y_n \neq h(\boldsymbol{x}_n)} D_t(n)$.

  Compute

$$\epsilon_t = \sum_{n: y_n \neq h_t(\boldsymbol{x}_n)} D_t(n) \qquad \text{and} \qquad \beta_t = \frac{1}{2} \log \frac{1 - \epsilon_t}{\epsilon_t}.$$

  Compute

$$D_{t+1}(n) = \frac{D_t(n) e^{-\beta_t y_n h_t(\boldsymbol{x}_n)}}{\sum_{n'=1}^N D_t(n') e^{-\beta_t y_{n'} h_t(\boldsymbol{x}_{n'})}}$$

  for each $n \in [N]$

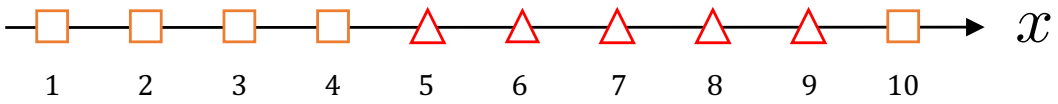

Figure 1: The 1-dimensional training set with 10 data. The square means the class of the data is $+1$, *i.e.* $y = +1$ and the triangle means $y = -1$. The number under each data is its x coordinate.

Now we are given a training set of 10 data as shown in Fig. 1. Each training data is 1-dimension and denoted as a square or a triangle in the figure, where the square means the class of the data is $+1$, *i.e.* $y = +1$ and the triangle means $y = -1$. You are going to experiment on the given training set with the learning process of the AdaBoost algorithm as shown in Alg. 1 for $T = 2$. The base classifier set $\mathcal{H}$ consists of all decision stumps, where each of them is parameterized by a pair $(s, b) \in \{+1, -1\} \times \mathbb{R}$ such that

$$h_{(s,b)}(x) = \begin{cases} s, & \text{if } x > b, \\ -s, & \text{otherwise.} \end{cases}$$

26. Please write down the pair (s, b) of the best decision stump h_1, and ϵ_1 at $t = 1$. If there are multiple equally optimal stump functions, just randomly pick **one** of them to be h_1. Write ϵ_1 in the form of **a fraction**. **(3 points)**

Ans: $s = -1$ and $b \in [4, 5)$ (any b in this range is fine). **(1 points)**

$\epsilon_1 = \frac{1}{10}$. **(2 points)**

27. Please write down the pair (s, b) of the best decision stump h_2, and ϵ_2 at $t = 2$. If there are multiple equally best stump functions, just randomly pick **one** of them to be h_2. Write ϵ_2 in the form of **a fraction**. You don't need to compute the actual value of $e^{\frac{1}{2}\log c}$, instead try to cancel them out. **(3 points)**

Ans:

$s = 1$ and $b \in [9, 10)$ (any b is this range is fine). **(1 points)**

$\epsilon_2 = \frac{\frac{4}{10}e^{-\frac{1}{2}\log 9}}{\frac{1}{10}e^{\frac{1}{2}\log 9} + \frac{9}{10}e^{-\frac{1}{2}\log 9}} = \frac{2}{9}$. **(2 points)**

28. Now we have h_1 and h_2, further calculate β_1 and β_2. Both β_1 and β_2 should be written in the form of $\frac{1}{2}\log c$, where c is **a integer** or **a fraction**. **(6 points)**

Ans: We can calculate β_1 and β_2 following line 6 of Alg. 1:

$$\beta_1 = \frac{1}{2}\log 9 \textbf{ (2 points)} \tag{1}$$

$$\beta_2 = \frac{1}{2}\log(\frac{1-\epsilon_2}{\epsilon_2}) \tag{2}$$

$$= \frac{1}{2}\log\frac{7}{2} \textbf{ (4 points)}$$

29. Write down the final classifier $H(\boldsymbol{x}) = \beta_1 h_1 + \beta_2 h_2$, the class predicted by $H(\boldsymbol{x})$ for each data point and calculate the final training accuracy. **(6 points)**.

Ans:

Substitute β_1 and β_2 into $H(\boldsymbol{x}) = sign(\beta_1 h_1 + \beta_2 h_2)$, we have that:

$$H(\boldsymbol{x}) = sign(\beta_1 h_1 + \beta_2 h_2) = sign(\frac{1}{2}(\log 9)h_1 + \frac{1}{2}(\log \frac{7}{2})h_2) \textbf{ (3 points)}$$

$H(\boldsymbol{x})$ without sign is also fine.

The class of each data is listed as following: **(1 points)**

Data	1	2	3	4	5	6	7	8	9	10
Class	+1	+1	+1	+1	-1	-1	-1	-1	-1	-1

Training accuracy is $\frac{9}{10} = 0.9$. **(2 points)**

5 GMM (10 points)

A GMM has the following density function:

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

where K is the number of Gaussian components, $\boldsymbol{\mu}_k$ and Σ_k are the mean and covariance matrix of the k-th Gaussian, ω_k's are the mixture weights, $\boldsymbol{\theta}$ is the parameters ($\boldsymbol{\mu}_k$'s, Σ_k's and ω_k's) and $\mathbf{x} \in \mathbb{R}^{D\times 1}$ is the observed random variable. We define the log likelihood of N samples as

$$P(\boldsymbol{\theta}) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n;\ \boldsymbol{\theta}),$$

and the soft-assignment as

$$\gamma_{nk} = p(z_n = k | \mathbf{x}_n;\ \boldsymbol{\theta}),$$

where z_n's are the latent variables.

30. Suppose we would like to use gradient descent to maximize the log-likelihood, find the gradient $\frac{\partial P(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k}$.

Ans:

$$\frac{\partial P(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} = \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \omega_k \frac{1}{\sqrt{(2\pi)^D |\Sigma_k|}} \exp\left[-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\right] \right\} \quad \text{expand } p(x_n;\theta) \quad \textbf{(2 points)}$$

$$= \sum_{n=1}^{N} \underbrace{\frac{\omega_k \frac{1}{\sqrt{(2\pi)^D |\Sigma_k|}} \exp\left[-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\right]}{\sum_{k'=1}^{K} \omega_{k'} \frac{1}{\sqrt{(2\pi)^D |\Sigma_{k'}|}} \exp\left[-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_{k'})^T \Sigma_{k'}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_{k'})\right]}}_{\textbf{(2 points)}} \underbrace{\Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)}_{\textbf{(2 points)}} \quad \textbf{(4 points)}$$

$$= \sum_{n=1}^{N} \frac{\omega_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k\right)}{\sum_{k'=1}^{K} \omega_{k'} \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_{k'}, \Sigma_{k'}\right)} \Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k) \quad \text{write as posterior or } \gamma_{nk} \quad \textbf{(2 points)}$$

$$= \sum_{n=1}^{N} \gamma_{nk} \Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k) \quad \textbf{(2 points)}$$

Other rubrics

- if recognize $\gamma_{nk} = \frac{\omega_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k'=1}^{K} \omega_{k'} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_{k'}, \Sigma_{k'})}$, +1
- if only write $\frac{1}{p(x_n;\theta)} \frac{\partial p(x_n;\theta)}{\partial \mu_k}$ without specifying Gaussian distribution, −1
- if not combine the 2 terms, $\frac{1}{2}\Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k) + \frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}$, −1
- $\Sigma_k^{-1}(x_n-\mu_k)$ without Σ_k^{-1} or write as $\frac{1}{|\Sigma_k|}$, −1

Wrong answer The following reasoning is incorrect for this question.

$$\ln p(\mathbf{x}_n;\boldsymbol{\theta}) \geq \mathbb{E}_{\boldsymbol{z}_n\sim q(\boldsymbol{z})}\left[\ln p(\mathbf{x}_n,\boldsymbol{z}_n;\boldsymbol{\theta})\right],$$

and then we take $q(\boldsymbol{z}_n = k) = \gamma_{nk}$ to get,

$$\begin{aligned}
\ln p(\mathbf{x}_n;\boldsymbol{\theta}) &\geq \sum_k \gamma_{nk} \ln\left(\omega_k \frac{1}{\sqrt{(2\pi)^D|\Sigma_k|}} \exp\left[-\frac{1}{2}(\mathbf{x}_n-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\right]\right)\\
\frac{\partial \ln p(\mathbf{x}_n;\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} &= \gamma_{nk}\Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\\
\frac{\partial P(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}_k} &= \sum_{n=1}^{N}\gamma_{nk}\Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k).
\end{aligned}$$

Note that,

$$\begin{aligned}
\ln p(\mathbf{x}_n;\boldsymbol{\theta}) &= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\mathbf{x}_n;\boldsymbol{\theta})\ln p(\mathbf{x}_n) = \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\mathbf{x}_n;\boldsymbol{\theta})\ln\left(\frac{p(\mathbf{x}_n,\boldsymbol{z}_n;\boldsymbol{\theta})}{p(\boldsymbol{z}_n|\mathbf{x}_n;\boldsymbol{\theta})}\right)\\
&= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\mathbf{x}_n;\boldsymbol{\theta})\ln p(\mathbf{x}_n,\boldsymbol{z}_n;\boldsymbol{\theta}) - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\mathbf{x}_n;\boldsymbol{\theta})\ln p(\boldsymbol{z}_n|\mathbf{x}_n;\boldsymbol{\theta})\\
&= \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\mathbf{x}_n;\boldsymbol{\theta})\ln p(\mathbf{x}_n,\boldsymbol{z}_n;\boldsymbol{\theta}) - \sum_k \gamma_{nk}\ln(\gamma_{nk})
\end{aligned}$$

The answer is incomplete if you did not explicitly justify why $\frac{\partial}{\partial\boldsymbol{\mu}_k}\sum_k \gamma_{nk}\ln(\gamma_{nk}) = 0$

You may use this page as scratch paper

You may use this page as scratch paper