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

Problem	1	2	3	4	5	Total
Points	20	20	20	20	20	100

Please read the following instructions carefully:

- The exam has a total of **13 pages** (including this cover and one 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 **2 hours and 20 minutes**. 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.
- Answers should be **concise** and written down **legibly**. All questions can be done within 5-12 lines.
- You must answer each question on the page provided. You can use the last two blank pages as scratch paper. Raise your hand to ask a proctor for more if needed.
- 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.
- You **may not** leave your seat **for any reason** unless you submit your exam at that point.

1 Multiple Choice, True or False

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

   (a) True.

   (b) False.

2. 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.

3. Which of the following statement is true? Ans: D

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

   (B) Boosting algorithm cannot select the same weak classifier more than once.

   (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

4. When applying a GMM of K components to a dataset of N points, if we representing γ_{nk} (which is the term used to update component weights) as a matrix of N rows and K columns, what is the sum of this matrix? Ans: C

   (A) 1

   (B) K

   (C) N

   (D) NK

5. What is the minimal number of freely learnable parameters in a GMM with K components and full covariance? Suppose that the dataset has N points, each being D dimensional. Ans: D. A covariance matrix is symmetric.

   (A) $KD + NK^2$

   (B) $K(N + D^2)$

   (C) $KD(D + 1)$

   (D) $K(D + D(D + 1)/2)$

6. Which of the following models has a continuous latent variable? Ans: B

   (a) Naive Bayes Classifier

   (b) Principal Component Analysis

   (c) Gaussian Mixture Model

   (d) Hidden Markov Model

7. The effectiveness of an SVM depends on: Ans: D

(a) Selection of Kernel

(b) Kernel Parameters

(c) Soft Margin Parameter C

(d) All of the above

8. Which of the following statements is NOT TRUE about Lagrangian duality? Ans: C

   (a) One purpose of Lagrange duality is to find a lower bound on a minimization problem or an upper bound for a maximization problem.

   (b) Duality allows us to formulate optimality conditions for constrained optimization problems.

   (c) The Lagrangian can be used to express the dual problem as an unconstrained one.

   (d) None of the above

9. Which of the following statements is TRUE about SVM? Ans: C

   (a) SVM gives us an unconstrained, smooth objective.

   (b) SVM gives calibrated probabilities that can be interpreted as confidence in a decision.

   (c) SVMs don't penalize examples for which the correct decision is made with sufficient confidence.

   (d) None of the above

10. Baum-Welsh algorithm can be used to maximize probability of an observation in an HMM model. Ans: A

    (a) TRUE

    (b) FALSE

11. Given the parameters of an HMM and an observation sequence O, we can determine the most likely state sequence. Ans: A

    (a) TRUE

    (b) FALSE

12. Which of the following statement about HMM is True? Ans: D

    (a) It can be used for Speech Recognition, and Vehicle Trajectory Prediction

    (b) It is a special case of the EM algorithm which is a iterative algorithm

    (c) The forward-backward algorithm can do completely unsupervised learning of the transition matrix A and emission matrix B parameters.

    (d) All of the above

2 Mixture Models (20 points)

Consider an Laplace Mixture Model with the following probability mass function:

$$p(x_n) = \sum_{k=1}^{K} p(x_n, z_n = k) = \sum_{k=1}^{K} p(z_n = k)p(x_n \mid z_n = k) = \sum_{k=1}^{K} \omega_k \pi(x_n \mid \mu_k, \sigma_k) = \sum_{k=1}^{K} \omega_k \frac{1}{2\sigma_k} e^{-\frac{|x_n - \mu_k|}{\sigma_k}}$$

where ω_k is the mixture weight such that $\sum_{k=1}^{K} \omega_k = 1$, $x_n \in \mathcal{R}$ is an observation, K is the number of mixtures, and $\{\mu_k, \sigma_k\}_{k=1}^{K}$ are the model parameters.

Similar to Gaussian Mixture Models, The expected complete log-likelihood is defined as

$$\mathcal{Q} = \sum_{n=1}^{N} \sum_{k=1}^{K} \big(\gamma_{nk} \log p(x_n, z_n = k) - \gamma_{nk} \log \gamma_{nk}\big),$$

where N is the number of data points.

13. To get the optimal ω_k, we have the following optimization problem:

$$\begin{aligned} &\arg_{\omega_k} \max \mathcal{Q}, \\ &s.t.\ \omega_k \geq 0, \\ &\sum_{k=1}^{K} \omega_k = 1. \end{aligned}$$

Write out the Langragian and find the optimal ω_k (treating all other variables constant). Note that you should directly NOT apply the results from the assignments. **(10 points)**

Ans:

$$\begin{aligned} \arg_{\omega_k} \max \mathcal{Q} &= \arg_{\omega_k} \max \sum_{n=1}^{N} \sum_{k=1}^{K} \big(\gamma_{nk} \log p(x_n, z_n = k) - \gamma_{nk} \log \gamma_{nk}\big) \quad \text{(definition)} \\ &= \arg_{\omega_k} \max \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log p(x_n, z_n = k) \quad (\gamma_{nk} \log \gamma_{nk} \text{ is unrelated to } \omega_k) \\ &= \arg_{\omega_k} \max \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \omega_k \frac{1}{2\sigma_k} e^{-\frac{|x_n - \mu_k|}{\sigma_k}} \quad \text{(definition)} \\ &= \arg_{\omega_k} \max \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \omega_k \quad (\text{remove terms unrelated to } \omega_k) \end{aligned}$$

Excluding the terms unrelated to ω_k, we can write the Lagrangian as

$$\mathcal{L} = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \omega_k + \mu(\sum_{k} \omega_k - 1) + \sum_{k} \mu_k \omega_k$$

Set the derivative to zero:

$$\nabla_{\omega_k} \mathcal{L} = \sum_{n} \gamma_{nk} \frac{1}{\omega_k} + \mu + \mu_k = 0. \tag{1}$$

By complementary slackness, we have $\mu_k = 0$ 1 and (1) becomes

$$\frac{1}{\omega_k}\sum_n \gamma_{nk} = -\mu \Rightarrow \omega_k = -\frac{1}{\mu}\sum_n \gamma_{nk} \tag{2}$$

By feasibility, we have

$$\sum_k \omega_k = 1 \Rightarrow \sum_k -\frac{\sum_n \gamma_{nk}}{\mu} = 1 \Rightarrow \mu = -\sum_n\sum_k \gamma_{nk} = N$$

Thus, (2) becomes

$$\omega_k = \frac{\sum_n \gamma_{nk}}{\sum_n \sum_k \gamma_{nk}} = \frac{1}{N}\sum_n \gamma_{nk}$$

Remarks:

- x partial credits: only the first few steps are correct.
- incorrect answer: everything before the final answer is correct.
- incomplete answer: mostly correct, with small defects such as not representing the denominator with N.

14. Find the optimal σ_k (treating all other variables constant). **(10 points)**

Ans: Set the derivative to zero, we have

$$\begin{aligned}\log p(x_n, z_n = k) &= \log \omega_k \frac{1}{2\sigma_k} e^{-\frac{|x_n-\mu_k|}{\sigma_k}} \qquad (3)\\ &= \log \omega_k + \log\frac{1}{2} - \log\sigma_k - \frac{|x_n - \mu_k|}{\sigma_k} \qquad (4)\end{aligned}$$

$$\begin{aligned}\arg_{\sigma_K}\max \mathcal{Q} &= \arg_{\sigma_k}\max \sum_{n=1}^{N}\sum_{k=1}^{K}\big(\gamma_{nk}\log p(x_n, z_n = k) - \gamma_{nk}\log\gamma_{nk}\big) \quad \text{(definition)}\\ &= \arg_{\sigma_k}\max \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\log p(x_n, z_n = k) \quad (\gamma_{nk} \text{ is unrelated to } \sigma_k)\\ &= \arg_{\sigma_k}\max \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(\log\omega_k + \log\frac{1}{2} - \log\sigma_k - \frac{|x_n-\mu_k|}{\sigma_k}) \quad \text{(from (3))}\\ &= \arg_{\sigma_k}\max \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}(-\log\sigma_k - \frac{|x_n-\mu_k|}{\sigma_k}) \quad \text{(remove terms unrelated to } \sigma_k)\\ &= \arg_{\sigma_k}\max \mathcal{Q}'\end{aligned}$$

$$
\begin{aligned}
\nabla_{\sigma_k}\mathcal{Q} &= \nabla_{\sigma_k}\mathcal{Q}^{'} \\
&= \sum_{n=1}^{N}\nabla_{\sigma_k}\sum_{k=1}^{K}\gamma_{nk}(-\log\sigma_k - \frac{|x_n-\mu_k|}{\sigma_k}) \\
&= \sum_{n}\gamma_{nk}(-\frac{1}{\sigma_k}+\frac{|x_n-\mu_k|}{\sigma_k^2}) = 0 \\
\Rightarrow \sigma_k &= \frac{\sum_n \gamma_{nk}|x_n-\mu_k|}{\sum_n \gamma_{nk}}
\end{aligned}
$$

Remarks:

- x partial credits: only the first few steps are correct.
- incorrect answer: everything before the final answer is correct.
- incomplete answer: mostly correct, with small defects such as an extra 2 in the solution.

3 Support Vector Machine (20 points)

Given an unlabeled set of examples $\{x_1, \ldots, x_N\}$, the one-class SVM algorithm tries to find a direction $\mathbf{w}$ that maximally separates the data from the origin. More precisely, it solves the following optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$s.t. \quad \mathbf{w}^T\mathbf{x}_n + b \geq 1 \;\; \forall n = \{1, \ldots, N\}$$

A new test example x is labeled 1 if $\mathbf{w}^T\mathbf{x} + b \geq 1$, and 0 otherwise.

15. Write down the corresponding dual optimization problem for the above. Your answer should not have any term of $\mathbf{w}$. **(10 points)**

Ans:

$$L\left(\mathbf{w}, \lambda_i\right) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{N} \lambda_i \left(1 - \mathbf{w}^T\mathbf{x}_i - b\right)$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \mathbf{w} - \sum_{i=1}^{N} \lambda_i \mathbf{x}_i = 0$$

$$\mathbf{w} = \sum_{i=1}^{N} \lambda_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \lambda_i = 0$$

Substituting we get,

$$\tilde{L}(\lambda_i) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \mathbf{x}_i^T \mathbf{x}_j$$

and dual objective would be

$$\max_{\lambda_i, \lambda_j \geq 0} \tilde{L}(\lambda_i)$$

16. Can the one-class SVM be kernelised in training? How? **(4 points)**

Ans: Since, the dual is expressed only in terms of dot product, it can be kernelized easily and dual objective would become

$$\tilde{L}(\lambda_i) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j k(\mathbf{x}_i, \mathbf{x}_j)$$

17. Can the one-class SVM be kernelised in testing? How? **(6 points)**

Ans: Note, that for testing we need to check value of $\mathbf{w}^T\mathbf{x}$, which is $\sum_{i=1}^{N} \lambda_i \mathbf{x}_i^T \mathbf{x} = \sum_{i=1}^{N} \lambda_i k(\mathbf{x}_i, \mathbf{x})$ Hence, it can be kernelized in testing too

4 Boosting (20 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 $n \in [N]$
```

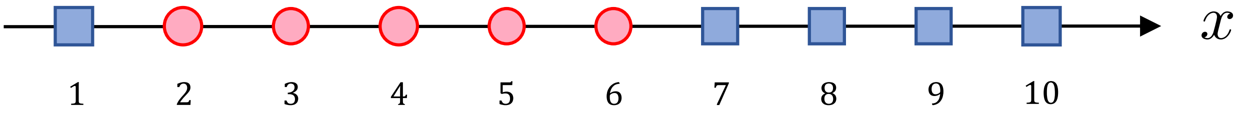

Figure 1: The 1-dimensional training set with 10 data. A square means the class of the data is -1, *i.e.* $y=-1$ and a circle means $y=+1$. The number under each data indicates its x coordinate.

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 circle in the figure, where the square refers the class of the data is -1, *i.e.* $y=-1$ and the circle refers $y=+1$. 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 the decision stumps is parameterized by a pair $(s,b) \in \{+1,-1\} \times \mathbb{R}$ such that

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

Throughout this problem, the natural logarithm which has the e (≈ 2.71828) as its base is applied for the $\log(\cdot)$.

18. 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 down which data points are mis-classified, and ϵ_1. ϵ_1 should be in the form of **a fraction**. **(5 points)**

Ans: $s = -1$ and $b \in [6, 7)$ (any b in this range is fine). **(2 points)**
The left-most circle will be mis-classified. **(1 points)**
$\epsilon_1 = \frac{1}{10}$ **(2 points)**

19. 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** following Alg. 1. You don't need to compute the actual value of $e^{\frac{1}{2}\log c}$, instead try to cancel them out with the $\exp(\cdot)$ and the $\log(\cdot)$ properties. **(7 points)**

Ans: $s = +1$ and $b \in [1, 2)$ (any b is this range is fine). **(3 points)**
The right-most 4 squares will be mis-classified.

$$\begin{aligned}
\epsilon_2 &= \frac{\frac{4}{10}e^{-\frac{1}{2}\log 9}}{\frac{1}{10}e^{\frac{1}{2}\log 9} + \frac{9}{10}e^{-\frac{1}{2}\log 9}} \\
&= \frac{4e^{-\frac{1}{2}\log 9}}{e^{\frac{1}{2}\log 9} + 9e^{-\frac{1}{2}\log 9}} \\
&= \frac{4}{e^{\log 9} + 9} \\
&= \frac{4}{9 + 9} \\
&= \frac{4}{18} = \frac{2}{9} \textbf{ (4 points)}
\end{aligned}$$

20. Now we have the final classifier $H(\boldsymbol{x}) = \text{sign}\,(\beta_1 h_1 + \beta_2 h_2)$: write down β_1, β_2, the class predicted by $H(\boldsymbol{x})$ for each data point and the final training accuracy. The training accuracy should be presented as a fraction. **(8 points)**.

Ans:

Substitute $\beta_1 = \frac{1}{2}\log 9$ and $\beta_2 = \frac{1}{2}\log\frac{7}{2}$ into $H(\boldsymbol{x}) = \text{sign}(\beta_1 h_1 + \beta_2 h_2)$, we have that:

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

(3 points)

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

The class of each data is listed in the table **(2 points)**

Training accuracy is $\frac{9}{10}$. **(3 points)**

5 HMM (20 points)

Recall a hidden Markov model is parameterized by

- initial state distribution $P\left(X_1 = s\right) = \pi_s$
- transition distribution $P\left(X_{t+1} = s' | X_t = s\right) = a_{s,s'}$
- emission distribution $P\left(O_t = o | X_t = s\right) = b_{s,o}$

21. Given a sequence of observations, our goal is to adjust the model parameters (π, A, B) to best fit the observations. Compute $\gamma_s(t)$ which is a probability of being at state $X_t = s$ at time t. **(7 points)**

$$\alpha_s(t) = P\left(X_t = s, O_{1:t} = o_{1:t}\right)$$
$$\beta_s(t) = P\left(O_{t+1:T} = o_{t+1:T} | X_t = s\right)$$

Ans:

$$\begin{aligned}
\gamma_s(t) &= P\left(X_t = s | O_{1:T}\right) \textbf{ (1 points)} \\
&= P\left(X_t = s, O_{1:T}\right) / P\left(O_{1:T}\right) \textbf{ (1 points)} \\
&= P\left(O_{t+1:T} | X_t = s, O_{1:t}\right) P\left(X_t = s, O_{1:t}\right) / P\left(O_{1:T}\right) \textbf{ (1 points)} \\
&= P\left(O_{t+1:T} | X_t = s\right) \alpha_s(t) / P\left(O_{1:T}\right) \textbf{ (1 points)} \\
&= \beta_s(t) \alpha_s(t) / P\left(O_{1:T}\right) \textbf{ (1 points)} \\
&= \frac{\beta_s(t)\alpha_s(t)}{\sum_s \alpha_s(T)} \textbf{(2 points)}
\end{aligned}$$

22. Consider a HMM model with states $X_t \in \{S_1, S_2, S_3\}$, observations $O_t \in \{A, B, C\}$, and parameters

$$\pi_1 = P(X_1 = S_1) = 1; \quad \pi_2 = P(X_1 = S_2) = 0; \quad \pi_3 = P(X_1 = S_3) = 0$$

$$\begin{aligned} a_{11} = 1/4, \quad a_{12} = 1/2, \quad a_{13} = 1/4 \\ a_{21} = 0, \quad a_{22} = 1/2, \quad a_{23} = 1/2 \\ a_{31} = 1, \quad a_{32} = 0, \quad a_{33} = 0 \end{aligned}$$

$$\begin{aligned} b_1(A) = 1/2, \quad b_1(B) = 1/2, \quad b_1(C) = 0 \\ b_2(A) = 1/2, \quad b_2(B) = 0, \quad b_2(C) = 1/2 \\ b_3(A) = 0, \quad b_3(B) = 1/2, \quad b_3(C) = 1/2 \end{aligned}$$

What is $P(X_3 = S_1)$? hint: you do not need to run Viterbi algorithm. **(6 points)**

Ans:

$$\begin{aligned} P(X_3 = S_2) = 1/4 \times 1/2 + 1/2 \times 1/2 = 3/8 \\ P(X_3 = S_3) = 1/4 \times 1/4 + 1/2 \times 1/2 = 5/16 \\ 1 - P(X_3 = S_2) - P(X_3 = S_3) = 1 - 3/8 - 5/16 = 5/16 \textbf{(6 points)} \\ \textit{or} \\ P(X_3 = S_1) = 1/4 \times 1/4 + 1/4 \times 1 = 1/16 + 1/4 = 5/16 \textbf{(6 points)} \end{aligned}$$

23. Write down the sequence of $X_{1:6}$ with the maximal posterior probability assuming the observation sequence is ABCABC. What is that posterior probability? hint: you do not need to run Viterbi algorithm. **(7 points)**

Ans: $S_1S_1S_3S_1S_1S_2$ **(4 points)**
posterior probability: 2/3 **(3 points)**

Calculation Process:
A B C A B C
S1 S1 S1 delete
S1 S1 S2 S2 S2 delete
S1 S1 S2 S2 S3 S1 delete
S1 S1 S2 S3 delete
S1 S1 S3 S1 S1 S2
S1 S1 S3 S1 S1 S3
S1 S1 S3 S1 S2 delete
S1 S1 S3 S1 S3 S1 delete
S1 S2 delete
S1 S3 delete

You may use this page as scratch paper, but nothing written on it will be graded.