CSCI-567 Spring 2019 Practice Final Exam

Section	Points Available
Multiple choice	26
Duality & SVM	20
Boosting	15
Gaussian mixture model / EM	19
Hidden Markov model	10
Principal component analysis	10
Total	100

Please read the following instructions carefully:

- The exam has a total of **11 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 **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.
- Answers should be **concise** and written down **legibly**. All questions can be done within 5-20 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 (26 points)

Lagrangian duality and SVM (6 points)

1. Which of the following is not a true statement about Lagrangian duality?

   (A) The lagrangian dual function is convex.

   (B) Duality lets us formulate optimality conditions for constrained optimization problems.

   (C) If strong duality holds we may have found an easier approach to our primal problem.

   (D) In the primal version of SVM, the Lagrangian is minimized with respect to optimization parameters: $\mathbf{w}$ and b.

2. In a soft margin SVM, what is the behavior of the width of the margin ($\frac{1}{\|\mathbf{w}\|}$) as $C \to 0$?

   (A) Behaves like hard margin

   (B) Goes to zero

   (C) Goes to infinity

   (D) None of the above

3. Why can SVMs train quickly?

   (A) They can be solved with convex optimization.

   (B) They can leverage the kernel trick.

   (C) They can be optimized in the dual space.

   (D) All of the above.

Boosting (4 points)

4. Which of the following statement is true?

   (A) The Adaboost algorithm will eventually give zero training error regardless of the type of weak classifier it uses, provided enough iterations are performed.

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

   (C) In the Adaboost algorithm, weights of the misclassified examples goes up.

   (D) All of the above.

5. Which of the following statement about Adaboost algorithm is true?

   (A) Adaboost is sensitive to outliers.

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

   (C) Adaboost is not robust to overfitting

   (D) None of the above

Gaussian Mixture Model / EM (6 points)

6. Which of the following statements of the Expectation-Maximization (EM) algorithm is correct?

   (A) The EM algorithm maximizes the expectation of latent variables $\mathbb{E}[z_n]$.

   (B) In the expectation (E) step, we can use an arbitrary distribution $q_n(z_n)$ to maximize the lower bound of the log-likelihood $P(\boldsymbol{\theta})$.

   (C) In the maximization (M) step, the objective likelihood is guaranteed to be increased by updating the model parameters (if not converged).

   (D) None of above.

7. Which of the following statements of Gaussian Mixture Model (GMM) is correct?

   (A) The parameters of a GMM can be learned via maximum-likelihood estimation (MLE).

   (B) Gradient descent cannot be used to learn a GMM.

   (C) GMM is a supervised learning method.

   (D) None of above.

8. Which of the following statements of GMM and K-means is correct?

   (A) Given a set of data points and a fixed number of clusters K, applying GMM and K-means will always result in same cluster centroids.

   (B) Given a set of data points and a fixed number of clusters K, applying GMM and K-means will always result in different cluster centroids.

   (C) Given a learned GMM, we assign a data point to a cluster if the distance from the data point to its centroid is the smallest.

   (D) Given a learned K-means model, we assign a data point to a cluster if the distance from the data point to its centroid is the smallest.

Hidden Markov Model (4 points)

9. Which is not true about the Baum-Welsh algorithm?

   (A) It is used to find unknown parameters of a hidden markov model.

   (B) It uses a forward-backward algorithm to maximize the probability of an observation.

   (C) It computes the most likely sequence of hidden states given an observation sequence.

   (D) It is a special case of the EM algorithm.

10. Which of the following is not a task for Hidden Markov Models?

    (A) Compute the probability that the observations are generated by the model.

    (B) Represent dependencies between hidden states.

    (C) Compute the most likely state sequence in the model that produced the observations.

    (D) Adjust parameters for prediction optimization.

PCA and dimensionality reduction (4 points)

11. Given d-dimensional data $\{x_i\}_{i=1}^{n}$, you run principle component analysis (PCA) and pick k principle components. In which of the following case, you can achieve loss-less compression of the data (zero reconstruction error)?

    (A) You can always achieve loss-less compression.

    (B) You can achieve loss-less compression if $k = d$.

    (C) You can achieve loss-less compression if $k = n$.

    (D) No, you can never achieve loss-less compression.

12. Which of the following statement is correct?

    (A) PCA can map high-dimensional data to low-dimensional data by non-linear transformation.

    (B) Let D be the dimension of original data. Suppose we project the data into D dimensional space using PCA. Then, the projected data are the same as the original data.

    (C) All principal components are orthogonal to each other.

    (D) Standard PCA is a supervised learning method.

(2 points)

13. Which of the following statement is correct?

    (A) Linear regression, k-nearest neighbors, support vector machines, logistic regression, perceptron algorithm and PCA can all be kernelized.

    (B) Cross validation can be used to select the number of weak learners added in boosting; this procedure may help reduce overfitting.

    (C) Gaussian mixture models provide a probabilistic interpretation for K-Means (but we only need to estimate the means in this case).

    (D) All of the above

2 Lagrangian Duality & SVM (20 points)

Consider a dataset consisting of points (x, y), where x is a real value, and $y \in \{-1, 1\}$ is the class label. The dataset contains three points $(x_1, y_1) = (0, 1), (x_2, y_2) = (-1, -1)$, and $(x_3, y_3) = (1, -1)$. Let the feature mapping $\boldsymbol{\phi}(u) = [1, \sqrt{2}u, u^2]^T$, and let $\mathbf{w} = [w_1, w_2, w_3]^T$. A max-margin SVM solves the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i[\mathbf{w}^T \boldsymbol{\phi}(x_i) + b] \geq 1, \quad i = 1, 2, 3 \end{aligned}$$

14. Please explain why all three points in the dataset are support vectors.

15. Using the method of Lagrange multipliers, show that the solution is $\mathbf{w}^* = [0, 0, -2]^T$, $b^* = 1$ and the margin is $\frac{1}{\|\mathbf{w}^*\|_2}$. Hint: Use the fact that $y_i[\mathbf{w}^T \boldsymbol{\phi}(x_i) + b] = 1, i = 1, 2, 3$.

3 Boosting (15 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
        $\epsilon_t = \sum_{n: y_n \neq h_t(\boldsymbol{x}_n)} D_t(n)$ and $\beta_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$.
    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 $n \in [N]$
```

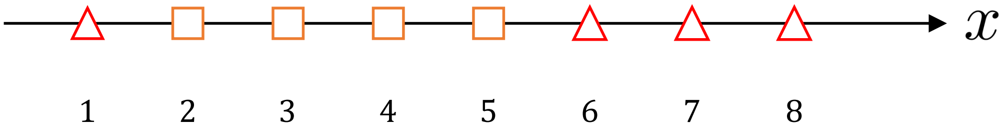


Figure 1: The 1-dimensional training set with 8 data. The square means the class of the data is $+1$, *i.e.* $y = +1$ and the triangle means $y = -1$. The number under each data is its x coordinate.

Now we are given a training set of 8 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

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

16. Please write down the pair (s, b) of the best decision stump h_1, ϵ_1 and the mis-classified data at $t = 1$. If there are multiple equally optimal stump functions, just randomly pick **ONE** of them to be h_1. **(6 points)**

17. Please write down the pair (s, b) of the best decision stump h_2, ϵ_2 and the mis-classified data at $t = 2$. If there are multiple equally best stump functions, just randomly pick **ONE** of them to be h_2. **(6 points)**

18. Suppose we run AdaBoost for two rounds and observe that β_1 and β_2 are both positive but not equal. Will the training accuracy of the final classifier H after these two rounds (see Line 2) be 1? Explain why or why not. **(3 points)**

4 GMM/EM (19 points)

The general Expectation-Maximization (EM) algorithm is summarized as follow:

```
Algorithm 2: General EM algorithm
  Step 0
    Initialize $\theta^{(1)}$, $t = 1$
  Step 1 (E-Step)
    1-1 Update the posterior of latent variables
        $q_n^{(t)}(\cdot) = p\left(\cdot|\mathbf{x}_n; \boldsymbol{\theta}^{(t)}\right)$
    1-2 Obtain Expectation of complete likelihood
        $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right) = \sum_{n=1}^{N} \mathbb{E}_{z_n \sim q_n^{(t)}} \left[\ln p\left(\mathbf{x}_n, z_n; \boldsymbol{\theta}\right)\right]$
  Step 2 (M-Step)
    Update the model parameter via Maximization
        $\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}\right)$
  Step 3
    $t \leftarrow t + 1$ and return to Step 1 if not converged
```

Consider a mixture model with the following density function

$$p(\mathbf{x}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \omega_k \nu_l \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_l\right).$$

To find all the parameters $\boldsymbol{\theta} = \{\omega_k, \nu_l, \boldsymbol{\mu}_k, \Sigma_l\}_{k=1}^{K}{}_{l=1}^{L}$ using MLE approach, we want to find

$$\arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \ln p\left(\mathbf{x}_n; \boldsymbol{\theta}\right).$$

We would like to use the EM algorithm to maximize the above likelihood, for which we introduce the latent variables $\mathbf{z}_n = [z_{n,1},\ z_{n,2}]^T$. The mixture weights are therefore decomposed as $\omega_k = p\left(z_{n,1} = k\right)$ and $\nu_l = p\left(z_{n,2} = l\right)$. We assume that $z_{n,1}$ and $z_{n,2}$ are independent, i.e. $p(\mathbf{z}_n; \boldsymbol{\theta}^{(t)}) = p(z_{n1}; \boldsymbol{\theta}^{(t)}) p(z_{n2}; \boldsymbol{\theta}^{(t)})$.

19. Derive the posterior of latent variables $p\left(\mathbf{z}_n = [k,\ l]^T|\ \mathbf{x}_n; \boldsymbol{\theta}^{(t)}\right)$ for the E-Step.

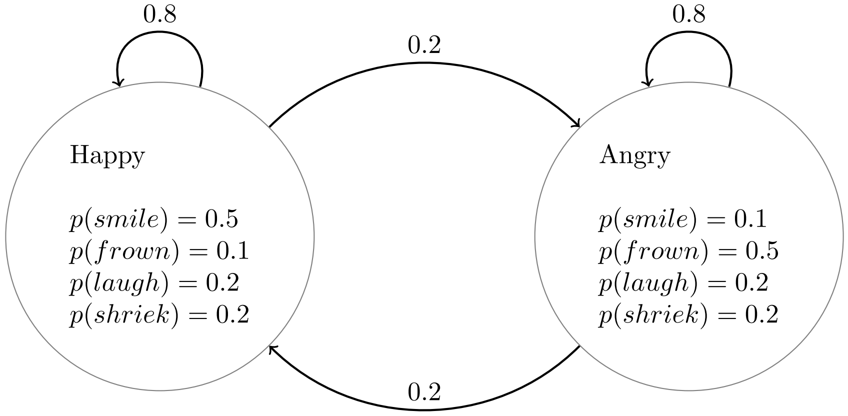
20. Recall that the solution of the simplex optimization problem

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

with $a_1, \cdots a_k \geq 0$ is $q_k^* = \frac{a_k}{\sum_{k'=1}^{K} a_{k'}}$. Derive the update of parameters ω_k, ν_l for the M-Step.

5 Hidden Markov Model (10 points)

Gollum lives a simple life. Some days he's Angry and other days he's Happy. But he hides his emotional state, so all you can observe is whether he smiles, frowns, laughs, or shrieks. Luckily, he is always happy on the first day. There is one transition per day.



Definitions:

X_t = state on day t

O_t = observation on day t

21. What is $P(X_2 = Happy)$?
22. What is $P(O_2 = frown)$?
23. What is $P(X_2 = Happy | O_2 = frown)$?
24. What is $P(O_{100} = shriek)$?
25. Assume that $O_1 = frown$, $O_2 = frown$, $O_3 = frown$, and $O_4 = frown$. What is the most likely sequence of states? *Hint: you do not need to run the Viterbi algorithm.*

6 Principal Component Analysis (10 points)

In this question, you are asked to perform PCA, using the dataset with 4 examples $\{\boldsymbol{x}_i\}_{i=1}^4 \in \mathbb{R}^2$ shown in Fig. 2. The 4 examples are

$$\begin{aligned}
\mathbf{x}_1 &= [3, 2]^{\mathrm{T}}, \\
\mathbf{x}_2 &= [-1, 2]^{\mathrm{T}}, \\
\mathbf{x}_3 &= [-1, 0]^{\mathrm{T}}, \\
\mathbf{x}_4 &= [3, 0]^{\mathrm{T}}.
\end{aligned}$$

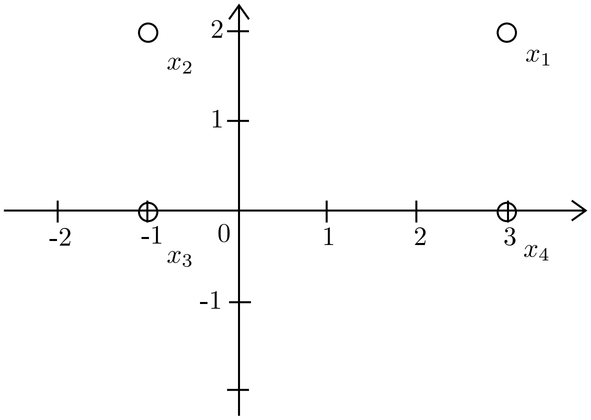


Figure 2: PCA dataset

26. Please finish the following steps to perform PCA to find the first principal component (PC1) $\boldsymbol{v}_1 \in R^2$ with the highest eigenvalue. Your answer should have ℓ_2 norm equal to 1 (i.e. $\|\boldsymbol{v}_1\|_2 = 1$).

Center the data	$\boldsymbol{\mu} =$
Covariance matrix	$C =$
Find PC1	$\boldsymbol{v}_1 =$

27. Now according to $\boldsymbol{v}_1$ you have computed, please find the coefficient on PC1 of a new data point $\boldsymbol{x}' = [2, 1]^{\mathrm{T}}$.

You may use this page as scratch paper