F243B8A7-5A9C-429A-B208-80ED7BDFCB82

csci567-fa19-exam2

#349 2 of 14

Q1 13

Multiple Choice, True or False

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

(a) True.
(b) False.

b 2. The AdaBoost algorithm will eventually reach zero training error regardless of the type of weak classifier it uses, when enough iterations are performed.

(a) True.
(b) False.

D 3. Which of the following statement is true?

(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

C 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?

(A) 1
(B) K
(C) N
(D) NK

C 5. How many learnable parameters are there in a GMM with K components and full covariance when the GMM is applied to a dataset of N points, each being D dimensional?

(A) $K(D + D(D-1))$
(B) $KD(D+1)$
(C) $K(N + D^2)$
(D) $KD + NK^2$

C 6. Which of the following models has a continuous latent variable?

(a) Naive Bayes Classifier
(b) Principal Component Analysis
(c) Gaussian Mixture Model
(d) Hidden Markov Model

a

7. Given the parameters of an HMM and an observation sequence O, we can determine the likelihood $P(O \mid \lambda)$. NOTE: λ represents the parameters of the HMM model.
   (a) True
   (b) False

B

8. Given an observation sequence O and the set of possible states in the HMM, we can learn the HMM parameters, including transition probability matrix and emission probabilities.
   (A) True
   (B) False

B

9. Which is TRUE about the Baum-Welsh algorithm?
   (A) It is used to find the real parameters of a hidden markov model.
   (B) It uses a forward-backward algorithm to maximize the probability of an observation.
   (C) The forward-backward algorithm can not do completely unsupervised learning of the transition matrix A and emission matrix B parameters.
   (D) It is a special case of the EM algorithm which is a recursive algorithm.

A

10. Which of the following statements is NOT TRUE about Lagrangian duality?
   (A) Optimal values of the primal and dual problems need to be equal.
   (B) Duality gives us an option of trying to solve our original (potentially nonconvex) constrained optimization problem in another way.
   (C) Duality allows us to formulate optimality conditions for constrained optimization problems.
   (D) One purpose of Lagrange duality is to find a lower bound on a minimization problem or an upper bounds for a maximization problem.

A

11. What do we mean by generalization error in terms of the SVM?
   (A) How far the hyperplane is from the support vectors
   (B) How accurately the SVM can predict outcomes for unseen data
   (C) The threshold amount of error in an SVM
   (D) None of the above.

A

12. Which of the following statements is NOT TRUE about the differences between SVM and logistic regression (LR)?
   (A) Both LR and SVM can give us an unconstrained, smooth objective.
   (B) SVMs have a nice dual form, giving sparse solutions when using the kernel trick.
   (C) They use different loss functions.
   (D) SVM is better at generalization since it doesn't penalize examples for which the correct decision is made with sufficient confidence.

40F326D6-0015-4C69-8E02-FBE9FEA855B4

csci567-fa19-exam2

#349 4 of 14

2 Mixture Models (25 points)

Consider a Poisson Mixture Model with the following probability mass function:

Q13 15

$$p(x_n) = \sum_{k=1}^{K} p(x_n, z = k) = \sum_{k=1}^{K} p(z = k)p(x_n \mid z = k) = \sum_{k=1}^{K} \omega_k \pi(x_n \mid \lambda_k) = \sum_{k=1}^{K} \omega_k \frac{\lambda_k^{x_n}}{x_n!} e^{-\lambda_k}$$

where ω_k is the mixture weight such that $\sum_{k=1}^{K} \omega_k = 1$, x_n is a non-negative integer, K is the number of mixtures, and $\lambda_k > 0$ is the parameter of a Poisson distribution.

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

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

where N is the number of data points.

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

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

Write out the Lagrangian and find the optimal ω_k (treating all other variables constant). **(15 points)**

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} (\gamma_{nk} \log p(x_n, z=k) - \gamma_{nk} \log \gamma_{nk}) - \sum_{i=1}^{K} \varepsilon_i w_i - \alpha\left(\sum_{k=1}^{K} w_k - 1\right)$$

s.t. $\varepsilon_i \geq 0, \forall i \in [k]$ and $\alpha \geq 0$

$$\boxed{\log p(x_n, z=k) = \log(w_k) + \log\left(\frac{\lambda_k^{x_n}}{x_n!} e^{-\lambda_k}\right)}$$

$$\nabla_{w_k} L = \sum_{n=1}^{N} \gamma_{nk} / w_k - \varepsilon_k - \alpha = 0$$

By complementary slackness, $\varepsilon_k = 0$,

$$\sum_n \gamma_{nk} / w_k - \alpha = 0$$

$$\frac{\sum_n \gamma_{nk}}{\alpha} = w_k \quad \overset{\text{given}}{\Rightarrow} \quad \sum_k w_k = 1, \ \sum_n \sum_k \gamma_{nk} = N$$

$$\alpha = N$$

therefore, $w_k^* = \frac{\sum_n \gamma_{nk}}{N}$

67ED66E0-EDA8-4124-9DF1-063DF41D5E55

csci567-fa19-exam2

#349 6 of 14

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

$$L = \sum^{N}\sum^{K} (\gamma_{nk} \cdot (\log(W_k) + \log(\lambda_k^{x_n}) - \log(x_n!) + \log(e^{-\lambda_k})) - \sum^{K} \alpha_i \lambda_i$$

Q14 4

$\alpha_i \geq 0 \quad \forall i \in [K]$

$$\nabla_{\lambda_k} = \sum^{N} \gamma_{nk} \cdot \left(\frac{x_n}{\lambda_k} - 1 \right) - \alpha_k = 0$$

Complementary slackness, $\alpha_k = 0$.

$$\sum^{N} \gamma_{nk} \left(\frac{x_n}{\lambda_k} - 1 \right) = 0$$

$$\sum^{N} \gamma_{nk} \left(\frac{x_n}{\lambda_k} \right) = \sum^{N} \gamma_{nk}$$

wrong expansion 4

by hint, x_n should be constant

$$\left(\sum^{N} \gamma_{nk} \right) \cdot \frac{x_n}{\lambda_k} = \sum^{N} \gamma_{nk}$$

$$\lambda_k^* = x_n$$

FCD227DB-7198-4131-BA3B-12F19CA7FC91

csci567-fa19-exam2

#349 7 of 14

Q15 10

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:

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

A new test example x is labeled 1 if $\mathbf{w}^T\mathbf{x} \geq 1$, 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)**

$L = \frac{1}{2}\|W\|_2^2 - \sum_N \alpha_n (W^T x_n - 1) \quad \alpha_n \geq 0 \ \forall n \in N$

$\nabla_W L = W - \sum_N \alpha_n x_n = 0$

$\sum_N \alpha_n \cdot x_n = W$

replacing w with $\sum_N \alpha_k x_n$,

10, >= 0 instead of > 0

$\frac{1}{2}\sum_{m,n} \alpha_m \alpha_n x_m^T x_n - \sum_{m,n} \alpha_n \alpha_m x_n^T x_m + \sum_N \alpha_n$

$= \sum^N \alpha_n - \frac{1}{2}\sum_{m,n} \alpha_n \alpha_m x_n^T x_m \quad \Leftarrow$ dual

s.t. $\alpha_n > 0 \ \forall n \in [N]$

52336548-D8A3-4B20-BA60-5C1DE59A832F

csci567-fa19-exam2

#349 8 of 14

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

Q16 2

Yes, we can have $k(x_n, x_m) = x_n^T x_m$ for the dual.

$= \sum_i^D x_{ni} \cdot x_{mi}$

D is x dimention.

Q17 2

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

Yes.

Missing reasons 2

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 $n \in [N]$

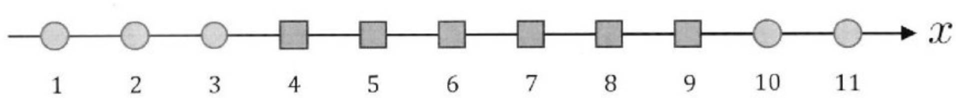

Figure 1: The 1-dimensional training set with 11 data. A square means the class of the data is $+1$, *i.e.* $y = +1$ and a circle means $y = -1$. The number under each data indicates its x coordinate.

Now we are given a training set of 11 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 $\log(\cdot)$.

3DC9E5CE-6F6D-4591-8375-1EA76F39FEF1

csci567-fa19-exam2

#349 10 of 14

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)**

Q18 5

h_1 $(s, b) \Rightarrow (1, 3.5)$

point 10 and point 11 will be misclassified. $\epsilon_1 = \frac{2}{11}$.

Q19 7

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)**

Q20 8

$$\beta_1 = \frac{1}{2}\log\frac{\frac{9}{11}}{\frac{2}{11}} = \frac{1}{2}\log\frac{9}{2}$$

$$D_2(x) = \begin{cases} \dfrac{\frac{1}{11}e^{\beta_1}}{\frac{2}{11}e^{\beta_1} + \frac{9}{11}e^{-\beta_1}} = \dfrac{1}{4} & x = 10 \text{ or } x = 11 \\ \dfrac{\frac{1}{11}e^{-\beta_1}}{\frac{2}{11}\cdot e^{\beta_1} + \frac{9}{11}e^{-\beta_1}} = \dfrac{1}{18} & \text{else.} \end{cases}$$

for h_2,

$(s, b) \rightarrow (-1, 9.5)$

$\epsilon_2 = 3 \times \frac{1}{18} = \frac{1}{6}$

point 1, 2, 3 are mis-classified.

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

$\beta_1 = \frac{1}{2}\log\frac{9}{2}$

$\beta_2 = \frac{1}{2}\log 5$

Accuracy $= \frac{11-3}{11} = \frac{8}{11}$

for $X_{1 \cdots 3}$,

$\text{sign}(H(x_{1\ldots3})) = -\frac{1}{2}\log\frac{9}{2} + \frac{1}{2}\log 5 \geq 0 \Rightarrow$ "+1" wrong

for $X_{4 \cdots 9}$,

$\text{sign}(H(x_{4\ldots9})) = \frac{1}{2}\log\frac{9}{2} + \frac{1}{2}\log 5 \geq 0 \Rightarrow$ "+1" correct

$\text{sign}(H(x_{10,11})) = \frac{1}{2}\log\frac{9}{2} - \frac{1}{2}\log 5 \leq 0 \Rightarrow$ "−1" correct.

10

5 HMM (15 points)

Recall a hidden Markov model is parameterized by

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

21. Given a sequence of observations $O_{1:t-1}, O_{t+1:T}$, O_t is missing for some reason. Compute the probability of O_t, which is $P(O_t = o | O_{1:t-1}, O_{t+1:T})$ in terms of forward message, backward message, transition probability, emission probability as needed. If $P(O_{1:t-1}, O_{t+1:T})$ appears in the denominator of your solution, you don't need to further express it in other forms. **(10 points)**

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

$$P(O_t = o \mid O_{1:t-1}, O_{t+1:T})$$

Q21 10

$$= \frac{P(O_t = o, O_{1:t-1}, O_{t+1:T})}{P(O_{1:t-1}, O_{t+1:T})}$$

$$= \frac{\sum_s P(O_t = o, X_t = s, O_{1:t-1}, O_{t+1:T})}{P(O_{1:t-1}, O_{t+1:T})}$$

$$= \frac{\sum_s P(O_t = o \mid X_t = s, O_{1:t-1}, O_{t+1:T}) \cdot P(O_{t+1:T} \mid X_t = s, O_{1:t-1}) \, P(X_t, O_{1:t-1})}{P(O_{1:t-1}, O_{t+1:T})}$$

$$= \frac{\sum_s P(O_t = o \mid X_t = s) \, P(O_{t+1:T} \mid X_t = s) \sum_{s'} P(X_t \mid X_{t-1} = s') P(X_{t-1}, O_{1:t-1})}{P(O_{1:t-1}, O_{t+1:T})}$$

$$= \frac{\sum_s b_{s,o_t} \beta_s(t) \cdot \sum_{s'} a_{s',s} \, \alpha_{s'}(t-1)}{P(O_{1:t-1}, O_{t+1:T})}$$

27E75ABB-CDEC-41DF-8E72-28613DA0DD3F

csci567-fa19-exam2

#349 12 of 14

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

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

$$a_{11} = 0, \quad a_{12} = 0, \quad a_{13} = 1$$
$$a_{21} = 1/2, \quad a_{22} = 1/2, \quad a_{23} = 0$$
$$a_{31} = 1/3, \quad a_{32} = 1/3, \quad a_{33} = 1/3$$
$$b_1(A) = 0, \quad b_1(B) = 1/2, \quad b_1(C) = 1/2$$
$$b_2(A) = 1/2, \quad b_2(B) = 0, \quad b_2(C) = 1/2$$
$$b_3(A) = 1/2, \quad b_3(B) = 1/2, \quad b_3(C) = 0$$

What is $P(X_3 = S_1)$? hint: you do not need to run Viterbi algorithm. **(5 points)**

$P(X_1 = S_3) = 1$

$P(X_2 = S_1) = P(X_2 = S_3) = \frac{1}{2}$ $\Rightarrow$ we know that $O_2 = B$, $X_2 \neq S_2$.

$P(X_3 = S_1) = \frac{1}{2} \times 0 + \frac{1}{2} \times \frac{1}{2}$

$= \frac{1}{4}$

Q22 1

12

You may use this page as scratch paper.

You may use this page as scratch paper.