41FD81B8-8583-4947-BA5E-86DF0E9F7F05

csci567-fa19-exam1

#141 2 of 16

Q1 7

1 Multiple Choice, True or False (30 points)

a 1. Given a dataset for binary classification, kNN with large k tends to have a smoother decision boundary (assuming $N >> k$). **(2 points)**

(a) True

(b) False

a 2. Because kNN is a very flexible non-parametric classifier, it can achieve near-perfect classification even for problems in which the true underlying data distributions overlap. **(2 points)**

(a) True

(b) False

b 3. Given a dataset which consists of a training set and a development set for tuning hyperparameter k. When we see that choosing a specific k results in very low training error but very high testing error, it is a good sign of underfitting. **(2 points)**

(a) True

(b) False

c 4. Which of the following penalty functions cannot be a good idea to regularize model complexity? **(3 points)**

(a) $R(\mathbf{w}) = \exp\{\sum_i |w_i|\}$

(b) $R(\mathbf{w}) = \exp\{-\sum_i |w_i|\}$

(c) $R(\mathbf{w}) = -\sum_i \log(|w_i|^{-1})$

(d) $R(\mathbf{w}) = \sum_i \exp\{|w_i|\}$

c 5. Suppose we are training a neural network with mini-batch SGD of batch size 50, and 50000 training samples. How many updates would there be while training during 5 epochs? **(3 points)**

(a) 50000

(b) 1000

(c) 5000

(d) 250000

$\frac{50000}{50}$

52AB07B8-A90B-4148-B7C5-F4F8087D9F36

csci567-fa19-exam1

#141 3 of 16

Q2 3

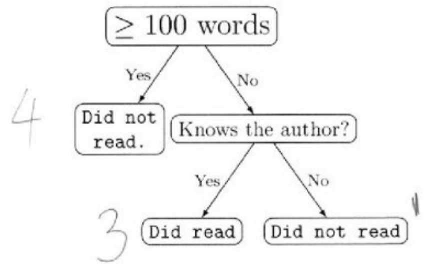
For questions 6 to 8 consider the following data and two decision trees below. We would like to build a decision tree classifier to answer the question "did Joe read the email ?"

Table 1: Did Joe read the email?

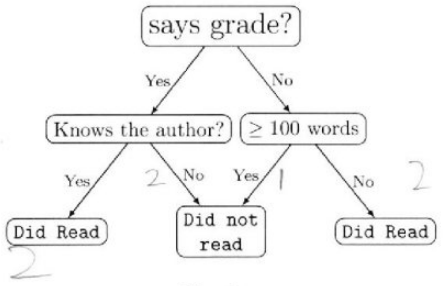
10

Knows the author?	≥ 100 words	Says research?	Says grade?	Says lottery	Did Joe read it?
no	no	yes	yes	no	no
yes	yes	no	yes	no	no
no	yes	yes	yes	yes	no
yes	yes	yes	yes	no	no
no	yes	no	no	no	no
yes	no	yes	yes	yes	yes
no	no	yes	no	no	yes
yes	no	no	no	no	yes
yes	no	yes	yes	no	yes
yes	yes	yes	yes	yes	no

Here are two decision trees that attempt to classify the emails:



Tree 1



Tree 2

b

6. Because decision trees learn to classify discrete valued outputs, it is impossible for them to overfit. **(3 points)**

(a) True

(b) False

C

7. How many points in the *training data* are correctly classified by tree 1? **(3 points)**

(a) 6

(b) 7

(c) 8

(d) 9

(e) 10

304B6C3E-F48A-4A8B-BA1A-BDFBA66347FB

csci567-fa19-exam1

#141 4 of 16

Q3 12

8. How many points in the *training data* are correctly classified by tree 2? **(3 points)**

   (a) 6

   (b) 7

   (c) 8

   (d) 9

   (e) 10

d

9. Suppose a convolution layer takes a $5 \times 7 \times 3$ image as input and outputs a $3 \times 4 \times 6$ tensor. Which of the following is a possible configuration of this layer? **(3 points)**

   (a) Three $2 \times 4 \times 3$ filters, stride 1, no zero-padding.

   (b) Three $3 \times 3 \times 3$ filters, stride 1, 1 zero-padding.

   (c) Six $3 \times 4 \times 3$ filters, stride 1, 1 zero-padding.

   (d) Six $3 \times 3 \times 3$ filters, stride 2, 1 zero-padding.

$(7-3) = 4$

$(9-3)/2 = 3$

d

10. For $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{2\times 1}$, which of the following bases $\phi(x)$ corresponds to the kernel defined as **(3 points)**

$$k(x, x') = e^{x_1 + x_1'} + e^{2(x_2 + x_2')}$$

   (a) $\phi(x) = [e^{x_1}, e^{x_2}]^T$

   (b) $\phi(x) = [e^{x_1}, \sqrt{2}e^{x_2}]^T$

   (c) $\phi(x) = [e^{x_1}, e^{\sqrt{2}x_2}]^T$

   (d) $\phi(x) = [e^{x_1}, e^{2x_2}]^T$

C

11. Consider the dataset consisting of points (x, y), given the basis function $\phi(x, y) = [x^2, 2xy, y^2]^T$, which of the following matrices is the kernel matrix of the three data points $(x_1, y_1) = (1, 0), (x_2, y_2) = (0, 1), (x_3, y_3) = (1, 1)$? **(3 points)**

   (a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 6 \end{bmatrix}$

   (b) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 6 \end{bmatrix}$

   (c) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 6 \end{bmatrix}$

   (d) $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \\ 2 & 1 & 6 \end{bmatrix}$

1 0 0 0 0 1

1 0 0 1 2 1

0 0 1 1 2 1

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 6 \end{bmatrix}$$

Q4 2

2 Linear Regression

Consider a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where each datapoint is associated with an importance weight $r_i > 0$, $i \in \{1, \cdots, n\}$, then the Weighted Residual Sum of Squares (WRSS) is defined as:

$w^T \to (1, d)$

$$\text{WRSS}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n} r_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \tag{1}$$

12. Define $\mathbf{X}$ as the matrix whose i-th row is $\mathbf{x}_i^T$, and $\mathbf{R}$ as a diagonal matrix where $\mathbf{R}_{ii} = r_i$ and 0 for all other entries. Write down the matrix form of the WRSS objective (Eq. 1). **(3 points)**

$$WRSS(w) = \frac{1}{2}\sum_{i=1}^{n} R_{ii}(y_i - w^T X_i)^2$$

$$= \frac{1}{2} R(w^T X - y)^2$$

0

13. Solve for the optimal $\mathbf{w}^*$ for WRSS in the matrix form. **(7 points)**

$\nabla_{w^*} WRSS(w^*) = R X^T(w^{*T} X - y) = 0$, $r_i > 0$, X is input

then $w^{*T} x = y \Rightarrow w^* = (X^T X)^{-1} X^T y$

2

Q5 3

3 Naive Bayes (15 points)

Naive Bayes classifiers are a family of simple "probabilistic" classifiers based on applying the Bayes' theorem with conditional independence assumptions between the features. In this problem, we will use a naive Bayes classifier with features $\{f_i\}$, where $i = 1, \cdots, d$, and label y , which is defined as follow:

$$\text{argmax}_y\, P(y|f_1, \cdots, f_d) = \text{argmax}_y\, P(y) \prod_{i=1}^{d} P(f_i|y)$$

Additionally, a linear classifier (for example, perceptron) can be defined as follow:

$$\text{argmax}_y \sum_{i=0}^{d} w_{y,i} \cdot f_i \quad \text{where } f_0 \text{ is a bias feature that is always 1 for all data}$$

14. For a naive Bayes classifier with binary-valued features, i.e. $f_i \in \{0, 1\}$, for $i = 0, \cdots, d$, *prove that the naive Bayes classifier is also a linear classifier* by defining weights $w_{y,i}$, for $i = 0, \cdots, d$, such that both classifiers above are equivalent. The weights should be expressed in terms of the naive Bayes probabilities: $P(y), P(f_i = 1|y)$, and $P(f_i = 0|y)$, $i = 1, \cdots, d$. You can assume that all these probabilities are non-zero. **(12 points)**

Hint: Using the log operation to convert products to summations, i.e. $\log \prod_{i=1}^{d} P(f_i|y) = \sum_{i=1}^{d} \log P(f_i|y)$

$\underset{y}{\text{argmax}}\, P(y) \prod_{i=1}^{d} P(f_i|y)$ is same as $\underset{y}{\text{argmax}} \log P(y) + \sum_{i=1}^{d} \log P(f_i|y)$

$\underset{y}{\text{argmax}} \sum_{i=1}^{d} \log(P(f_i|y) \cdot P(y))$

$W_{y,i} = \max(P(f_i=1|y) \cdot P(y),\ P(f_i=0|y) \cdot P(y))$

This page intentionally left blank.

Q7 3

15. If we are given the dataset as in Table 2, and f_1, f_2 are both binary features. Can we use a naive Bayes classifier to correctly classify all four data? **(3 points)**

f_1	f_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Table 2: Four data and their labels.

Describe the classifier to prove if your answer is yes, or briefly justify to disprove if your answer is no. (you answer should be within 3 lines)

$$\underset{y}{\arg\max}\ P(y) \cdot \prod_{i=1}^{d} P(f_i \mid y)$$

$P(f_1=0 \mid y=0) = \frac{1}{2}$

$P(f_1=1 \mid y=1) = \frac{1}{2}$

$P(f_2=0 \mid y=0) = \frac{1}{2}$

$P(f_2=1 \mid y=1) = \frac{1}{2}$

$P(y=0) = \frac{1}{2}$

$P(y=1) = \frac{1}{2}$

✓

$\Longrightarrow$ No, naive Bayes is a linear classifier, the relation is XOR, it cannot work on XOR.

Q8 4

4 Logistic Regression (25 points)

Given a dataset $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_n)\}$ where $\boldsymbol{x}_n \in \mathbb{R}^D$, and $y_n \in \{0, 1\}$. Consider this prediction model

$$P(y_n = 1|\boldsymbol{x}_n; \boldsymbol{w}) = \Phi(\boldsymbol{w}^T\boldsymbol{x}_n),$$

where

$$\Phi(z) = \int_{-\infty}^{z} \phi(z)dz,$$

$$e^{-\frac{z^2}{2}} = 7\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \cdot (-z) +$$

and

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp{(-\frac{z^2}{2})}.$$

The shape of $\Phi(z)$ is very similar to the sigmoid activation function we used in logistic regression. Because $\Phi(z)$ is called the *probit function*, we thus call this model *Probit Regression*.

16. The cross-entropy loss of a binary classifier over a dataset is defined as follows:

$$H(y, p) := -\sum_{(\boldsymbol{x}_n, y_n) \in \mathcal{D}} (y_n \ln p_n + (1 - y_n) \ln(1 - p_n)),$$

where $p_n = P(y_n = 1|\boldsymbol{x}_n; \boldsymbol{w})$. Our goal is to minimize cross-entropy in our binary classification problem. Please derive $\nabla_{\boldsymbol{w}} H(y, P(y|\boldsymbol{x}, \boldsymbol{w}))$, express it with $y_n, \boldsymbol{x}_n, \Phi(\cdot)$ and $\phi(\cdot)$, and reduce it to the simplest form. **(15 points)**

$$\nabla_w H(y, P(y|x, w)) = -\sum -y_n \cdot \frac{1}{\Phi(w^T x)} \cdot \phi^2(w^T x_n) \cdot (w^T x_n) \cdot x_n + (1 - y_n) \cdot \frac{1}{1 - \Phi(w^T x_n)} \cdot \phi^2(w^T x_n)(w^T x_n) \cdot x_n$$

$$\nabla_w \phi(w^T x_n) = \phi(w^T x_n) \cdot (-(w^T x_n) \cdot x_n)$$

basic derivatives 4

$$\nabla_w H = \sum y_n \cdot \frac{1}{\Phi(w^T x_n)} \cdot \phi^2(w^T x_n) \cdot (w^T x_n) \cdot x_n - (1 - y_n) \cdot \frac{1}{1 - \Phi(w^T x_n)} \cdot \phi^2(w^T x_n) \cdot (w^T x_n) \cdot x_n$$

$$= \sum_{\in D} (2y_n - 1) \cdot \frac{\phi^2(w^T x_n)}{\Phi(w^T x_n)} (w^T x_n) \cdot x_n$$

This page intentionally left blank.

Q10 4

17. The activation function that the prediction model uses determines how much the model is *sensitive* to outliers. Outliers are non-typical data points that deviates far away from typical ones with the same label. For example, if the data points $\boldsymbol{x}_n \in \mathcal{R}$ with label $y_n = 1$ are mostly within the range $[2, 4]$, then a data point with value 10 is considered an outlier.

Suppose that we add an outlier $\boldsymbol{x}_{N+1}$ to the dataset $\mathcal{D}$. Please derive $\nabla_{\boldsymbol{x}_{N+1}} H(y, P(y|\boldsymbol{x}, \boldsymbol{w}))$ for Probit Regression. **(10 points)**

Note that the cross-entropy is now a summation over $N+1$ points, i.e.,

$$H(y,p) := -\sum_{n=1}^{N+1} (y_n \ln p_n + (1-y_n)\ln(1-p_n)),$$

$$\nabla_{x_{N+1}} H(y, P(y|x,w)) = -\sum (-y_n \cdot \frac{1}{\Phi(w^T x_n)} \cdot \phi^2(w^T x_n) \cdot (w^T x_n) \cdot w + (1-y_n) \cdot \frac{1}{1-\Phi(w^T x_n)} \cdot \phi^2(w^T x_n) \cdot (w^T x_n) \cdot w)$$

$$= \sum_{n=1}^{N+1} (y_n \cdot \frac{\phi^2(w^T x_n)}{\Phi(w^T x_n)} \cdot (w^T x_n) \cdot w - (1-y_n) \cdot \frac{\phi^2(w^T x_n)}{1-\Phi(w^T x_n)} \cdot (w^T x_n) \cdot w)$$

-2 for not removing summation

-2 for having extra terms

-2 for not simplifying (check rubric)

Q11 4

Neural Network (20 points)

Consider the following neural network, LeNet-5, that consists of two convolution layers (rectangles with 'conv'), two average-pooling layers and three fully connected layers (right most three rectangles). The neural net takes image of size (32 × 32 × 1) and outputs a prediction vector of probabilities for 10 classes.

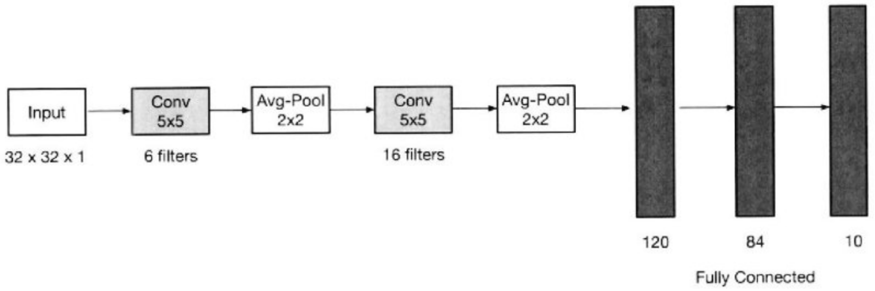


Figure 1: LeNet-5

18. For all convolutional layers, padding is zero, stride is 1. There is no bias in all layers. How many parameters do we need to learn for this network? (Your answer should tell your calculation process step by step). **(10 points)**

$6 \times (5 \times 5 \times 1)$ ✓ $+ 16 \times$ ✓ $(5 \times 5 \times 6)$

$= 6 \times 25 + 16 \times 25 \times 6$

$= 2550$

Q12 9

Suppose we have a binary-class Convolutional Neural Network defined below.

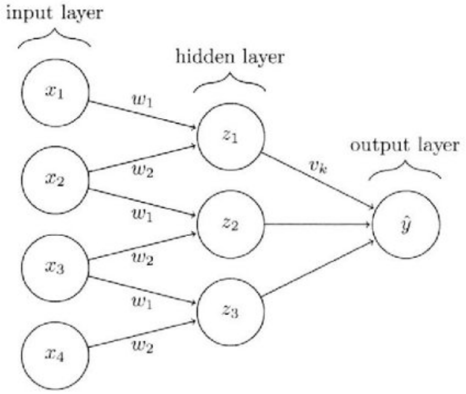


Figure 2: A neural network with one hidden layer.

For binary classification, the forward propagation can be expressed as:

input layer	x_i	(2)
hidden layer	$z_k = ReLU(w_1 x_k + w_2 x_{k+1})$, where $ReLU(x) = max\{0, x\}$	(3)
output layer	$\hat{y} = \sigma(\sum_{k=1}^{3} v_k z_k)$, where $\sigma(z) = \frac{1}{1+\exp(-z)}$	(4)
loss function	$L(y, \hat{y}) = -[y \log \hat{y} + (1-y)\log(1-\hat{y})]$, where $\hat{y}$ is prediction, y is ground truth	(5)

19. Please write down $\frac{\partial L}{\partial v_k}$ and $\frac{\partial L}{\partial w_1}$ in terms of only x_k, z_k, v_k, y, and/or $\hat{y}$ using backpropagation. **(10 points)**

wrong dL/dy -1

Hint: the derivative of the ReLU function is $H(a)$ = $\mathbb{I}[a > 0]$. You can directly use $H(a)$ in your answer.

$$\frac{\partial L}{\partial V_k} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial V_k} = \left(-\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right) \cdot 6\left(\sum_{k=1}^{3} V_k Z_k\right) \cdot \left(1 - 6\left(\sum_{k=1}^{3} V_k Z_k\right)\right) \cdot Z_k$$

$$\frac{\partial L}{\partial \hat{y}} = -\left[\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right]$$

$$\frac{\partial \hat{y}}{\partial V_k} = 6\left(\sum_{1}^{3} V_k Z_k\right) \cdot \left(1 - 6\left(\sum_{1}^{3} V_k Z_k\right)\right) \cdot Z_k$$

$$\frac{\partial L}{\partial W_1} = \sum_{K=1}^{3} \frac{\partial L}{Z_K} \cdot \frac{\partial Z_K}{\partial W_1} = \sum_{K=1}^{3} \left(-\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right) \cdot 6\left(\sum_1^3 V_K Z_K\right) \cdot \left(1-6\left(\sum_1^3 V_K Z_K\right)\right) \cdot V_K \cdot H(W_1 X_K + W_2 X_{K+1}) \cdot X_K$$

$$\frac{\partial Z_K}{\partial W_1} = H(W_1 X_K + W_2 X_{K+1}) \cdot X_K$$

$$\frac{\partial L}{Z_K} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial Z_K} = \left(-\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right) \cdot 6\left(\sum_1^3 V_K Z_K\right) \cdot \left(1-6\left(\sum_1^3 V_K Z_K\right)\right) \cdot V_K$$

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

9947B9A4-525F-4681-B515-904AF6C7EDB4

csci567-fa19-exam1

#141 16 of 16