CX4240 Computing for Data Analysis - Midterm Exam

Instructor: Bo Dai

Time Limit: 75 Minutes

Please write down your name and GT-ID on every page.

Name: _____ **GT-ID:** _____

Please read the following instructions carefully.

- The exam consists of four problems, **each worth 25 points**.
- This is a **closed-book exam**. No external resources or communication with others is allowed.
- You are allowed to bring **one double-sided US-letter-size cheatsheet**.
- By submitting this exam, you confirm that you have upheld the Georgia Tech Honor Code.

Question	Full Points	Points Earned
Q1	25	
Q2	25	
Q3	25	
Q4	25	
Total	100	

Name: _____ **GT-ID:** _____

1 Linear Regression [25pt]

A bike sharing company wants to predict the number of bikes rented each afternoon. For each day, the company records the following information:

- `temperature`: The daily temperature
- `tourists`: Number of tourists in the area
- `weekend`: A binary variable where 1 indicates weekend and 0 indicates weekday
- `rentals`: Number of bikes rented that afternoon

The company plans to use **linear regression** to predict the number of bike rentals. In this problem, the dependent variable is

$$y = \text{rentals}.$$

(a) [5pt] List all independent variables used to predict the outcome.

Your answer:

(b) [5pt] Why is linear regression more appropriate than logistic regression for this problem?

Your answer:

(c) [10pt] Please answer True or False. *No explanation is required.*

1. Linear regression predicts continuous numerical values.

Your answer:

2. In linear regression, the model assumes the noise term follows a Gaussian distribution.

Your answer:

3. Least squares regression minimizes the sum of squared prediction errors.

Your answer:

4. Increasing the degree in polynomial regression can sometimes lead to overfitting.

Your answer:

5. In the model $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$, θ_0 represents the intercept.

Your answer:

(d) [5pt] Suppose the learned linear regression model is

$$\hat{y} = \theta_0 + \theta_1(\text{temperature}) + \theta_2(\text{tourists}) + \theta_3(\text{weekend}),$$

where

$$\theta_0 = 1, \qquad \theta_1 = 0.2, \qquad \theta_2 = 0.5, \qquad \theta_3 = 2.$$

For a day with temperature = 25, tourists = 3, and weekend = 1, compute the predicted value $\hat{y}$.

Your answer:

Name: _____ **GT-ID:** _____

2 Naive Bayes [25pt]

About 2/3 of your email is spam, so you downloaded an open-source spam filter based on word occurrences that uses the Naive Bayes classifier. Assume you collected the following regular and spam mails to train the classifier, and only three words are informative for this classification. Each email is represented as a 3-dimensional binary vector whose components indicate whether the respective word is contained in the email.

`study`	`free`	`money`	Category
1	0	0	Regular
0	0	1	Regular
1	0	0	Regular
1	1	0	Regular
1	0	0	Spam
1	0	0	Spam
0	1	0	Spam
0	1	0	Spam
0	1	1	Spam
0	1	1	Spam
0	1	1	Spam
0	1	1	Spam

(a) [5pt] The spam filter uses a prior $P(\text{spam}) = 0.1$. Explain in one sentence why this might be reasonable.

Your answer:

(b) [10pt] Using the training data above, compute the following model parameters:

$P(\text{study}|\text{spam})$, $P(\text{study}|\text{regular})$, $P(\text{free}|\text{spam})$, $P(\text{free}|\text{regular})$, $P(\text{money}|\text{spam})$, $P(\text{money}|\text{regular})$.

Note: $P(study|spam)$ stands for $P(study = 1|spam)$, similar for the others.

Your answer:

(c) [10pt] Based on the prior distribution in part (a) and the conditional probabilities in part (b), compute the probability $P(\text{spam} \mid s)$ that the sentence

$$s = \textit{money for psychology study}$$

is spam.

Your answer:

Name: _____ **GT-ID:** _____

3 K-Means Clustering [25pt]

Notation:

- μ_k: The centroid (mean) of cluster k.
- r_{nk}: Binary indicator variable (hard assignment). $r_{nk} = 1$ if data point $\mathbf{x}_n$ is assigned to cluster k, and $r_{nk} = 0$ otherwise.

You are running K-Means to estimate the centroids μ_k. The algorithm iteratively performs the assignment step and the update step:

- **Assignment step:** Assigns each data point to the nearest cluster centroid based on the Euclidean distance $\boldsymbol{\mu}_k$:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

- **Update step:** Updates the centroids μ_k using the current assignments:

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

Suppose in the current iteration, you have three data points: $\mathbf{x}_1 = \begin{pmatrix}1\\3\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}5\\5\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}3\\1\end{pmatrix}$. There are two clusters in total, and their centroids are $\boldsymbol{\mu}_1 = \begin{pmatrix}1\\1\end{pmatrix}$ and $\boldsymbol{\mu}_2 = \begin{pmatrix}4\\4\end{pmatrix}$.

Questions :

(a) [8pt] Perform the assignment step, i.e., assign each data point to the nearest cluster and write down the assignments r_{nk}.

Your answer:

(b) [8pt] Perform the update step, i.e., calculate the updated centroid of the clusters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Your answer:

(c) [9pt] True or False (No explanations needed):

- K-Means can be understood as Gaussian Mixture Models with hard cluster assignment.

  Your answer:

- The K-Means algorithm can converge to difference results under different initializations.

  Your answer:

- Apart from the Euclidean distance, $D(\boldsymbol{x}, \boldsymbol{\mu}) = \sum_{i=1}^{d} |x_i + \mu_i|$ is also a proper distance metric for K-Mean algorithm.

  Your answer:

Name: _____ **GT-ID:** _____

4 Machine Learning Pipeline [25pt]

An online learning platform is developing several machine learning models to better support students. The platform collects the following types of data:

- Student profile features (such as age, major, and average homework score)
- Images of handwritten math answers
- Word sequences from forum posts

Answer the following questions:

(a) [4pt] A standard machine learning pipeline can be summarized as

$$(1) \rightarrow (2) \text{ Learning Algorithm} \rightarrow (3).$$

Fill in the missing components (1) and (3).

Your answer:

(b) [5pt]

1. What algorithm is commonly used to update the parameters of a neural network during training? [2 pt]

   Your answer:

2. What is the role of backpropagation in training a neural network? [3 pt]

Your answer:

(c) [6pt] For each of the following scenarios, choose the most suitable architecture from MLP, CNN, and RNN, and briefly justify your choice.

1. The input is a fixed-length feature vector containing a student's age, major, and average homework score. [2 pt]

Your answer:

2. The input is a grayscale image of a handwritten digit from a student's answer sheet. [2 pt]

Your answer:

3. The input is a sequence of words from a student's forum post, in the original order. [2 pt]

Your answer:

(d) [5pt] A handwritten-answer image has size $28 \times 28 \times 1$. We apply a convolution layer with

$$F = 3, \qquad P = 0, \qquad S = 1,$$

where F denotes the filter size, P denotes the padding, and S denotes the stride.

Using the formula

$$W_{\text{out}} = \frac{W - F + 2P}{S} + 1,$$

compute the height and width of the output activation map.

Your answer:

(e) [5pt] Suppose we are building a classifier to detect "requests for help" in forum posts. Consider the two sequences below, which contain an identical set of tokens but differ in their temporal ordering:

Sequence A: "The dog bit the man." Sequence B: "The man bit the dog"

1. Why can word order matter in this task? [2 pt]

Your answer:

2. Why is an RNN generally more suitable than a plain MLP for this task? [3 pt]

Your answer:

Name: _____ **GT-ID:** _____

Solution for Problem 1

Solution:

(a) The independent variables are

temperature, tourists, weekend.

(b) Linear regression is more appropriate because the target variable, the number of rentals, is a continuous numerical quantity, whereas logistic regression is used for binary classification.

(c)

1. True
2. True
3. True
4. True
5. True

(d) Substituting the values into the model,

$$\hat{y} = 1 + 0.2(25) + 0.5(3) + 2(1).$$

Thus,

$$\hat{y} = 1 + 5 + 1.5 + 2 = 9.5.$$

Therefore,

$$\boxed{\hat{y} = 9.5}.$$

Solution for Problem 2

Solution:

(a) A prior such as $P(\text{spam}) = 0.1$ may be reasonable because falsely labeling important regular emails as spam can be more costly than allowing some spam emails through.

(b) There are 8 spam emails and 4 regular emails.
For spam:

$$P(\text{study}|\text{spam}) = \frac{2}{8} = \frac{1}{4}, \qquad P(\text{free}|\text{spam}) = \frac{6}{8} = \frac{3}{4}, \qquad P(\text{money}|\text{spam}) = \frac{4}{8} = \frac{1}{2}.$$

For regular:

$$P(\text{study}|\text{regular}) = \frac{3}{4} = \frac{3}{4}, \qquad P(\text{free}|\text{regular}) = \frac{1}{4}, \qquad P(\text{money}|\text{regular}) = \frac{1}{4}.$$

(c) The sentence $s =$ *money for psychology study* contains `study`=1, `free`=0, `money`=1. Thus,

$$P(s|\text{spam}) = P(\text{study}|\text{spam})\,(1 - P(\text{free}|\text{spam}))\,P(\text{money}|\text{spam})$$

$$= \frac{1}{4} \cdot \left(1 - \frac{3}{4}\right) \cdot \frac{1}{2} = \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{32}.$$

Including the prior,

$$P(\text{spam})P(s|\text{spam}) = 0.1 \cdot \frac{1}{32} = \frac{1}{320}.$$

Similarly,

$$P(s|\text{regular}) = P(\text{study}|\text{regular})\,(1 - P(\text{free}|\text{regular}))\,P(\text{money}|\text{regular})$$

$$= \frac{3}{4} \cdot \left(1 - \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{9}{64}.$$

Including the prior,

$$P(\text{regular})P(s|\text{regular}) = 0.9 \cdot \frac{9}{64} = \frac{81}{640}.$$

Therefore,

$$P(\text{spam} \mid s) = \frac{\frac{1}{320}}{\frac{1}{320} + \frac{81}{640}} = \frac{2}{83}.$$

Solution for Problem 3

Solution:

(a)

$$\text{For } \mathbf{x}_1 : \|\mathbf{x}_1 - \boldsymbol{\mu}_1\|^2 = 4, \|\mathbf{x}_1 - \boldsymbol{\mu}_2\|^2 = 10 \implies r_{11} = 1, r_{12} = 0$$
$$\text{For } \mathbf{x}_2 : \|\mathbf{x}_2 - \boldsymbol{\mu}_1\|^2 = 32, \|\mathbf{x}_2 - \boldsymbol{\mu}_2\|^2 = 2 \implies r_{21} = 0, r_{22} = 1$$
$$\text{For } \mathbf{x}_3 : \|\mathbf{x}_3 - \boldsymbol{\mu}_1\|^2 = 4, \|\mathbf{x}_3 - \boldsymbol{\mu}_2\|^2 = 10 \implies r_{31} = 1, r_{32} = 0$$

(b)

$$\boldsymbol{\mu}_1 = \frac{\mathbf{x}_1 + \mathbf{x}_3}{2} = \frac{1}{2}\left[\begin{pmatrix}1\\3\end{pmatrix} + \begin{pmatrix}3\\1\end{pmatrix}\right] = \begin{pmatrix}2\\2\end{pmatrix}$$
$$\boldsymbol{\mu}_2 = \frac{\mathbf{x}_2}{1} = \begin{pmatrix}5\\5\end{pmatrix}$$

(c) 1. True
2. True
3. False

Solution for Problem 4

Solution:

(a) A standard pipeline is

Training Data → Learning Algorithm → Target Function: Predictor/Classifier/Representation..

(b)

1. A common algorithm is gradient descent, often stochastic gradient descent.
2. Backpropagation computes the gradients of the loss with respect to the network parameters so that the parameters can be updated during training.

(c)

1. **MLP**. The input is a fixed-length feature vector, which is well suited for a multilayer perceptron.
2. **CNN**. A convolutional neural network is appropriate for images because it captures spatial and local patterns.
3. **RNN**. A recurrent neural network is appropriate for sequential data because it models word order.

(d) Using

$$W_{\text{out}} = \frac{W - F + 2P}{S} + 1$$

with $W = 28$, $F = 3$, $P = 0$, and $S = 1$, we get

$$W_{\text{out}} = \frac{28 - 3 + 0}{1} + 1 = 25 + 1 = 26.$$

So the output activation map has height 26 and width 26, i.e.

$$\boxed{26 \times 26}.$$

(e)

1. Word order can affect the meaning of a sentence, so changing the order can change the classification result.
2. An RNN is more suitable because it processes the input sequentially and can model dependencies based on word order, whereas a plain MLP does not naturally capture sequence order.