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 Logistic Regression [25pt]

A food delivery company wants to predict whether a customer will order dinner using their mobile app. For each customer, we collect two features:

x_1 = distance to the nearest restaurant (km), x_2 = number of previous orders this month.

Let the label be

$$y = \begin{cases} 1, & \text{if the customer places an order,} \\ 0, & \text{otherwise.} \end{cases}$$

We model the probability using logistic regression:

$$P(y = 1 \mid x; \theta) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2) = \frac{1}{1 + \exp\left(-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)\right)}.$$

(a) [5pt] What is the range of the sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Your answer:

(b) [5pt] In logistic regression, what does the value

$$P(y = 1 \mid x; \theta)$$

represent?

Your answer:

Name: ____________________ **GT-ID:** ____________________

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

1. Logistic regression is commonly used for binary classification problems. [2 pt]

Your answer:

2. The sigmoid function can output negative values. [2 pt]

Your answer:

3. Logistic regression produces a linear decision boundary in the feature space. [2 pt]

Your answer:

4. Maximum Likelihood Estimation (MLE) only depends on the training data, while MAP estimation also incorporates a prior on the parameters. [2 pt]

Your answer:

5. Logistic regression always has a closed-form solution similar to linear regression. [2 pt]

Your answer:

(d) [5pt] Suppose the learned parameters are

$$\theta_0 = -1, \qquad \theta_1 = -0.5, \qquad \theta_2 = 1.$$

For a customer with $x_1 = 2$ and $x_2 = 3$, compute

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2.$$

Then compute

$$P(y = 1 \mid x; \theta) = \sigma(z).$$

Finally give the predicted class if the classification threshold is 0.5.

Name: ____________________ **GT-ID:** ____________________

Your answer:

Name: ____________________ **GT-ID:** ____________________

2 Chain-Augmented Naive Bayes [25pt]

A clinic is using a machine learning model to detect a specific **Respiratory Infection (I)** based on two clinical observations: **High Fever (F)** and **Low Oxygen (O)**.

While a standard Naive Bayes assumes symptoms are independent given the infection, clinical data suggests that High Fever often causes physiological stress that leads to Low Oxygen. We model this as a **Chain-Augmented Naive Bayes** with the structure

$$P(I, F, O) = P(I)P(F|I)P(O|I, F),$$

i.e., F depends on I, while O depends on both F and I.

The clinic provides the following records for 10 patients:

Patient	**High Fever (F)**	**Low Oxygen (O)**	**Infection (I)**
1	Yes	Yes	Yes
2	Yes	Yes	Yes
3	Yes	No	Yes
4	Yes	Yes	Yes
5	No	No	Yes
6	Yes	Yes	No
7	No	Yes	No
8	No	No	No
9	No	No	No
10	No	No	No

(a) [5pt] Explain the difference between discriminative models and generative models. Based on this distinction, are Naive Bayes and Chain-Augmented Naive Bayes discriminative or generative models?

Your answer:

Name: ____________________ **GT-ID:** ____________________

(b) [10pt] Using the table above, calculate:

- $P(I = \text{Yes})$
- $P(F = \text{Yes} \mid I = \text{Yes})$
- $P(F = \text{Yes} \mid I = \text{No})$
- $P(O = \text{Yes} \mid F = \text{Yes}, I = \text{Yes})$
- $P(O = \text{Yes} \mid F = \text{Yes}, I = \text{No})$

Your answer:

(c) [10pt] A new patient arrives with **High Fever = Yes** and **Low Oxygen = No**. Calculate the probability that the patient is infected:

$$P(I = \text{Yes} \mid F = \text{Yes}, O = \text{No}).$$

Your answer:

Name: ____________________ GT-ID: ____________________

3 Gaussian Mixture Model (GMM) and EM Algorithm [25pt]

Notation:

- $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$: mixing coefficient, mean, and covariance of component k
- γ_{nk}: responsibility (posterior probability) of component k for data point $\mathbf{x}_n$

You are running the Expectation-Maximization (EM) algorithm for a Gaussian Mixture Model (GMM) to estimate the parameters $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$. The EM algorithm iteratively performs:

- **E-step**: estimates the responsibilities

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_{k'}, \Sigma_{k'})}$$

- **M-step**: updates the parameters

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}, \qquad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} \gamma_{nk}},$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top}{\sum_{n=1}^{N} \gamma_{nk}}.$$

Suppose in the current iteration, you have three data points:

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \qquad \mathbf{x}_2 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \qquad \mathbf{x}_3 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

After the E-step, the calculated responsibilities for component $k = 2$ are

$$\gamma_{12} = 0.5, \qquad \gamma_{22} = 0.4, \qquad \gamma_{32} = 0.1.$$

(a) [5pt] Write down the log-likelihood $\log p(\{\mathbf{x}_n\}_{n=1}^N)$ of a dataset $\{\mathbf{x}_n\}_{n=1}^N$ for a GMM with K components.

Your answer:

Name: ____________________ **GT-ID:** ____________________

(b) [6pt] Please answer True or False. *No explanation is required.*

1. The summation of the responsibilities over the samples $\sum_{n=1}^{N} \gamma_{nk}$ should equal 1. [3 pt]

Your answer:

2. The mixing coefficients π_k updated in the M-step must sum to 1 across all K components. [3 pt]

Your answer:

(c) [14pt] Calculate the updated mean and covariance matrix for component $k = 2$.

Your answer:

Name: ____________________ **GT-ID:** ____________________

4 Machine Learning Pipeline [25pt]

A startup is building an AI system to help an online bookstore understand user behavior and recommend books. The system collects the following types of data:

- User profile information (age, preferred genres, membership duration)
- Browsing history (books viewed, time spent on pages)
- Purchase history (whether the user bought a recommended book)

(a) [5pt] 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] The company wants to build a model that predicts whether a user will purchase a recommended book.

Two example features are:

$$x_1 = \text{number of books viewed in the past week}, \qquad x_2 = \text{average time spent on book pages (minutes)}.$$

The label is

$$y = \begin{cases} 1, & \text{if the user buys the recommended book,} \\ 0, & \text{otherwise.} \end{cases}$$

Hint. Think about whether the output is continuous or discrete, and recall models discussed in class for such prediction tasks.

1. What type of machine learning problem is this (classification, regression, representation, etc.)? [2 pt]

Your answer:

2. Name one model that could be used for this task. [3 pt]

Name: ____________________ **GT-ID:** ____________________

Your answer:

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

1. The input is a user's profile represented by a fixed-length feature vector, including age, membership duration, and average spending. [2 pt]

   Your answer:

2. The input is the image of a book cover, and the system wants to classify the book into genres such as mystery, romance, or science fiction. [2 pt]

   Your answer:

3. The input is the sequence of books a user clicked on during one browsing session, in chronological order. [2 pt]

   Your answer:

Name: ________________ **GT-ID:** ________________

(d) [**4pt**] Suppose we want to classify book-cover images. Why is a CNN usually more suitable than directly flattening the image into a vector and using a plain MLP?

Your answer:

(e) [**5pt**] A book-cover image has size $32 \times 32 \times 3$. We apply a convolution layer with

$$F = 5, \qquad P = 0, \qquad S = 1$$

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

Use the formula

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

to compute the height and width of the output activation map.

Your answer:

Name: ____________________ **GT-ID:** ____________________

Solution to Problem 1

Solution:

(a) The sigmoid function satisfies

$$0 < \sigma(z) < 1.$$

Thus, its range is

$$(0, 1).$$

(b) The quantity $P(y = 1 \mid x; \theta)$ represents the probability that the label is $y = 1$ given the input features x under the logistic regression model.

(c)

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

(d) First compute

$$z = -1 + (-0.5)(2) + 1(3) = 1.$$

Then

$$P(y = 1 \mid x; \theta) = \sigma(1) = \frac{1}{1 + e^{-1}}.$$

Since $\frac{1}{1+e^{-1}} > 0.5$, the predicted class is

$$\hat{y} = 1.$$

Solution to Problem 2

Solution:

(a) Discriminative models learn $P(y \mid x)$ directly, while generative models model the data generation process, such as $P(x, y)$ or $P(x \mid y)$ together with $P(y)$. Both Naive

Bayes and Chain-Augmented Naive Bayes are generative models.

(b) From the table:

$$P(I = \text{Yes}) = \frac{5}{10} = \frac{1}{2}.$$

Among the 5 infected patients, 4 have high fever, so

$$P(F = \text{Yes} \mid I = \text{Yes}) = \frac{4}{5}.$$

Among the 5 non-infected patients, 1 has high fever, so

$$P(F = \text{Yes} \mid I = \text{No}) = \frac{1}{5}.$$

Among the infected patients with $F = \text{Yes}$, 3 out of 4 have low oxygen, so

$$P(O = \text{Yes} \mid F = \text{Yes}, I = \text{Yes}) = \frac{3}{4}.$$

Among the non-infected patients with $F = \text{Yes}$, the only one has low oxygen, so

$$P(O = \text{Yes} \mid F = \text{Yes}, I = \text{No}) = 1.$$

(c) We compute the posterior up to proportionality:

$$P(I = \text{Yes})P(F = \text{Yes} \mid I = \text{Yes})P(O = \text{No} \mid F = \text{Yes}, I = \text{Yes})$$

$$= \frac{1}{2} \cdot \frac{4}{5} \cdot \left(1 - \frac{3}{4}\right) = \frac{1}{2} \cdot \frac{4}{5} \cdot \frac{1}{4} = \frac{1}{10}.$$

For $I = \text{No}$,

$$P(I = \text{No})P(F = \text{Yes} \mid I = \text{No})P(O = \text{No} \mid F = \text{Yes}, I = \text{No})$$

$$= \frac{1}{2} \cdot \frac{1}{5} \cdot (1 - 1) = 0.$$

Thus,

$$P(I = \text{Yes} \mid F = \text{Yes}, O = \text{No}) = 1.$$

Solution to Problem 3

Solution:

(a) The log-likelihood of the dataset is

$$\log p(\{\mathbf{x}_n\}_{n=1}^{N}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) \right).$$

(b)

1. False
2. True

(c) First compute the denominator:

$$\sum_{n=1}^{3} \gamma_{n2} = 0.5 + 0.4 + 0.1 = 1.$$

So the updated mean is

$$\boldsymbol{\mu}_2 = 0.5 \begin{pmatrix} 1 \\ 3 \end{pmatrix} + 0.4 \begin{pmatrix} 5 \\ 5 \end{pmatrix} + 0.1 \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 2.8 \\ 3.6 \end{pmatrix}.$$

Now compute deviations:

$$\mathbf{x}_1 - \boldsymbol{\mu}_2 = \begin{pmatrix} -1.8 \\ -0.6 \end{pmatrix}, \quad \mathbf{x}_2 - \boldsymbol{\mu}_2 = \begin{pmatrix} 2.2 \\ 1.4 \end{pmatrix}, \quad \mathbf{x}_3 - \boldsymbol{\mu}_2 = \begin{pmatrix} 0.2 \\ -2.6 \end{pmatrix}.$$

Thus

$$\Sigma_2 = 0.5 \begin{pmatrix} 3.24 & 1.08 \\ 1.08 & 0.36 \end{pmatrix} + 0.4 \begin{pmatrix} 4.84 & 3.08 \\ 3.08 & 1.96 \end{pmatrix} + 0.1 \begin{pmatrix} 0.04 & -0.52 \\ -0.52 & 6.76 \end{pmatrix}.$$

This gives

$$\Sigma_2 = \begin{pmatrix} 3.56 & 2.26 \\ 2.26 & 1.64 \end{pmatrix}.$$

Solution to Problem 4

Solution:

(a) A standard machine learning pipeline is

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

(b)

1. Binary classification.
2. Example answers include logistic regression and neural network.

(c)

1. **MLP**. The input is a fixed-length feature vector.
2. **CNN**. CNNs are suitable for images because they preserve spatial structure and extract local patterns.
3. **RNN**. RNNs are suitable for sequential data because order matters.

(d) A CNN is more suitable because it preserves the spatial structure of the image and learns local visual patterns through convolution filters. A plain MLP on a flattened image loses much of this spatial information.

(e) Using

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

with $W = 32$, $F = 5$, $P = 0$, and $S = 1$,

$$W_{\text{out}} = \frac{32 - 5 + 0}{1} + 1 = 27 + 1 = 28.$$

So the output activation map has height and width

$$28 \times 28.$$