CS4641 Machine Learning - 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 five problems, each worth 25 points. You are required to **complete four out of the five problems**. If you answer all five, only the top four highest-scoring problems will be counted.
- This is a **closed-book exam. No notes, external resources, or communication with others is allowed**.
- 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	
Q5	25	
Total (Top 4 Scores)	100	

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

1 Least Square Regression [25pt]

Consider the following training data with inputs (x_1, x_2) and output y:

x_1	x_2	y
0	0	0
0	1	1.5
1	0	2
1	1	2.5

Assume these points come from a linear model:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + (\text{noise}).$$

Use *least squares regression* to estimate the model $\vec{\theta} = [\theta_0, \theta_1, \theta_2]^\top$.
Hint I. For matrix inverse,

$$\begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 3/4 & -1/2 & -1/2 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}.$$

Hint II. The least square regression has the objective function as:

$$\min_\theta \|y - X\vec{\theta}\|_2^2. \tag{1}$$

Minimizing Eq. 1, we obtain the close-form estimation of $\vec{\theta}$ as

$$(X^\top X)\,\vec{\theta} = X^\top y \;\Rightarrow\; \vec{\theta}^* = (X^\top X)^{-1} X^\top y.$$

Your answer: From our input data, we have

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad y = \begin{pmatrix} 0 \\ 1.5 \\ 2 \\ 2.5 \end{pmatrix}.$$

Then, we have

$$X^\top X = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}, \quad X^\top y = \begin{pmatrix} 6.0 \\ 4.5 \\ 4.0 \end{pmatrix}.$$

Given the inverse of $X^\top X$:

$$(X^\top X)^{-1} = \begin{pmatrix} 3/4 & -1/2 & -1/2 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix},$$

we can compute the estimation $\vec{\theta}^*$ as

$$\begin{aligned}\vec{\theta}^* &= (X^\top X)^{-1} X^\top y \\ &= \begin{pmatrix} 3/4 & -1/2 & -1/2 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix} * \begin{pmatrix} 6 \\ 4.5 \\ 4 \end{pmatrix} \\ &= \begin{pmatrix} 0.25 \\ 1.5 \\ 1 \end{pmatrix}.\end{aligned}$$

2 Logistic Regression [25pt]

A city government is studying the factors influencing whether residents install rooftop solar panels to formulate more effective promotion policies. The government has collected data from 500 residents, including the following information for each individual:

- `income`: Annual income, measured in thousand dollars.
- `electricity_cost`: Average monthly electricity cost, measured in dollars.
- `env_awareness`: Environmental awareness, a binary variable where 1 indicates high environmental awareness, and 0 otherwise.
- `home_ownership`: Homeownership status, a binary variable where 1 means the resident owns a home, and 0 means they rent.
- `install_solar`: Solar panel installation decision, a binary outcome variable that takes the value 1 if the resident has installed solar panels and 0 otherwise.

The government aims to use **logistic regression** to predict whether a resident will install solar panels. In this context, the dependent variable y represents the installation decision:

$$y = \begin{cases} 1, & \text{if the resident has installed solar panels} \\ 0, & \text{otherwise} \end{cases}$$

(a) [3pt] List all independent variables used to predict the outcome.

Your answer: income, electricity_cost, env_awareness, home_ownership

(b) [3pt] Why is logistic regression more appropriate for this problem than linear regression?

Your answer: Logistic regression is more appropriate because the outcome is binary. It models probabilities between 0 and 1 using a sigmoid function. Linear regression assumes a continuous outcome and can predict values outside [0, 1], which are not valid probabilities.

(c) True/False [9pt] Please directly answer with True/False. *No justification is needed.*

1. (T/F) Maximum Likelihood estimation (MLE) only considers the data, while Maximum a posteriori (MAP) incorporates both the data and a prior distribution of parameters. [3pt]

   Your answer: True

2. (T/F) MAP estimation selects the optimal parameters by maximizing the likelihood function, whereas MLE selects the optimal parameters by maximizing the posterior probability. [3pt]

   Your answer: False

3. (T/F) A sigmoid function never outputs the value 0 or the value 1.[3pt]

   Your answer: True

(d) Calculation [10pt] Given the logistic regression model:

$$P(\texttt{install\_solar} = 1) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 \texttt{income} + \beta_2 \texttt{electricity\_cost} + \beta_3 \texttt{env\_awareness} + \beta_4 \texttt{home\_ownership})\right)}$$

Suppose the estimated coefficients are: $\beta_1 = 0.05$, $\beta_2 = -0.01$, $\beta_3 = 0.8$, $\beta_4 = 1.2$, and $\beta_0 = -4$. Given a resident with the following characteristics: $\texttt{income} = 60$ (thousand dollars), $\texttt{electricity\_cost} = 100$ (dollars), $\texttt{env\_awareness} = 1$, and $\texttt{home\_ownership} = 1$. Calculate the probability that this resident will install solar panels.

Your answer:

$$\begin{aligned} z &= -4 + (0.05 \times 60) + (-0.01 \times 100) + (0.8 \times 1) + (1.2 \times 1) \\ &= -4 + 3 - 1 + 0.8 + 1.2 \\ &= 0 \end{aligned}$$

$$P(\texttt{install\_solar} = 1) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{0}} = \frac{1}{1 + 1} = \frac{1}{2} = 0.5$$

3 Naive Bayes [25pt]

Background (Bayes' Theorem).

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

(a) Calculation [20pt] In a certain population, 1% of the individuals have a disease D. A medical test for this disease has the following characteristics:

- If a person has the disease D, the test is positive 90% of the time (i.e., $P(\text{Positive} \mid \text{Disease}) = 0.9$).
- If a person does not have the disease D, the test is negative 90% of the time (i.e., $P(\text{Negative} \mid \text{not Disease}) = 0.9$).

Suppose a randomly selected individual from this population is tested and the test result is positive. Using Bayes' Theorem, compute the probability that this individual actually has the disease, given that they tested positive (i.e., $P(\text{Disease} \mid \text{Positive})$).

Hint In this problem, the Bayes' Theorem is given by

$$P(\text{Disease} \mid \text{Positive}) = \frac{P(\text{Positive} \mid \text{Disease})\, P(\text{Disease})}{P(\text{Positive})}$$

Your answer:

$$\begin{aligned} P(Positive) &= P(Positive \mid Disease)P(Disease) + P(Positive \mid Healthy)P(Healthy) \\ &= 0.108 \end{aligned}$$

Using Bayes' theorem, the probability that an individual has the disease given a positive test result is:

$$P(Disease \mid Positive) = \frac{P(Positive \mid Disease)P(Disease)}{P(Positive)}$$

Thus, we have:

$$P(Disease \mid Positive) = \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} = \frac{1}{12}$$

(b) Log-Likelihood [5pt] Let the log-likelihood for a single observation be defined as

$$\ell(\theta; x) = \log p(x \mid \theta)$$

For a dataset with n independent observations, which statement **best describes** how the overall log-likelihood is constructed from these single-observation log-likelihoods? *No justification is needed.* [5pt]

A. The overall log-likelihood is the sum of the single-observation log-likelihoods.

B. The overall log-likelihood is the product of the single-observation log-likelihoods.

C. The overall log-likelihood is the average of the single-observation log-likelihoods.

D. The overall log-likelihood is the maximum of the single-observation log-likelihoods.

Your answer: A

4 GMM/K-Means [25pt]

GMM Background. A Gaussian Mixture Model (GMM) assumes data is generated from k Gaussian distributions. The EM algorithm iteratively:

- **E-Step**: Estimates responsibilities (probabilities of data points belonging to each component).
- **M-Step**: Updates component parameters using these responsibilities.

Notation and Definitions

- y_j^l: Responsibility of component l for data point x_j
- The Euclidean distance between two points (a_1, b_1) and (a_2, b_2) is given by:

$$\textbf{Distance} = \sqrt{(a_2 - a_1)^2 + (b_2 - b_1)^2}$$

(a) E-Step Responsibility [3pt] For a Gaussian Mixture Model (GMM) with responsibilities y_j^l (probability that data point x_j belongs to component l), which statement(s) **must be** true? *No justification is needed.*

A. For each data point x_j, $\sum_{l=1}^{k} y_j^l = 1$.

B. For each component l, $\sum_{j=1}^{N} y_j^l = 1$.

Your answer: A

(b) K-means: Cluster Assignment [6pt] Given:

- Centroid A: $(1, 2)$
- Centroid B: $(4, 6)$
- Point P: $(3, 4)$

Compute Euclidean distances (can be left in square root form) between (P, A) and (P, B) and assign P to the nearest cluster.

Your answer:

$$d_A = \sqrt{(3-1)^2 + (4-2)^2} = \sqrt{8}, \quad d_B = \sqrt{(3-4)^2 + (4-6)^2} = \sqrt{5}$$

Assignment: Cluster $\boxed{B}$.

(c) True/False [16pt] Please answer ”True” or ”False” for the following statements. *No justification is needed.*

1. (T/F) The EM algorithm always converges to the global minima. [4pt]

   Your answer: **False**. EM converges, but not necessarily to the global minima.

2. (T/F) The EM algorithm maximizes the Evidence Lower Bound (ELBO) of MLE. [4pt]

   Your answer: **True**

3. (T/F) K-means algorithm starts by randomly assigning points to clusters. [4pt]

   Your answer: **False**. K-means starts by initializing centroids.

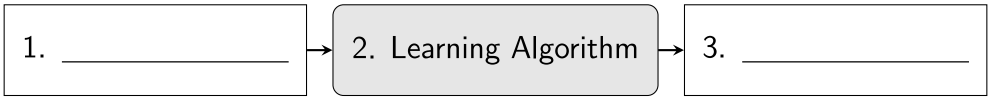
4. (T/F) K defines the number of neighbors in a cluster in K-means.[4pt]

   Your answer: **False**. K defines the number of clusters in K-means.

5 Machine Learning Pipeline

You are a machine learning engineer developing an **AI-powered medical diagnosis system** that integrates multiple types of patient data to assist doctors in detecting and predicting diseases. Your task is to **design the appropriate ML pipeline for each task and justify your choices**.

(a) General Pipeline [3pt]



The machine learning pipeline can be broken down into three key components, as illustrated in the figure above. Complete the first and third parts in the figure.
1. Training Data
3. Target Function: Predictor/Classifier/Representation...

(b) Case Study I [9pt] Your diagnosis system handles three tasks as follows:

- **Task 1 - Heart Rate Anomaly Detection.** Your diagnosis system is designed to detect heart rate anomalies from electrocardiogram (ECG) sequence data from a patient's heart activity. The data consists of time-series signals, where each sample is a sequence of voltage readings recorded at regular intervals.
- **Task 2 - Pneumonia Detection from X-ray Images.** Your system is built to detect pneumonia from chest X-ray images, where each sample consists of a grayscale medical image labeled as pneumonia or normal.
- **Task 3 - Blood Pressure Prediction.** Your system aims to predict a patient's blood pressure as continuous outputs based on structured health data, including age, weight, cholesterol level, and prior medical history.

For each task, answer the following:

1. What is the training data for each task?
2. What should the model predict (classification, regression, representation, etc.)?

Your answer:

- **Task 1**
  1. Electrocardiogram (ECG) sequence data
  2. Classification (if labeled data is available) or Representation (clustering)
- **Task 2**
  1. Chest X-ray images
  2. Binary classification

- **Task 3**
  1. Structured health data, including age, weight, cholesterol level, and prior medical history
  2. Regression (blood pressure as continuous outputs)

(c) Hyperparameter [3pt] We decide to adopt a convolutional neural network (CNN) as the backbone for the Pneumonia Detection task described in (b), list **three possible hyperparameters** to tune for the training procedure.

Your answer:
Example answers:

- **Learning Rate**: Determines the step size for updating weights during gradient descent.
- **Number of Convolutional Filters**: Controls the number of feature detectors in each convolutional layer.
- **Batch Size**: Defines the number of training samples processed in one iteration.
- **Kernel Size**: Defines the receptive field of convolutional filters.
- **Number of Layers / Depth**: Determines the complexity of the model.
- **Dropout Rate**: Specifies the proportion of neurons randomly dropped during training to prevent overfitting.
- **Weight Initialization**: Affects the starting point of training.
- **Activation Function**: Defines the non-linearity introduced at each layer.
- **Optimizer Choice**: Determines how model parameters are updated.
- **L2 Regularization (Weight Decay)**: Penalizes large weights to reduce overfitting and improve generalization.

(d) Case Study II [10pt] The Pneumonia Detection task can be seen as the binary classification task of classifying images as normal vs. pneumonia. Your colleague designed a CNN with a single output neuron. Let the output of this neuron be z. The final output of your network, y is given by:

$$y = \texttt{sigmoid}(\texttt{ReLU}(z)) \qquad (2)$$

Then, your colleague classifies all inputs with a $y \geq 0.5$ as pneumonia. What problem is this algorithm going to encounter?

Hint. The sigmoid activation function is defined as: $\sigma(x) = \frac{1}{1+e^{-x}}$, while the ReLU activation function is defined as $f(x) = \max(0, x)$. Think about how a classifier determines its prediction and what the outcome will be in this setting.

Your answer: Using ReLU then sigmoid will cause all predictions to be positive, i.e.,

$$y = \sigma(\text{ReLU}(z)) \geq 0.5,\ \forall z.$$