AUS | الجامعة الأميركية في الشارقة
American University of Sharjah

American University of Sharjah
School of Engineering
Department of Computer Engineering
P. O. Box 26666
Sharjah, UAE

Instructor: Alex Aklson
Office: ESB-2172
Phone: 971-6-515 4893
E-mail: aaklson@aus.edu
Semester: Fall 2024

MLR503 – Data Mining and Knowledge Discovery

Make-up Midterm Exam

November 7, 2024 (2 hours)

Student Name: lujain khalil **Student ID:** _____

Instructions:

- Write both your *name and ID* above.
- Read the questions carefully; write your answers clearly in the space provided.
- This is a **closed book and closed notes exam**. **Only use of calculator is permitted.**

For Instructor's Use Only:

/ 10	MCQs	Q1
/ 10	Data Types	Q2
/ 20	Linear Regression	Q3
/ 20	Logistic Regression	Q4
/ 15	Decision Trees – Implementation	Q5
/ 10	Decision Tree Ensemble – Theory	Q6
/ 85	**Total**	

Question 1 (10 marks)

1. **(1 mark)** What happens if the learning rate is too large in gradient descent?
   a) The algorithm may converge too slowly.
   b) The cost function may overshoot the minimum, increase or oscillate.
   c) The algorithm will always find the optimal solution.
   d) The cost function will not decrease at all.

2. **(1 mark)** Which of the following is true regarding the normal equation and gradient descent? (Select all that apply)
   a) The normal equation requires iterations, while gradient descent does not.
   b) Gradient descent is faster for very large datasets compared to the normal equation.
   c) The normal equation requires solving a matrix inversion.
   d) Gradient descent always converges faster than the normal equation.

3. **(1 mark)** Which of the following best explains why the linear regression cost function is not suitable for logistic regression?
   a) It is not convex for logistic regression.
   b) It does not penalize errors for classification.
   c) It leads to probabilities greater than 1.
   d) It causes the model to overfit the data.

4. **(1 mark)** Which of the following methods will guarantee finding the global minimum of the cost function for linear regression with gradient descent?
   a) Using a small learning rate.
   b) Starting with randomly initialized parameters.
   c) Running gradient descent for an infinite number of iterations.
   d) The cost function for linear regression is a parabola or a convex function, so gradient descent always finds the global minimum regardless of the learning rate or initialization.

5. **(1 mark)** In logistic regression, which of the following is true about the cost function?

   a) It is non-convex and can have multiple local minima.

   b) It is convex, ensuring a single global minimum.

   c) It increases indefinitely with the number of features.

   d) It can also be minimized using the Normal Equation.

6. **(1 mark)** How does logistic regression handle non-linearly separable data?

   a) By transforming the features using the sigmoid function.

   b) By increasing the number of iterations in gradient descent.

   c) By using higher-order polynomial features.

   d) By reducing the decision threshold.

7. **(1 mark)** What is the primary purpose of regularization in linear regression?

   a) To improve feature scaling.

   b) To prevent overfitting by penalizing large coefficients.

   c) To reduce the dataset size.

   d) To maximize R-squared value.

8. **(1 mark)** *Fill in the blank*: In linear regression, the metric that measures the proportion of variance explained by the model is called ____R-squared____.

9. **(2 marks)** *Fill in the blank:* In evaluating a classification model, ____recall____ measures the proportion of actual positives correctly identified, while ____precision____ measures the proportion of positive predictions that are actually correct.

Question 2 (10 marks)

Below is a list of attributes.

Measurement Scale	Type	Attribute
ratio	numerical, discrete	Number of Children in a Family
interval	numerical, discrete	Year of Birth
—	categorical, nominal	Martital Status (e.g., Single, Married, Divorced)
ratio	numerical, continuos	Purchase Amount (in $)
—	categorical, ordinal	Income Level (e.g., Low Income, Hight Income)

a) **(5 marks)** For each attribute, identify whether it is categorical (nominal or ordinal) or numerical (continuous or discrete). If the attribute is numerical, also specify the measurement scale (interval or ratio).

b) **(2 marks)** For the attribute "Income Level", explain how you would encode it for use in a machine learning model.

- ordinal encoder: Low ⇒ 1
  middle ⇒ 2
  high ⇒ 3

- doesnt lose information regarding order of categories

c) **(3 marks)** Describe how you would encode the "Marital Status" attribute if you were to include it in a regression model. Discuss a potential issue and how to address it.

- one-hot encoder, avoiding association of categories w/ order
- potential issue : multi-colinearity b/w features
- address : drop some categories or group them. for example, single or divorced → single
  married → married

Question 3 (20 marks)

You are a data scientist working at a real estate company. Your manager asked you to help them price a house for sale that has 3 bedrooms and a lot size of 1500 sq. meters. You ask your manager for a sample data, and he provides you with the following:

Price ($1000)	# of Bedrooms	Size (sq. meters)	House Number
200	1	900	House 1
330	3	1600	House 2
400	4	1875	House 3

$h_w(x) = w^T X$

You decide to build a model **analytically** to automatically predict the price of any house given its size (in sq. meters) and number of bedrooms.

a) **(2 marks)** Define your matrix X and your vector y.

$$y = \begin{bmatrix} 200 \\ 300 \\ 400 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 900 \\ 1 & 3 & 1600 \\ 1 & 4 & 1875 \end{bmatrix}$$

b) **(8 marks)** Solve the normal equation to build your model and write the equation of the resulting hypothesis. Show all intermediate steps.

$$J(w) = \frac{1}{2m} (Xw - y)^T (Xw - y)$$

$$J(w) = \frac{1}{2m} \left(w^T X^T X w - 2 w^T X^T y + y^T y \right)$$

$$\frac{\partial}{\partial w} J(w) = \frac{1}{2m} \left(2 X^T X w - 2 X^T y \right)$$

$$= \frac{1}{m} \left(X^T X w - X^T y \right)$$

setting $\frac{\partial}{\partial w} J(w) = 0$

$$\frac{1}{m}(X^T X w - X^T y) = 0$$

$$X^T X w = X^T y$$

$$w = (X^T X)^{-1} X^T y$$

$$y = \begin{bmatrix} 200 \\ 300 \\ 400 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 900 \\ 1 & 3 & 1600 \\ 1 & 4 & 1875 \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 4 \\ 900 & 1600 & 1875 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 4 \\ 900 & 1600 & 1875 \end{bmatrix} \begin{bmatrix} 200 \\ 300 \\ 400 \end{bmatrix} = \begin{bmatrix} 900 \\ 2700 \\ 1\,410\,000 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 4 \\ 900 & 1600 & 1875 \end{bmatrix} \begin{bmatrix} 1 & 1 & 900 \\ 1 & 3 & 1600 \\ 1 & 4 & 1875 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 8 & 4375 \\ 8 & 26 & 13200 \\ 4375 & 13200 & 6885625 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} \frac{3829}{18} & \frac{1066}{9} & \frac{-163}{450} \\ \frac{1066}{9} & \frac{1213}{18} & \frac{-46}{225} \\ \frac{-163}{450} & \frac{-46}{225} & \frac{7}{1250} \end{bmatrix}$$

$$(X^T X)^{-1} X^T y = \begin{bmatrix} \frac{1550}{3} \\ \frac{850}{3} \\ -\frac{2}{3} \end{bmatrix} = w$$

$$h_w(x) = \frac{1550}{3} + \frac{850}{3} \cdot \text{bedrooms} - \frac{2}{3} \cdot \text{size}$$

bedrooms = 3 size = 1500

c) **(2 marks)** Predict the price that you would recommend to your manager now using linear regression.

$$\frac{1550}{3} + \frac{850}{3}(3) - \frac{2}{3}(1500) = 366.67$$

$\Rightarrow \$ 366\ 670$

d) **(5 marks)** Your manager is not happy with the price, and wants to sell the house for $360,000. Based on your hypothesis, what can you recommend the manager do to the house to bring its price to $360,000 from the price that your model predicted, without changing the number of bedrooms?

increasing the size of the house by 1 m^2 would decrease the price by around \$ 666.7 · to get it down to \$ 360,000 he will need to decrease the size by 10 m^2

$$\frac{\Delta \text{ price}}{\Delta \text{ size}} = -\frac{2}{3}$$

$$\Delta \text{size} = (366.67 - 360) \cdot -\frac{3}{2} = -10\ m^2$$

e) **(3 marks)** Does your recommendation make sense? If not, explain what could be causing this counterintuitive recommendation.

↑ size cannot logically imply ↓ price. this could be due to high multicollinearity in the data that the model cannot explain, and/or unscaled features.

feature engineering & scaling methods would help mitigate this

Question 4 (20 marks)

You are a data scientist working at an IT company, and you are tasked with building a model to predict whether an employee will *churn* (i.e., leave the company) or not. The prediction is based on the employee's income and age. You have collected the following data on four employees, detailing their income (in thousands of dollars), age, and whether they churned (1 = churned, 0 = did not churn):

Approved (0 or 1)	Age (Years)	Income ($1000s)	Individual Number
0	25	45	1
1	35	60	2
0	28	50	3
1	50	90	4

You want to build a model to predict whether an employee would churn given their income and age.

a) **(2 marks)** Define a suitable cost function that you will target to optimize for this problem.

$$J(w) = \frac{-1}{m} \sum_{i=1}^{m} y \log(h_w(x)) + (1-y) \log(1 - h_w(x))$$

where $h_w(x) = \frac{1}{1 + e^{-w^T x}}$

why tf is it 3 marks?

b) **(3 marks)** How many coefficients will you need to optimize in this case?

3 coefficients: Age, Income, and the bias term / intercept

c) **(6 marks)** Using the cost function you defined in part (a), derive the update rule for each coefficient that needs to be optimized. Show your work by deriving the expression for the derivative of the cost function with respect to each coefficient.

$$w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(w)$$

$$J(w) = \frac{-1}{m} \sum_{i=1}^{m} y \log(h_w(x)) + (1-y) \log(1 - h_w(x))$$

we know that: $\frac{d}{d\sigma} \log(\sigma) = \frac{\sigma'}{\sigma}$

$$\frac{d}{dz} \sigma(z) = \sigma(z) \cdot (1 - \sigma(z)) \cdot z'$$

$$\frac{d}{w_j} w^T x = x_j$$

$$\Rightarrow \frac{d \log(\sigma(w^T x))}{d\, \sigma(w^T x)} \cdot \frac{d\, \sigma(w^T x)}{d\, w^T x} \cdot \frac{d\, w^T x}{dw_j} = \frac{1}{\cancel{\sigma}} \cdot \cancel{\sigma} \cdot (1 - \sigma) \cdot x_j$$

$$= (1 - \sigma) \cdot x_j$$

we know that: $\frac{d}{d\sigma} \log(1 - \sigma) = \frac{-\sigma'}{1 - \sigma}$

$$\Rightarrow \frac{d \log(1 - \sigma(w^T x))}{d\, \sigma(w^T x)} \cdot \frac{d\, \sigma(w^T x)}{d\, w^T x} \cdot \frac{d\, w^T x}{dw_j} = \frac{-1}{\cancel{+\sigma}} \cdot \sigma \cdot \cancel{(+\sigma)} \cdot x_j$$

$$= -\sigma \cdot x_j$$

$$\frac{d}{dw_j} J(w) = \frac{-1}{m} \sum_{i=1}^{m} y(1 - \sigma) \cdot x_j + (1-y)(-\sigma) \cdot x_j$$

$$= \frac{-1}{m} \sum_{i=1}^{m} \left[y - \cancel{y\sigma} - \sigma + \cancel{y\sigma} \right] x_j$$

$$= \frac{1}{m} \sum_{i=1}^{m} (\sigma - y) \cdot x_j$$

$$= \frac{1}{m} \sum_{i=1}^{m} (h_w(x^i) - y^i) \cdot x_j$$

$$w_j := w_j - \frac{\alpha}{m} \sum_{i=1}^{m} (h_w(x^i) - y^i) \cdot x_j$$

d

$w_0 = -2$

$w_1 = 0.02$

$w_2 = 0.02$

$\alpha = 0.01$

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$w^T x = -2 + 0.02 \begin{bmatrix} 25 \\ 35 \\ 28 \\ 50 \end{bmatrix} + 0.02 \begin{bmatrix} 45 \\ 60 \\ 50 \\ 90 \end{bmatrix} = \begin{bmatrix} -0.6 \\ -0.1 \\ -0.44 \\ 0.8 \end{bmatrix}$$

$$h_w(x) = \begin{bmatrix} 0.3543 \\ 0.475 \\ 0.3917 \\ 0.69 \end{bmatrix} \qquad (h_w(x) - y) = \begin{bmatrix} 0.3543 \\ -0.525 \\ 0.3917 \\ -0.31 \end{bmatrix}$$

$$w_0 := -2 - \frac{0.01}{4}(-0.089) = -1.9998$$

$$w_1 := -0.02 - \frac{0.01}{4}(-14.0499) = 0.0151$$

$$w_2 := -0.02 - \frac{0.01}{4}(-23.8715) = 0.0397$$

d) **(6 marks)** Assuming an initial value of -2.0 for first the coefficient, and 0.02 for all other coefficients, run one iteration of gradient descent using a learning rate of 0.01.

$h_w(x) = -1.9998 + 0.0151(\text{age}) + 0.0397(\text{income})$

e) $h_w(x)$ w/ updated coeff:

$$w^T x = \begin{bmatrix} 0.1642 \\ 0.9107 \\ 0.408 \\ 2.3282 \end{bmatrix} \quad h_w(x) = \begin{bmatrix} 0.541 \\ 0.7131 \\ 0.66996 \\ 0.91119 \end{bmatrix}$$

$$\Rightarrow y\text{-pred} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$\Rightarrow \text{accuracy} = \frac{2}{4} = 0.50$$

$$= 50\%$$

e) **(3 marks)** Calculate the model accuracy using the updated coefficients.

Question 5 (15 marks)

You are a data scientist working for a company that sells home solar panel systems. The marketing team wants to predict whether a homeowner will purchase a solar panel system based on certain characteristics. By understanding customer behavior, the company aims to target its sales efforts more effectively.

You are given a dataset containing information about previous homeowners who were offered solar panel systems, that looks like the following:

Purchased	High Credit Score	Home Ownership	Has Garage	Customer ID
Yes	Yes	Own	Yes	1
Yes	No	Own	Yes	2
No	Yes	Rent	Yes	3
No	No	Rent	Yes	4
Yes	Yes	Own	No	5
No	No	Own	No	6
No	Yes	Rent	No	7
No	No	Rent	No	8
Yes	Yes	Own	Yes	9
Yes	Yes	Own	No	10

a) **(14 marks)** Using entropy as the measure of impurity at each node, build a decision tree classifier that can predict whether a customer will purchase a solar panel system. Recall that entropy is calculated as,

$$H(p_1) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

where p_1 is the fraction of datapoints belong to class 1.

yes = 5/10 NO = 5/10

$$H(p_{root}) = -\frac{5}{10}\log\left(\frac{5}{10}\right) - \frac{5}{10}\log\left(\frac{5}{10}\right) = 1$$

high credit score

Yes → $P_1 = \frac{4}{6}$, $w = 6/10$, $H = 0.9183$

No → $P_1 = 4/4$, $w = 4/6$, $H = 0.8113$

$$\text{info gain} = 1 - \left[\frac{6}{10}(0.9183) + \frac{4}{6}(0.8113) \right] = 0.1245$$

home ownership

own → $P_1 = 5/6$, $w = 6/10$, $H = 0.65$

rent → $P_1 = 0/4$, $w = 4/10$, $H = 0$

$$\text{info gain} = 1 - \left[\frac{6}{10}(0.65) + 0 \right] = 0.61$$

has garage

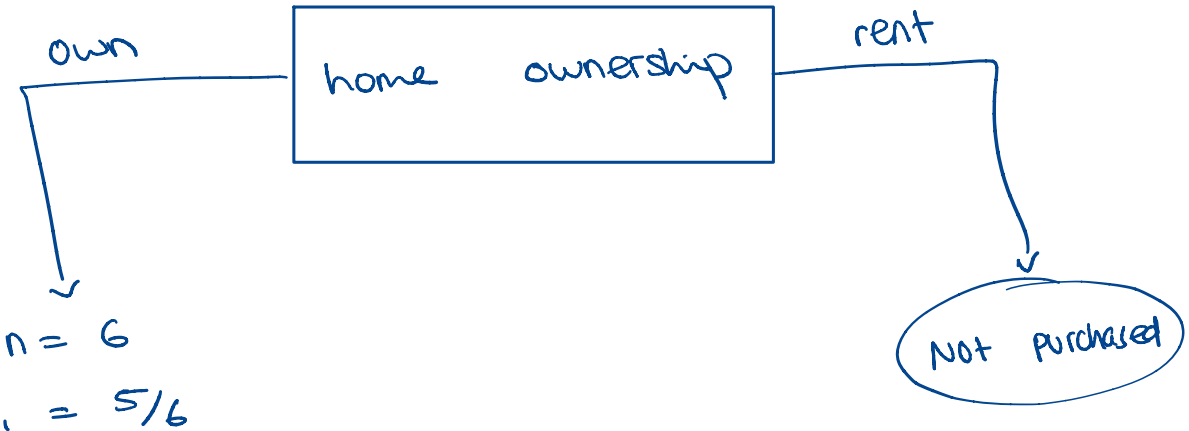
Yes → $P_1 = 3/5$, $w = 5/10$, $H = 0.971$

NO → $P = 2/5$, $w = 5/10$, $H = 0.971$

$$\text{info gain} = 1 - \left[\frac{5}{10}(0.971) + \frac{5}{10}(0.971) \right]$$

$$= 0.029$$

highest info gain : home ownership



$n = 6$

$P_1 = 5/6$

$H = 0.65$

and so on ...

b) **(1 mark)** Given a new homeowner with the following characteristics:

- Doesn't have a garage,
- Rents, and,
- Doesn't have a high credit score,

use your decision tree to predict whether this customer will purchase the solar panel system.

Rents ⇒ will not purchase

Question 6 (10 marks)

a) **(1 mark)** Discuss a serious limitation of a single decision tree.

- sensitive to small changes in data
- not very robust

b) **(4 marks)** Explain the difference between bagging and boosting in the context of creating decision tree ensembles.

bagging:

- descion trees are trained in parallel
- bootstrapped data for each D.T. where the data is sampled w/ replacement
- final decision is made by aggregating all decisions
- less computationally expensive

boosting:

- decision trees traineel sequentially
- misclassified samples from a tree are given higher weights when passed on to next tree
- more computationally expensive

c) **(2 marks)** Give one example of a bagging decision tree ensemble.

random forests, where both the samples & features are bootstrapped, then decisions are aggregated

d) **(3 marks)** In XGBoost, how does the training process differ from standard boosting ensembles in terms of target variable and handling of misclassified points?

- trees are trained sequentially & based on residuals of previous tree
- instead of giving higher weights to misclassified points, the loss gradient & hessian are used for better handling of misclassification
- loss function in XGBoost is regularized, avoiding overfitting

not sure about this at all lol