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

Midterm Exam

October 31, 2024 (2 hours)

Student Name: Wooooooood **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
/ 5	Decision Tree Ensemble – Theory	Q6
/ 80	**Total**	

Question 1 (10 marks)

1. **(1 mark)** Which of the following is an example of multivariate linear regression?

   a) Predicting the amount of rainfall based on the temperature.

   b) Predicting a person's monthly electricity bill based on the number of residents, square footage, and average daily usage.

   c) Predicting a student's exam score based on their study hours.

   d) Predicting whether an email is spam or not based on its word count.

Because multiple input features.

2. **(1 mark)** Logistic regression is best suited for which type of problem?

   a) Predicting continuous outcomes.

   b) Classifying data into binary categories.

   c) Finding clusters in unlabeled data.

   d) Predicting the price of a house based on features such as size, number of floors, and age.

Self explanatory

3. **(1 mark)** What happens if the learning rate is too large in gradient descent?

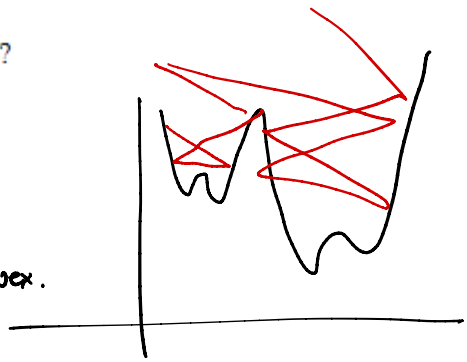
   a) The algorithm may converge too slowly.

   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.

Recall: non convex.



4. **(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. X

   b) Gradient descent is faster for very large datasets compared to the normal equation. $n \geq 1000$ features

   c) The normal equation requires solving a matrix inversion. $(X^T X)^{-1}$

   d) Gradient descent always converges faster than the normal equation. X

5. **(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.

Not entirely sure why this is the case.

6. **(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.

Smaller updates are better, but also much slower.
Infinite should be right by nature.

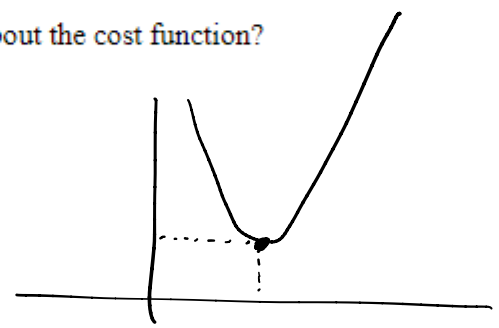
7. **(1 mark)** In logistic regression, which of the following is true about the cost function?

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.



8. **(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.

Not sure about this one either.

9. **(1 mark)** How can you extend logistic regression for multi-class classification?

a) Use a different threshold for each class.

b) Use one-vs-all or one-vs-rest approach. Self explanatory.

c) Train a binary classifier for each pair of classes.

d) No extension is necessary, as logistic regression can naturally handle multi-class classification.

10. **(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. Self explanatory again.

Question 2 (10 marks)

Below is a list of attributes.

Measurement Scale	Type	Attribute
Ratio scale	Numerical, discrete	Number of Children in a Family
Interval, because no true zero	Numerical, discrete	Year of Birth
_____	Categorical, nominal	Martital Status (e.g., Single, Married, Divorced)
Ratio (true zero)	Numerical, continuous	Purchase Amount (in $)
_____	Categorical (Cordinal)	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.

Label encoding. Why? Because it is an ordinal feature. one-hot encoding would not work.

0	High inc.
1	Med. inc.
2	Low inc.

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

Marital status: categorical, nominal ⟶ we can do one-hot encoding.

Choose $I_3 \in \mathbb{R}^{3\times 3}$: $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ ⟶ single, ⟶ married, ⟶ divorced

This is good b/c we maximize the Euclidean distance between each of the three levels.

Issue: Multicollinearity. May not contribute meaningfully to the outcome of prediction.

Maybe it could be dropped, regularized, etc...

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 a lot size of 1500 sq. ft. You ask your manager for a sample data, and you are provided with the following:

not population!!

Price ($1000)	Size (sq. ft)	House Number
180	850	House 1
310	1400	House 2
450	2000	House 3

a) **(5 marks)** You decide to first normalize your data to have 0 mean and 1 standard deviation. Normalize the data and enter the normalized values in the table below:

σ = standard deviation

σ^2 = variance.

μ = mean.

To get mean 0 std 1,

$\rightarrow \frac{x-\mu}{\sigma}$

$\mu = \frac{850 + 1400 + 2000}{3} = 1416.669$

$\sqrt{\frac{(850 - 1416.669)^2 + (1400 - 1416.69)^2 + (2000 - 1416.69)^2}{n-1}}$; $n = 3$.

$\therefore \sigma = 600.247$.

$\Rightarrow \frac{800 - 1416.69}{600.347} = -1.0271$

Price ($1000)	Size (sq. ft)	House Number
180	-1.0271	House 1
310	-0.0278	House 2
450	0.9717	House 3

Do not scale the target variable!

b) **(8 marks)** Use Ordinary Least Squares (OLS) to build your model using the **normalized data** and write the equation of the resulting hypothesis. Recall that the OLS equations are the following:

$$w_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2}$$

These formulae are given?

Normalized data means $\mu = 0, \sigma = 1$.

$$w_0 = \bar{y} - w_1 \bar{x}$$

Thus: $w_1 = \frac{\sum_{i=1}^{m}(x_i)(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i)^2} = \frac{180(-1.0291) + 310(-0.0298) + 450(0.9919)}{(1.0291)^2 + (-0.0298)^2 + (0.9919)^2}$

$= \frac{269.841}{1.9999} \approx 134.929$

$w_0 = \bar{y} - w_1 \bar{x}$

$= 313.333$

$h_w(x) = w_0 + w_1 x = 313.333 + 134.929\ x$

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

At 1500 sq. ft, we first transform: $\frac{1500 - 1416.667}{600.149} = 0.1388$

$\Rightarrow$ plug it in: $313.333 + (0.1388)(134.929) = \$332,060$

d) **(5 marks)** Your manager is not happy with the price, and wants to sell the house for $350,000. By leveraging the size coefficient in your hypothesis, what can you recommend the manager do to the house to bring its price to $350,000 from the price that your model predicted.

Hint: Consider how the coefficient translates to the original scale of the "Size" feature.

Let's transform back. We have $\hat{x} = \frac{x-\mu}{\sigma} \Rightarrow \sigma\hat{x} + \mu = x$.

$\therefore \quad 313.333 + \left(\frac{x-\mu}{\sigma}\right) 134.929 = y$

$\therefore$ Same stuff happens here but it's actually much simpler.

$313.333 + x\,(134.929) = 350$

$\therefore \quad x = 0.2918 \longrightarrow \hat{x} = \frac{x-\mu}{\sigma} \Rightarrow \sigma(0.2918) + 1416.667$

$= 1579.81$

House has to be 1579 sq. ft. to be worth $350,000.

Question 4 (20 marks)

You are a data scientist at a company that provides email security solutions. The company has developed two logistic regression models, *Model A* and *Model B*, to classify incoming emails as "Spam" or "Not Spam". Both models have been tested on the same dataset, and their performance is summarized in the confusion matrices below.

Model A:

		Predicted: Not Spam	Predicted: Spam
Actual	Not Spam	TN 80	20 FP
	Spam	FN 15	85 TP

Model B:

		Predicted: Not Spam	Predicted: Spam
Actual	Not Spam	TN 90	FP 10
	Spam	FN 25	TP 75

a) **(5 marks)** Calculate the Precision and Recall for both *Model A* and *Model B*. Show your calculations.

Model B: Precision: $\frac{TP}{FP + TP} = \frac{75}{85} = 0.88$ Recall: $\frac{TP}{TP+FN} = \frac{75}{75+25} = 0.75$

Model A: Prec.: $\frac{85}{105} = 0.81$ $\frac{85}{100} = 0.85$

b) **(5 marks)** Your company is considering deploying one of these models based on different business priorities. For each of the following scenarios, indicate which model (*Model A* or *Model B*) would be more suitable and justify your choice.

Precision: of all positive predictions, how many were actually positive?

Recall: of all actual positives, how many predicted positive?

Scenario 1:
The company wants to minimize the number of legitimate emails incorrectly marked as spam to avoid inconveniencing users.

We want a model w/ better precision. Therefore model B.

Scenario 2:
The company aims to maximize the detection of all spam emails to protect users from potential threats, even if it means some legitimate emails are mistakenly marked as spam.

Recall more important now. Therefore model A.

c) **(10 marks)** Assume that for your company the cost of a false positive is $2 per incident, and the cost of a false negative is $10 per incident. Calculate the total cost associated with each model based on the confusion matrices provided. Which model results in a lower total cost?

Costs: $2 for FP
$10 for FN.
} Just do the math.

Model A: $190
Model B: $290

Model A cheaper eh?

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:

Target

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

Don't care.

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.

Calculate $H(p_1) = -\left[\frac{5}{10}\log_2\left(\frac{5}{10}\right) + \frac{5}{10}\log_2\left(\frac{5}{10}\right)\right]$

$= 1$ ——▷ this is the maximum impurity.

Look at the other splits:

yes: 6/10 = w; 4/6 = ✓

$H = -\left[\frac{4}{6}\log_2(\frac{4}{6}) + \frac{2}{6}\log_2(\frac{2}{6})\right]$

$= 0.9182$

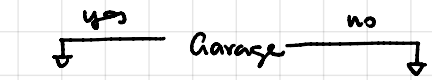
no: w = 4/10; 1/4 = ✓; 3/4 = ×

$H = -\left[\frac{1}{4}\log_2(\frac{1}{4}) + \frac{3}{4}\log_2(\frac{3}{4})\right]$

$= 0.8112$

$\longrightarrow 1 - \left[\frac{6}{10}(0.9182) + \frac{4}{10}(0.8112)\right]$

$= 0.1245$ = information gain



yes: w = 5/10; y = 3/5; n = 2/5; H = 0.991

no: w = 5/10; y = 2/5; n = 3/5; H = 0.991

$I.G = 1 - \left[\frac{1}{2}0.991 + \frac{1}{2}0.991\right]$

$= 0.029$

own: 6/10 w; y = 5/6; n = 1/6; H = 0.65

rent: 4/10 w; y = 0/4; n = 4/4; H = 0.

$I.G. = 1 - \frac{6}{10}(0.65) = \underline{0.61}$

Based on root homeowner.

$H(p_-) = -\left[\frac{5}{6}\log_2(\frac{5}{6}) + \frac{1}{6}\log_2(\frac{1}{6})\right]$

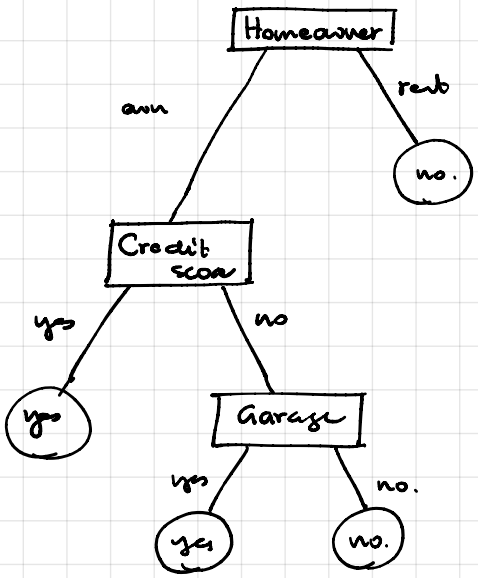
$= \underline{0.65}$

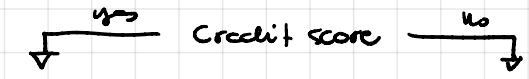
yes: w = 3/6; y = 2/3; n = 0/3; H = 0

no: w = 3/6; y = 2/3; n = 1/3; H = 0.9183

$IG = 0.65 - \frac{1}{2}(0.9183) = 0.191$

Root node: Homeowner (highest information gain).





yes: w = 4/6; y = 4/4; n = 0/4; H = 0

no: w = 2/6; y = 1/2; n = 1/2; H = 1

$I.G. = 0.65 - \frac{2}{6}(1) = \underline{0.2166}$

Split on credit score next.

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.

If they rent, immediately no.

Question 6 (5 marks)

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

Highly sensitive to small changes in data, not very robust. Bad at generalizing. There's only a few. Not good for unstructured data.

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

Bagging: for some $b := 1, 2, \ldots, B$ (with $B = 64$ to 128 typically), sample with replacement to get B new training subsets of size m. Probability of sampling from each of these new subsets is $(\frac{1}{m})$.

Boosting: $b := 1, 2, \ldots, B \Rightarrow$ subsets of training data w/ size m, but probability of picking is no longer $(\frac{1}{m})$. It is now more likely to choose to train on examples that were misclassified in previous trees.

Example: Don't worry about it.