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
Assignment 2
Due: October 27, 2024 by 11:59PM

Question 1

Consider the following data on experience (in years) and salary (in thousands of dollars) of a sample of employees:

$x_i = \bar{x}_{mean} \; y_{mean} \; y_i$
total total total total

Experience (Years)	Salary ($1000)
1.2	42
2.5	46
3.1	51
3.9	58
5.2	62

a) Build a model that can help you automatically predict the salary of an employee given their years of experience. Write down the equation of your model's hypothesis.

b) Use your hypothesis to predict the salary of an employee with 4.5 years of experience.

c) Interpret the coefficients of your hypothesis.

Question 2

You have the following data on size (in square feet) and number of bedrooms of 3 houses and their respective prices (in thousands of dollars):

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

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

a) Define your matrix $\boldsymbol{X}$ and your vector $\boldsymbol{y}$.

$$X = \begin{bmatrix} 1 & 900 & 1 \\ 1 & 1600 & 3 \\ 1 & 1875 & 4 \end{bmatrix} \quad y = \begin{bmatrix} 200 \\ 330 \\ 400 \end{bmatrix}$$

b) Solve the normal equation and write the equation of the resulting hypothesis.

c) Predict the price of a house that is 1500 sq. ft. and has 3 bedrooms.

$$h_w(x) = w_0 + w_1 x_1 + w_2 x_2$$

d) Interpret the coefficients of your hypothesis.

$$h_w(x) = (X^T X)^{-1} X^T \cdot y$$

Question 3

You have the following data on the income (in thousands of dollars) and age of 6 individuals and whether they were approved for a credit card (1 = approved, 0 = not approved):

Individual Number	Income ($1000s)	Age (Years)	Approved (0 or 1)
1	45	25	0
2	60	35	1
3	75	40	1
4	50	28	0
5	90	50	1
6	100	60	1

You want to build a model to predict whether a new applicant will be approved for a credit card given their income and age.

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

b) How many coefficients will you need to optimize in this case?

c) 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.

d) Assuming an initial value of 1 for the coefficient, 2 for the second coefficient, 3 for the third coefficient, 4 for the fourth coefficient, and so on, run one iteration of gradient descent using a learning rate of 0.01.

e) Calculate the model accuracy using the updated coefficients.

Question 4

You built a model that can predict whether a customer would churn (i.e., leave the company) (1 = churn, 0 = no churn) using a number of features, with one them being x_1 (salary). In the dataset you leveraged to build the model, only 10% of the customers churned. The remaining 90% did not churn. The confusion matrix for applying the model to a test set is as follows:

		Predicted	
		No Churn (0)	Churn (1)
Actual	No Churn (0)	180	20
	Churn (1)	70	30

a. Calculate the accuracy, precision, recall, and F1 score for the model.

b. Interpret the F1 score and explain when it is more useful than accuracy.

c. Suppose the coefficient corresponding to the feature "salary" in the model you built is -0.6. For an employee with a salary of $\$30{,}000$, the probability of churning is 0.63. What would the probability of churning be if the company were to increas this employee's salary to $\$31{,}000$?

Question 5

You want to automatically decide whether today would be a good day to play tennis or not based on the day's outlook, temperature, and humidity. You have some recorded historical data, as shown below:

Outlook	Temperature	Humidity	Play Tennis
Sunny	Hot	High	No
Sunny	Hot	Normal	Yes
Overcast	Hot	High	Yes
Overcast	Mild	High	Yes
Sunny	Mild	Normal	No

a. Using entropy as the measure of impurity at each node, build a decision tree classifier that can predict whether today would be a good day to play tennis or not based on outlook, temperature, and humidity. 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 belonging to class 1 (Yes to Playing Tennis).

b. Use the decision tree classifier to predict whether to play tennis or not for a day with the following features:
   - Outlook: Overcast
   - Temperature: Hot
   - Humidity: Normal