CS 284 Machine Learning

Midterm Examination

Name: ____ Semester:_____ AY:____________

Instructions: Submit this questionnaire along with a PDF version containing your answers to each item. When you are required to provide code for a specific item, include a link next to the item label in your answer sheet. For answers requiring analytical solutions or proofs, you must write them by hand and scan them (or take a clear picture). Then, include the scanned pages as part of your PDF submission. **Submit only a single PDF file containing all your answers.**

1. **[6 pts]** A company is considering launching a new product. It has collected data on the number of units sold (Y) and the amount spent on advertising (X) in dollars for similar products in the past, as shown in the table below. The data points in red constitute the test set.

Table 1.1. Dataset for item number 1.

X	6	35	77	4	27	7	16	61	14	84	91	30	72	12	77	80	4	48	90	95
Y	105	117	128	101	113	104	106	123	103	138	132	108	129	108	128	130	95	108	143	137

1.1. Determine the simple linear regression equation that best fits the data.

1.2. Create a scatter plot representing the relationships in the tabulated data. Include the regression line in the plot.

1.3. Calculate the training and testing R^2.

1.4. What would be the expected number of units sold if the company does not spend anything on advertising?

1.5. If the company spends 50 dollars, what would be the expected number of units it can sell?

1.6. Suppose the company spends 1000 dollars on advertising. What would be the expected number of units sold? Explain why you would or would not be willing to trust this prediction.

2. **[7 pts]** Suppose we modify the objective function of the OLS formulation of the multiple linear regression as follows:

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \underset{\beta_0, \beta_1, \dots, \beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where n is the number of samples and p is the number of features. Assuming $\lambda \geq 0$ is an arbitrary value, provide an analysis on the effect on the parameter estimates under the following settings:

2.1. Setting $\lambda = 0$

2.2. Setting $\lambda \ll \min_{1 \leq j \leq p} |\beta_j|$

2.3. Setting $\lambda \gg \max_{1 \leq j \leq p} |\beta_j|$

Note: *The symbol "<<" means much less compared to a reference point while the symbol ">>" means much greater compared to a reference point.*

3. **[7 pts]** Suppose that we modify the objective function of the OLS formulation of the multiple linear regression as follows:

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \underset{\beta_0, \beta_1, \dots, \beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where n is the number of samples and p is the number of features. Assuming $\lambda \geq 0$ is an arbitrary value, provide an analysis on the effect on the parameter estimates under the following settings:

3.1. Setting $\lambda = 0$

3.2. Setting $\lambda \ll \left(\min\limits_{1 \leq j \leq p} |\beta_j| \right)^2$

3.3. Setting $\lambda \gg \left(\max\limits_{1 \leq j \leq p} |\beta_j| \right)^2$

Note: *The symbol “<<” means much less compared to a reference point while the symbol “>>” means much greater compared to a reference point.*

4. **[8 pts]** Suppose you have a random sample $X_1, X_2, \dots, X_n$ drawn from the following distribution with probability mass function:

$$p(X_i = x_i; r, q) = \binom{x_i + r - 1}{x_i} q^r (1 - q)^{x_i} \text{ , } x_i = 0, 1, 2, \dots$$

where $r > 0$ (known) and $p \in (0,1)$.

4.1. Construct the likelihood function

4.2. Construct the log-likelihood function

4.3. Find the maximum likelihood estimator for the unknown parameter q.

Note*: as in combinatorics,* $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ *, also known as “n choose r.”*

4.4. Show whether the MLE for q is a biased or unbiased estimator.

5. **[5 pts]** You are given the following moment generating function of a real-valued random variable $X \sim D(p)$below:

$$M_X(t) = \frac{p}{1 - (1 - p)e^t}$$

5.1. Calculate $E[X]$.

5.2. Calculate $E[X^2]$.

5.3. Calculate the variance of X.

5.4. Find the method of moments estimator for p.

6. **[10 pts]** For this exercise, you are required to provide the link to your Jupyter Notebook and its PDF version as proof of your solution. Failure to provide proof will result in no points being awarded for this exercise. You are given the following datasets to model the relationship of a dependent variable **Y** with three independent variables **X1, X2,** and **X3**.

CS 284 Machine Learning
Midterm Examination

Name: ____ Semester:_____ AY:___________

Table 6.1. Training dataset for item number 6

	X1	X2	X3	Y
1	10.99	10.69	4.52	58.76
2	9.72	10.65	4.81	58.14
3	11.3	11.29	3.89	58.96
4	13.05	12.52	3.8	66.8
5	9.53	9.94	5.81	53.91
6	9.53	8.92	6.36	49.25
7	13.16	13.26	4.93	71.62
8	11.53	10.56	6.0	61.26
9	9.06	8.4	5.36	48.69
10	11.09	11.18	4.35	59.6
11	9.07	9.44	5.36	48.99
12	9.07	9.15	6.54	51.29
13	10.48	10.43	4.96	56.52
14	6.17	6.02	6.56	35.36
15	6.55	5.81	2.38	32.59
16	8.88	8.52	5.82	49.95
17	7.97	7.74	5.09	48.02
18	10.63	11.16	4.7	59.79
19	8.18	8.36	5.09	47.05
20	7.18	6.29	3.01	36.09

Table 6.2. Testing dataset for item number 6

	X1	X2	X3	Y
1	12.93	13.09	4.78	66.07
2	9.55	9.36	5.36	52.49
3	10.14	9.8	6.48	56.28
4	7.15	7.46	4.48	46.09
5	8.91	9.43	4.19	49.92
6	10.22	10.69	4.5	57.61
7	7.7	7.28	5.92	43.09
8	10.75	10.6	5.33	56.29
9	8.8	8.96	4.47	51.24
19	9.42	9.9	5.51	55.55

6.1. Fit a multiple linear regression model using the training data. **(1 pt)**

6.2. Create a table that compares the predictions of your model to the target values of the training set.

6.3. Calculate the adjusted training R^2.

6.4. Create a table that compares the predictions of your model to the target values of the testing set.

6.5. Calculate the adjusted testing R^2.

6.6. Calculate the variance inflation factor of each feature.

6.7. Determine if there are any features in the data that are potentially causing multicollinearity. If there are any, discuss ways of handling this multicollinearity.

6.8. Provide the updated model after having performed the suggested remedy in the previous item (item g). If you don't think the model needs to be updated, then provide an explanation.

7. **[10 pts]** For this exercise, you are required to provide the link to your Jupyter Notebook and its PDF version as proof of your solution. *Failure to provide proof will result in no points being awarded for this item*. Given the simulated dataset provided in the following link (the training data is named **"item_7_training_data.csv"** and the testing data is named **"item_7_testing_data.csv"**), perform the following tasks:

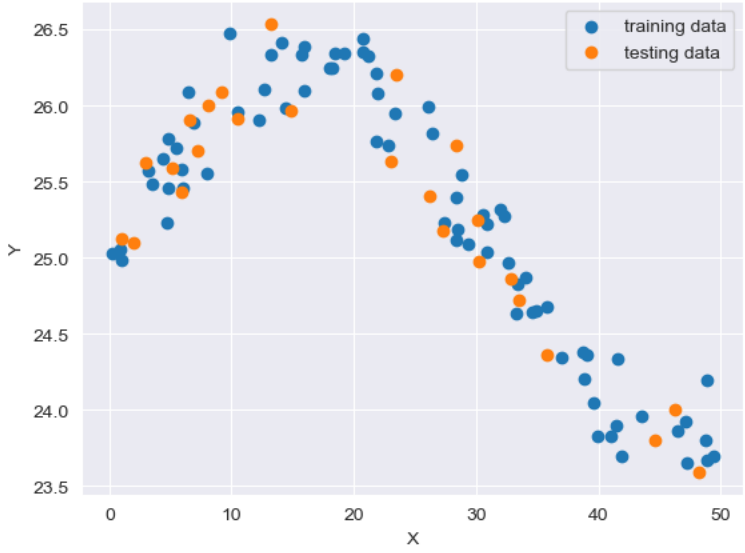

Figure 7.1 Scatter plot of the simulated data

7.1. Implement the code to fit a polynomial regression model to the given data.

7.2. Fit polynomial regression models with degrees from 1 up to 7.

7.3. Using the test data, compare the adjusted R^2 values of each model. Create a line plot showing the adjusted R^2 for each model, and include the model degrees in the legend.

7.4. Based on the adjusted R^2 criterion, determine which model is the best for generalizing to unseen data.

7.5. Write the full equation of the best polynomial regression model. **For example:**

$$\hat{Y} = 1.2 + 3.4X + 1.2X^2$$

8. **[5 pts]** Apple and Samsung sometimes launch similar products in the same business period. The events A and B are defined as follows:

$A =$ the event that Apple launches a product.
$B =$ the event that Samsung launches a product.

CS 284 Machine Learning

Midterm Examination

Name: ____ Semester:_____ AY:___________

Given $P(A) = 0.25,\ P(A \cap B) = 0.15,\quad P(A' \cap B') = 0.17$, answer the following questions *(note that A' is the complement of A):*

8.1. Find the probability that Samsung will launch a product on a randomly selected business period.

8.2. Find the probability that Samsung will launch a product given that Apple launched a product.

8.3. Determine if A and B are independent events.

9. **[5 pts]** You are given the following probability distribution given by a discrete random variable X.

Table 9.1. Probability distribution for item number 9

X	0	1	2	3
$P(X = x)$	0.4	0.3	0.2	0.1

9.1. Find the expectation of X.

9.2. Calculate the second theoretical moment of X.

9.3. Calculate the variance of X.

9.4. Calculate the variance of a new random variable $Y = 5 - 3X$.

10. **[10 pts]** For this item, you are expected to develop a case scenario that demonstrates your creativity in applying multiple linear regression to a realistic, real-world situation. You DO NOT NEED to implement this in code. Your narrative and the logical flow of your discussion should be sufficient.

"Imagine that you are a data scientist tasked with analyzing the factors that influence home prices in a specific real estate market. You will use multiple linear regression to model this relationship."

10.1. Case Scenario: Develop a detailed case scenario describing the real estate market you are analyzing. Specify the dependent variable (home prices) and at least three independent variables (e.g., square footage, number of bedrooms, location) that you believe could significantly impact home prices. Justify your choice of these variables based on relevant literature or market knowledge.

10.2. Model Specification: Write the equation for the multiple linear regression model you will use to analyze the data. Explain the meaning of each term in the equation, including the intercept and coefficients for each independent variable. This equation can be fictitious.

10.3. Data Collection: Describe how you would collect data for your analysis. Discuss potential sources of data (e.g., real estate listings, public property records) and any challenges you might face in obtaining accurate and comprehensive data.

10.4. Model Fitting and Evaluation: Outline the steps you would take to fit the multiple linear regression model to your data. Discuss how you would evaluate the model's performance, including metrics such as R-squared, adjusted R-squared, and significance of coefficients.

10.5. Interpretation of Results: Explain how you would interpret the results of your regression analysis in terms of the impact of each independent variable on home prices. What insights could this provide to stakeholders in the real estate market?

10.6. Practical Implications: Reflect on how your findings could influence decision-making for various stakeholders, such as home buyers, sellers, and real estate agents. Consider any limitations of multiple linear regression and how they might affect your conclusions.