CompSci 526: Intro To Data Science **Fall 2025**

Homework 2

Due: Tuesday 10/7 at 11:59 pm ET (submit via Gradescope)

Submission: Please submit your responses in a single PDF file. Ensure that all written content is typed; hand-drawn figures are acceptable. Code snippets or outputs should also be included within the PDF. When submitting on Gradescope, you will be prompted to label pages for each question. Make sure to complete this labeling accurately to expedite the grading process.

Overview: This homework has 4 questions, for a total of 60 points.

Policy: Collaboration is encouraged, but each student must submit their own individual work. If you collaborate with others, please acknowledge your collaborators by explicitly listing their names below.

Names of your collaborators:

1. Let us conduct regression analysis on the bike sharing data set. The data set description is available at `https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset`, and the data set can be downloaded at `https://archive.ics.uci.edu/ml/machine-learning-databases/00275/Bike-Sharing-Dataset.zip`.

   We will use only the data in file `day.csv`.

   (a) (2 points) Plot the marginal distribution for the count of bike rentals and the conditional count distribution given the weather situation (`weathersit`).

   (b) (4 points) Fit a linear regression for ride count as a function of the weather situation variable. Carefully consider how to incorporate `weathersit` into the model, and justify your choice of modeling approach. Report the coefficients of your model and explain the implications of the approach you used for the `weathersit` variable.

   (c) (2 points) What is the difference in expected ride counts when the weather is clear (1) versus wet (3)?

   (d) (3 points) What are the in-sample residual sum of squares (RSS), R^2, and estimated standard deviation of the residual errors for the ride-weather regression?

   (e) (4 points) Fit a linear regression for the ride counts onto all of the weather variables (`weathersit`, `temp`, `hum`, `windspeed`). You can ignore the date/time variables. What is the impact on expected ride count due to a 10 degree increase in the temperature?

   (f) (4 points) Fit a logistic regression to model ride counts using weather variables (`weathersit`, `temp`, `hum`, `windspeed`), excluding date/time variables. To categorize the total rental bike count, establish qualitative labels: 'Low Demand' and 'High Demand.' Specifically, designate counts less than or equal to 4000 as 'Low Demand' and counts greater than 4000 as 'High Demand.' Evaluate and report the performance of the logistic regression model using a train-test split. Include the accuracy of the model on both the training and test sets.

   (g) (5 points) Choosing an appropriate threshold in logistic regression hinges on the specific context of your problem and analysis goals. In contrast to the 4000 threshold employed in the previous question, opt for a different threshold guided by some criteria. Clearly articulate the criterion you consider for this choice. Evaluate and report this logistic regression model's performance, including accuracy and F-1 score metrics, in comparison to the previous question. Additionally, provide a rationale for any observed improvement or lack thereof in performance.

   Note that the F-1 score is calculated using the formula: $2 * TP \Big/ (2 * TP + FP + FN)$, where TP represents true positives, FP represents false positives, and FN represents false negatives.

2. Explore the following theoretical aspects of simple linear regression:

   (a) (3 points) Show that in simple linear regression, the mean of the residuals e_i is always zero.

   (b) (3 points) Show that in simple linear regression, the residuals e_i are orthogonal to the predictor variable X_i, i.e., their dot product sums to zero $\sum X_i e_i = 0$.

   (c) (3 points) In statistics, two variables are uncorrelated if their covariance is zero. Show that in simple linear regression, the residuals are uncorrelated with the predicted responses, i.e., their covariance is zero: $\text{Cov}(e, \hat{Y}) = \frac{1}{n}\sum(e_i - \bar{e})(\hat{Y}_i - \overline{\hat{Y}}) = 0$.

   (d) (3 points) Show that in simple linear regression, the mean of the predicted responses equals the mean of the observed responses, i.e., $\overline{\hat{Y}} = \bar{Y}$.

   (e) (3 points) Show that in simple linear regression, R^2 can be expressed as the ratio of the explained sum of squares (ESS) to the total sum of squares (TSS) by starting from its definition: $R^2 = 1 - \frac{RSS}{TSS}$. Here, ESS quantifies the variance explained by the regression model. It measures how much of the total variation in Y is captured by the fitted regression line: $ESS = \sum(\hat{Y}_i - \bar{Y})^2$.

   (f) (3 points) Show that in the case of simple linear regression with Y as the response variable and X as the predictor variable, the R^2 statistic is equal to the square of the correlation between X

and Y, without assuming that the means of X and Y are zero. The correlation coefficient r_{XY} is defined as:

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}.$$

3. Consider a logistic regression model that predicts the probability of developing a particular health condition based on daily sugar intake. The model uses the sigmoid function as its hypothesis:

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 \cdot x)}},$$

where x represents the daily sugar intake (in grams), θ is the parameter vector, and $h_\theta(x)$ is the predicted probability of developing the health condition. The log-likelihood function for logistic regression is expressed as:

$$\ell(\theta) = \sum_{i=1}^{m} \left[y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right],$$

where m is the number of training examples, $x^{(i)}$ is the daily sugar intake for the i-th training example, and $y^{(i)}$ is the corresponding label (1 for developing the condition, 0 for not developing it). The goal is to maximize this log-likelihood with respect to θ.

Given the following simplified training set, explore how to manually calculate and maximize the log-likelihood for this logistic regression model:

Sugar Intake (g)	Condition Developed (Yes=1/No=0)
30	0
50	0
70	1
90	1

(a) (2 points) Calculate $h_\theta(x)$ for each training example using the initial parameter values $\theta_0 = 0$ and $\theta_1 = 0$.

(b) (2 points) Calculate the log-likelihood $\ell(\theta)$ using the initial parameter values.

(c) (4 points) Explain the process of maximizing the log-likelihood. Discuss how gradient ascent (the optimization technique opposite to gradient descent) can be applied to adjust the parameters θ_0 and θ_1 to increase $\ell(\theta)$. Then, perform the first iteration of gradient ascent to update θ_0 and θ_1, assuming a learning rate of $\alpha = 0.01$.

4. Consider a travel application designed to recommend activities based on a user's preferences. The app classifies activities into three categories: cultural experiences (0), adventure outings (1), and relaxation spots (2). Each user's preferences are represented by two features: their `Excitement Level` (x_1) on a scale from 1 to 10 and their `Budget` (x_2) in dollars.

In this scenario, you will use a multiclass perceptron to predict the `Activity Category` based on user preferences.

Training data:

Sample #	`Excitement Level` (x_1)	`Budget` (x_2)	`Activity Category`
1	3	100	0
2	8	300	1
3	5	150	2

The initial weights (W) for the multiclass perceptron are set randomly as:

$$W = \begin{bmatrix} 0.4 & 0.1 & -0.3 \\ 0.3 & -0.2 & 0.5 \\ -0.1 & 0.3 & 0.2 \end{bmatrix},$$

where the first column represents the bias terms, and each row corresponds to the weight parameters for a specific category: the first row for category 0, the second row for category 1, and the third row for category 2.

(a) (9 points) For each sample in the training dataset, use the multiclass perceptron learning rule to update the weights if the predicted `Activity Category` does not match the actual category. In the event of a tie (multiple categories with the highest score), resolve it by selecting the category with the smallest number. Perform a single update pass for each training sample, starting from sample 1 and continuing through sample 3.

(b) (1 point) Using the updated weights from the previous question, predict the `Activity Category` for a new set of user preferences: `Excitement Level = 6` and `Budget = 200`.