CS 284 Machine Learning
Final 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. For this exercise, you need to **provide a link to your Jupyter notebook**. Given the training dataset and testing dataset, implement a class named "CustomPolynomialRegression" with the following methods:
    a. **fit(x,y)**: takes a 1-D vector x containing the features and a 1-D vector y containing the targets, and fits a polynomial regression.
    b. **test_score(x,y)**: takes a 1-D vector x containing the features and a 1-D vector y containing the targets, and returns the "root mean squared error" and the "mean absolute error."
    c. **goodness_of_fit(x,y):** takes in a 1-D vector x containing the feature and a 1-D vector y containing the target and returns the Akaike Information Criterion (AIC):

$$AIC = n \times \ln\left(\frac{SSE}{n}\right) + 2k$$

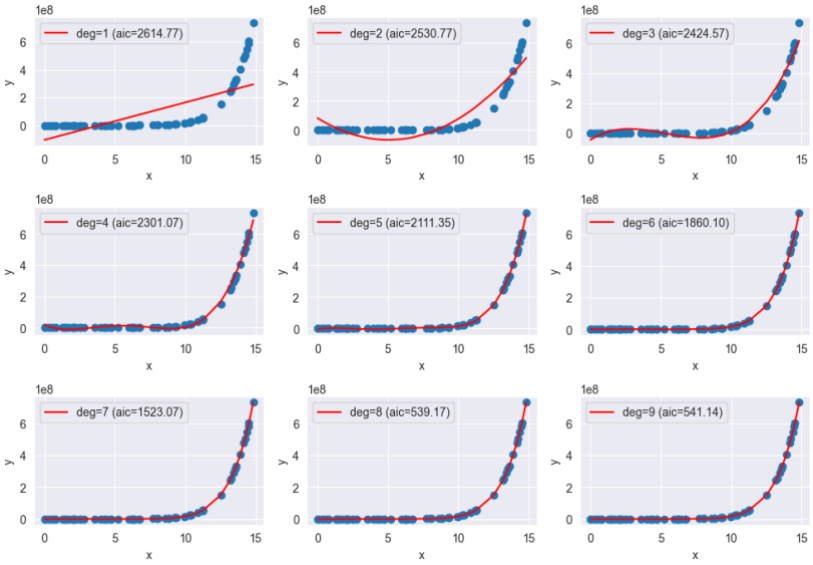
    where n is the number of samples, SSE is the sum of squares error, and k is the number of model parameters.
    d. **predict(x):** takes in a 1-D vector x containing the feature and returns the predictions

    Answer the following questions:

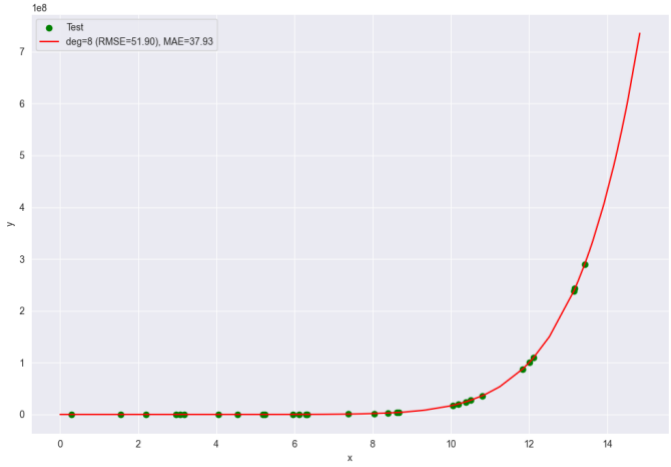
    a. Using the **goodness_of_fit** function for model selection, create a 3-by-3 subplot that displays each model from degree 1 to degree 9. Each subplot must show a scatter plot of the training data alongside the predictions of the fitted model, including its AIC. For example, the subplot layout might look like the figure below (NOTE: this is for demonstration purposes only; the plots shown do not correspond to the data for this item). *What are the degree and AIC of the best model?*

CS 284 Machine Learning
Final Examination

Name: _____ Semester:_____ AY:_____



b. What are the root mean squared error (RMSE) and the mean absolute error (MAE) of the best model?
c. Using the selected model from **item a**, create a plot depicting the test data points and the model's predictions. For example, the plot would look something like:



d. Provide the regression equation for the best model.
e. Discuss whether the best model has the same degree as the data-generating process (i.e., the true functional form generating the data). What do you think the functional form of the data-generating process is? Justify your answer.

2. **Create** a toy dataset and **discuss** the Bias-Variance Tradeoff. You may choose any of the models discussed to explain the bias-variance tradeoff. For example, you may use the Cubic Splines Regression Model as the model of choice and vary the number of knots as

Name: _____ Semester:_____ AY:_____

shown in class. However, you should generate your own toy dataset for this. **Provide a link to the Jupyter notebook supporting your discussions.**

3. You are given the breast cancer dataset in the following link. Create a logistic regression model and perform the following. **Provide a link to your Jupyter notebook.**
    a. To simplify the preprocessing steps, simply remove rows with missing values and NaN. Furthermore, standardize your data using the z-score normalization (or StandardScaler in scikit-learn).
    b. Create a correlation heat map depicting the pairwise correlation of each feature.
    c. Fit a logistic regression model on the training data without performing regularization. Tabulate the coefficients of the fitted logistic regression model.
    d. Propose a ranking of the features (from most important to least important) based on their relationship with the class. *Hint: Remember the logit function or log-odds.*
    e. Rank the positively associated features with breast cancer from most associated to least associated.
    f. Rank the negatively associated features with breast cancer from most associated to least associated.
    g. For every one unit increase in "mean texture" how much percentage does the odds of getting breast cancer change? Indicate whether it increases or decreases and its corresponding percentage.
    h. For every one unit increase in "fractal dimension error" how much percentage does the odds of getting breast cancer change? Indicate whether it increases or decreases and its corresponding percentage.
    i. What is the "accuracy" of the fitted logistic regression model on the test set.
4. You are a data scientist tasked with estimating parameters for a model of your choosing using maximum likelihood estimation (MLE). **Note:** *As discussed, MLEs are applied not only to the simple distributions covered in class but also to linear regression, splines, and logistic regression. Therefore, you have many options to choose from for this exercise. There is no need to code this exercise; you only need to explain the concepts with pen and paper.*
    a. **Case Scenario**: Create a case scenario where you apply MLE to estimate parameters for a given dataset. Specify the type of data you are analyzing (e.g., heights, test scores, or sales data) and the underlying distribution you believe fits this data (e.g., normal, exponential, or binomial).
    b. **Likelihood Function**: Describe how you would formulate the likelihood function based on the chosen distribution. Explain the steps you would take to derive this function from your data.
    c. **Parameter Estimation**: Outline the process of finding the maximum likelihood estimates for the parameters of your chosen distribution. Include any necessary

Name: _____ Semester:_____ AY:_____

calculations or derivatives you would perform. If there is no analytical solution to the optimization problem, discuss how you would go about solving the problem.

d. **Interpretation of Results**: Discuss how you would interpret the estimated parameters in the context of your data. What insights could these estimates provide regarding the phenomenon you are studying?
e. **Practical Implications**: Reflect on how your findings could be applied in real-world scenarios. Consider any limitations of MLE and how they might affect your conclusions.