Homework 2

MA 506 Probability and Statistical Inference: Fall 2022

Due: October 19 (Wednesday), 11:59pm

100 points

Question 1: Exploratory Data Analysis(30 points)

For this problem we will be using the *wine* dataset. The dataset is available directly from sklearn (*load_wine*()). Here, just consider the 13 attributes (ignoring the class). The idea is to predict the *Alcohol* value using other 12 attributes. So, in essence you have 12 features and 1 target. For this dataset, do the following:

1. (5 points) Visualize the dataset as a pandas dataframe with proper column names. Here:
   - First 12 columns should represent the 12 features
   - 13^{th} column should be the alcohol value
2. (5 points) Plot scatterplots of alcohol value vs each of the features. Since there are 12 features, plot the 12 scatterplots in a grid of 3×4.
3. (5 points) Looking at the scatterplots above, which feature do you think will be most useful in predicting the alcohol value ? Which feature will be least helpful ? Explain.
4. (15 points) In your opinion, what information, the following observations will give regarding the alcohol content (increase or decrease). You can use any method of your choice to make this decision. Make sure to explain your reasoning
   - Increased value of Color intensity
   - Reduced value of Proline
   - Increased Magnesium but reduced Ash value

Question 2: Regression (70 points)

Now, just considering the feature 'proline' to predict the alcohol value

1. (10 points) Fit a straight line: $alcohol = \beta_0 + \beta_1 proline$. Plot this line over the scattered data. Compute the R^2 metric for this model. Please note that R^2 quantifies the quality of a regression model and is defined as

$$R^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

   Here y_i, are the observed y values, $\hat{y}$ are the corresponding predicted y values and $\bar{y}$ is the average of all y values. Additionally, n is the number of data points. Please note that R^2 values vary between 0 and 1 with 1 representing a perfect fit to the data points.

2. (20 points) Draw another plot, but this time with the above straight line model, also plot the graph of polynomials with degree 2,3,4 and 5 fitted over the same scattered data (alcohol vs proline). The graph legend should include the R^2 value of all 5 models.

3. (10 points) Looking at the 5 models in previous part, which model do you think is the best. Explain. You can use any criteria of your choice to choose the best model.

4. (10 points) Consider just the quadratic polynomial model ($alcohol = \beta_0 + \beta_1 proline + \beta_1 proline^2$), plot the quadratic curve fitted to this data along with 95% t-confidence interval and 95% t-prediction interval. Use different colors and opaqueness to improve the visibility of the different intervals. Why do you think the confidence interval is contained in the prediction interval ?

5. (20 points) With the same *alcohol vs proiline* data, plot a figure with 6 subplots. Here subplot 1 to 6 should show polynomial model of degree 1 to 6 respectively with corresponding 95% t-confidence intervals. Comment on the thickness of these intervals. For which model is the confidence interval most broad ? What information does this plot provide you ?