ECE M148 | Homework 2
Introduction to Data Science | Due: April 19, 12:00 p.m
Instructor: Lara Dolecek | TAs: Harish G V, Jayanth Shreekumar

**Please upload your homework to Gradescope by April 19, 12:00 PM.
You can access Gradescope directly or using the link provided on BruinLearn.
You may type your homework or scan your handwritten version. Make sure all the work is discernible.**

1. Recall the KNN Regression on the data set from homework 1.

   Compute the root mean squared error (RMSE)) on the training data in homework 1 for each choice of K for $K = 1$, 2, 3, and 6. What choice of K would you use? Now, consider the following test data points $A = \{(x, y)\} = \{(1.25, 2), (3.4, 5), (4.25, 2.5)\}$. Using the KNN Regression model trained in homework 1, perform regression on the points in A and calculate the test RMSE for each choice of K for $K = 1$, 2, 3, and 6. Does your choice of K change now based on this new test data RMSE?

2. Suppose we have the following data points with coordinates $(x, y) : \{(1, 1), (2, 2), (3, 3), (4, 3.5)\}$.

   (a) Suppose you want to fit the model $Y = \beta_0 + \beta_1 \times X$ by minimizing the mean square error (MSE) $\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 \times x_i)^2$. Write down the conditions for the derivative of the MSE that is necessary for β_0, β_1 to be optimal. From the conditions on the derivative, derive the formulae for β_0 and β_1. You do not need to re-derive the exact equations shown in class but you must derive a closed form solution for β_0 and β_1 in terms of the data (x, y).

       Hint: The following equalities may prove useful $\sum_{i=1}^{n}(\overline{x})^2 - \overline{x}x_i = 0$ and $\sum_{i=1}^{n} \overline{y}\,\overline{x} - y_i\overline{x} = 0$ where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

   (b) Fit the model $Y = \beta_0 + \beta_1 \times X$ based on the given data points by minimizing MSE. Compute R^2 for this model and briefly explain the meaning of the parameter β_1.

3. In class, we learned about one hot encoding.

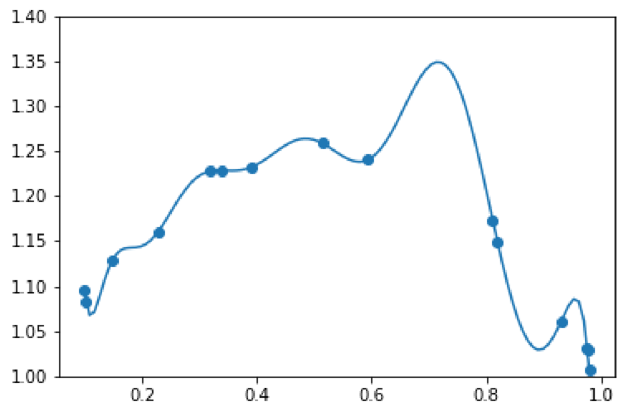
   (a) Explain what is one hot encoding and where it can be used.

   (b) Consider a housing dataset that contains information about homes in California. Briefly justify if one hot encoding is appropriate for the following example data features in the housing data:

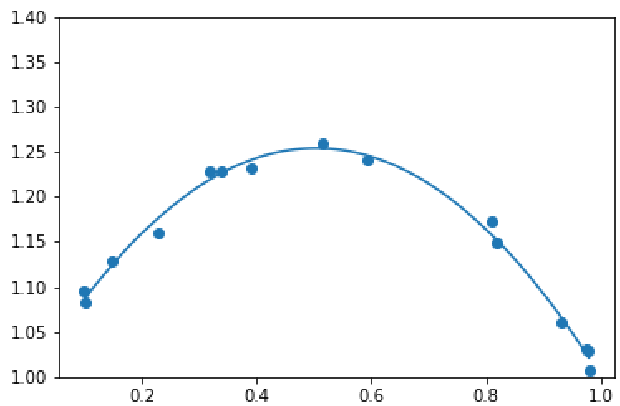
      (i) Zipcode of the house
      (ii) Price of the house
      (iii) City of the house
      (iv) Name of homeowner (assume each homeowner owns only one home)
      (v) Year the house was built

4. For each of the following plots, decide if the model is overfitted, underfitted, or it provides a good fit.

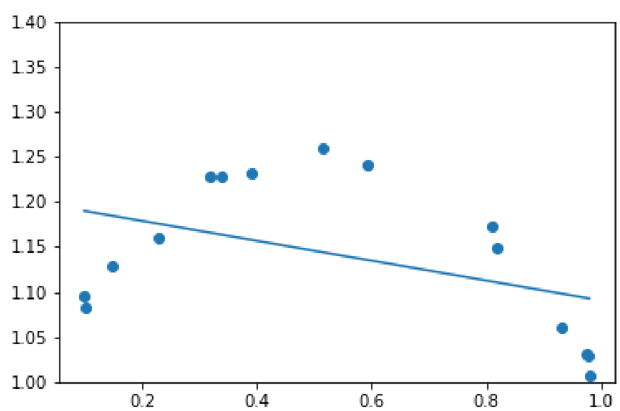
   (a) Example 1



   (b) Example 2



   (c) Example 3



5. True and False questions. For each statement, decide whether the statement is True or False and provide justification (full credit for the correct justification).

   (a) We can solve the problem of linear regression by trying all possible values for the model parameters and select the ones that minimize the MSE.

   (b) We can detect that a model is over-fitting when the training error is larger than the testing error.

   (c) For regression problems, R^2 is used as a measure of how much of the variability in the data is explained by the model and can never be greater than 1.

   (d) Multi-linear regression is a special case of polynomial regression.

   (e) KNN is more likely to overfit the data as K gets larger.