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

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?

   Solution:

   RMSE of Training Data:

   K=1: RMSE $= 0$

   K=2: RMSE ≈ 1.242

   K=3: RMSE ≈ 1.333

   K=6: RMSE ≈ 1.633

   From the training data, $K = 1$ is the best.

   The RMSE of the Test Data:

   K=1: RMSE ≈ 0.866

   K=2: RMSE ≈ 0.408

   K=3: RMSE ≈ 1.2361

   K=6: RMSE ≈ 1.322

   From the test data, $K = 2$ is the best which is different than the one we optimized by purely looking at the training data.

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.

Solution:

(a) To solve this problem, we differentiate the MSE in terms of the parameters and find the parameter values that cause the derivative of the MSE to be zero.
Let $L = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 \times x_i)^2$. For β_0, β_1 to be optimal, the following conditions must hold:

$$\frac{dL}{d\beta_0} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{dL}{d\beta_1} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

Now, we set the derivatives to zero and solve for the parameters:

$$\frac{dL}{d\beta_0} = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\implies \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\implies \beta_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i) = \overline{y} - \beta_1\overline{x}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
Now, we solve for β_1.

$$
\begin{aligned}
\frac{dL}{d\beta_1} &= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \\
&\implies \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \\
&\implies \sum_{i=1}^{n}(y_i - (\overline{y} - \beta_1\overline{x}) - \beta_1 x_i)x_i = 0 \\
&\implies \sum_{i=1}^{n}(y_i x_i - \overline{y}x_i + \beta_1\overline{x}x_i - \beta_1 x_i^2) = 0 \\
&\implies \sum_{i=1}^{n}(y_i x_i - \overline{y}x_i) - \beta_1\sum_{i=1}^{n}(x_i^2 - \overline{x}x_i) = 0 \\
&\implies \beta_1 = \frac{\sum_{i=1}^{n}(y_i x_i - \overline{y}x_i)}{\sum_{i=1}^{n}(x_i^2 - \overline{x}x_i)}
\end{aligned}
$$

Now, it is fine to stop here since this gives us a closed form solution but we can go a little further to relate the parameters to well known functions in probability theory.

First, note that $\sum_{i=1}^{n}(\overline{x})^2 - \overline{x}x_i = 0$ and $\sum_{i=1}^{n}\overline{y}\,\overline{x} - y_i\overline{x} = 0$. Thus,

$$
\begin{aligned}
\beta_1 &= \frac{\sum_{i=1}^{n}(y_i x_i - \overline{y}x_i)}{\sum_{i=1}^{n}(x_i^2 - \overline{x}x_i)} \\
&= \frac{\sum_{i=1}^{n}(y_i x_i - \overline{y}x_i) + \sum_{i=1}^{n}(\overline{y}\,\overline{x} - y_i\overline{x})}{\sum_{i=1}^{n}(x_i^2 - \overline{x}x_i) + \sum_{i=1}^{n}((\overline{x})^2 - \overline{x}x_i)} \\
&= \frac{\sum_{i=1}^{n}(y_i x_i - \overline{y}x_i - y_i\overline{x} + \overline{y}\,\overline{x})}{\sum_{i=1}^{n}(x_i^2 - 2\overline{x}x_i + (\overline{x})^2)}
\end{aligned}
$$

By completing the squares, we get

$$
\begin{aligned}
\beta_1 &= \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}
\end{aligned}
$$

Hence, we can see that the denominator is the empirical estimate of the variance of x (i.e. $Var(X)$ where X is the random variable representing the input data x) and that the numerator is the empirical estimate of the covariance between x and y (i.e. $Cov(X, Y)$ where X and Y are random variables).

(b) By using the equations in part (a), we get $\beta_1 = 0.85$ and $\beta_0 = 0.25$. With this, we get $R^2 \approx 0.9792$.

One interpretation of β_1 is that it measures how the average change in the independent variable x linearly affects the dependent variable y.

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

   Solution:

   (a) One hot encoding is the process of converting categorical data into binary columns where each column represents each of the categories. When converting the categorical data point into the binary columns, a single 1 is placed in the column representing the category of the data point and the rest are set to 0.

   This process is used to encode categorical data into real valued data that a ML model can then use. This technique should be used when:

   - The categorical data does not have an order since one hot encoding does not preserve this. An example of such a categorical data is age.
   - The number of categories is much smaller than the dataset. For example, if each data point was given an id number, this would not be something we would want to perform one hot encoding on since no two data points have the same id.

   (b) (i) Zipcode of the house: **Appropriate**
   Zipcode data groups a lot of houses together and, in general, there is no order to zipcode data.

   (ii) Price of the house: **Not Appropriate**
   This is real valued data

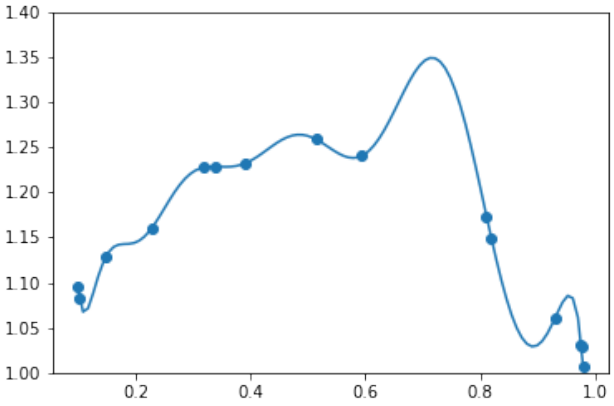
   (iii) City of the house: **Appropriate**
   There are much fewer cities than there are houses so this is appropriate.

   (iv) Name of homeowner (assume each homeowner owns only one home): **Not Appropriate**
   Since each name only belongs to one data point, there is no point to one-hot encode it.

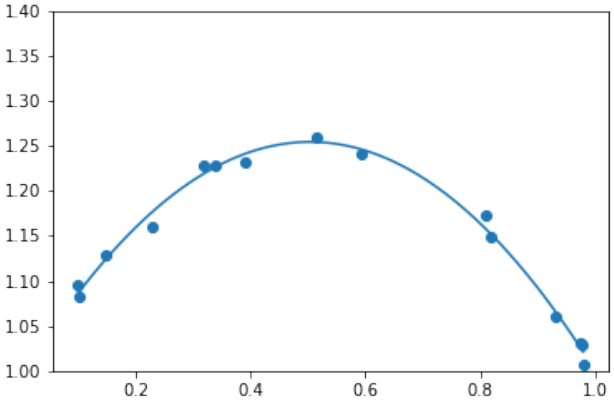
   (v) Year the house was built: **Not Appropriate**
   Years have an order to them and are real valued numbers

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

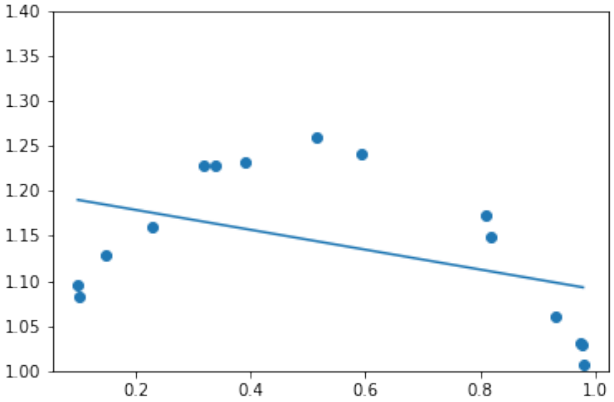
   (a) Example 1



   (b) Example 2



   (c) Example 3



Solution:

(a) This model looks very overfitted since it tries to capture every point on the training data and seems to be trying to fit the noise as well.

(b) This model looks like a good fit because it gets the general shape of the curve while not fluctuating to the slight discrepancies caused by noise..

(c) This is a very under fitted model since it is not even close to any of the points.

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.

   Solution:

   (a) **False**. Since the model parameters are real valued, there is an infinite number of choices which makes it infeasible to check by brute force.

   (b) **False**. Overfitting happens when testing error is much larger than training error. This happens because the model is overfitting to the training data which results in low training error and does not generalize well which results in high testing error.

   (c) **True**: That is one interpretation of R^2 and it can never be greater than 1 since $R^2 = 1 - a$ where a is the ratio of the sum of squares between the model and mean and, thus, can never be negative.

   (d) **False**: Multi-linear regression models the dependent variable y as a linear function of the inputs $\{x_i\}_{i=1}^n$, i.e $y = \sum_{i=1}^n \beta_i x_i + \beta_0$. Polynomial regression models the dependent variable y as a linear function of the powers of the independent variable x, i.e $y = \sum_{i=1}^n \beta_i x^i + \beta_0$. Clearly, polynomial regression is a special case of multi-linear regressions by setting $x_i = x^i$.

   (e) **False**: As k increases, KNN converges to the mean which is a low complexity model which underfits. For small k, the model is very susceptible to outliers and can overfit to the training data.