Assignment 1

Statistical Learning – Fall 2020

Assignment Date: 1399/07/24 Due Date: 1399/08/04

1) **Q1 & Q2 of chapter 2.**

2) **Q5 & Q7 of chapter**

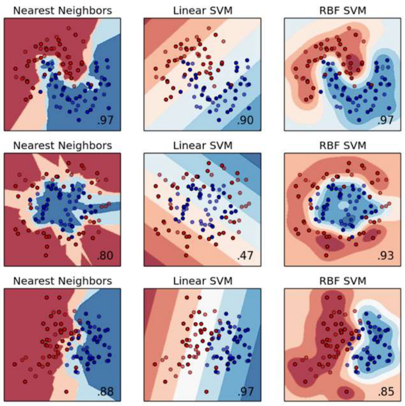
3) **Bias-variance tradeoff I:**

(a) Create a synthetic dataset (with both features and targets). Use the `make_moons` module with the parameter `noise=0.35` to generate 1000 random samples.

(b) Scatterplot your random samples with each class in a different color

(c) Create 3 different data subsets by selecting 150 of the 1000 data points at random. For each of these 150-sample datasets, fit three k-Nearest Neighbor classifier with: $k \in \{4, 25, 50\}$. This will result in 9 combinations (3 datasets, with 3 trained classifiers).

(d) For each combination of dataset trained classifier, in a 3-by-3 grid, plot the decision boundary (similar in style to Figure 2.15 from *Introduction to Statistical Learning*). Each column should represent a different value of k and each row should represent a different dataset.
Something like this but all of them are KNN, rows show different dataset, columns show different **k.**



(e) What do you notice about the difference between the rows and the columns. Which decision boundaries appear to best separate the two classes of data? Which decision boundaries vary the most as the data change?

(f) Explain the bias-variance tradeoff using the example of the plots you made in this exercise.

4) **Model Selection:** which “k” leads to the best model?

- **(a)** Use `make_moons` with same parameters of part “a” to create a new set of 500 random samples. Consider this as the test test while the 1000 samples of part “a” is your train dataset
- **(b)** Use train dataset to train a kNN classifier for $k = 1, 2, \ldots, 500$. For each “k” compute the test and train error and make a figure like Fig. 2.17 of the ISLR book.
- **(c)** What values of k represent high bias and which represent high variance?
- **(d)** What is the optimal value of k and why?

5) **Regression model:**
you can *use scikit-learn* *LinearRegression* *module.*

- **(a)** You have the Train and test data below. Visualize them using a scatter plot.
- **(b)** Fit the best linear regression model for the training data. Report the model coefficients and both the R^2 value and mean square error for the fit of that model for the training data.
- **(c)** To improve the model, first look at the scatter plot of the data and decide what type of non-linear parameter might be helpful. Use extended version of the linear model, i.e., $y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n$ where x_i could be any function of x, like $\frac{1}{x}$, $\log x, x^2$. Find the best parameters and also report the R^2 and mean square error of the fit for the training data.
- **(d)** Add the two curves of model in (b) and (c) to the scatter plot of part (a). The figure will be similar to Fig. 3.8 of ISLR.
- **(e)** Use both model and apply them hem to the test data and estimate the R^2 and mean square error of the test dataset.
- **(f)** Which models perform better on the training data, and which on the test data?

```
x_train = [3.19,9.26,9.38,8.77,7.91,3.79,3.18,7.61,2.36,6.26,6.62,1.53,6.2
5,7.93,7.07,4.58,4.14,2.14,9.04,4.56,3.99,6.71,2.51,0.84,6.13,5.22,0.25,3.
60,1.36,5.59,4.81,1.14,0.36,2.31,1.37,5.86,4.23,9.48,2.26,0.77,4.33]
y_train = [46.40,172.16,209.00,203.31,82.88,62.57,14.38,177.00,8.01,82.35,
84.84,-5.59,54.96,167.17,83.09,-21.63,94.64,63.97,106.57,38.99,88.26,66.99
,-11.12,-0.44,65.08,61.47,-0.61,23.85,10.55,83.69,54.35,51.82,-18.63,1.98,
4.90,55.44,50.09,155.66,45.57,18.12,30.58]

x_test = [5.65,0.07,8.84,5.14,6.65,1.35,5.45,7.39,3.35]
y_test = [98.52,16.09,198.45,75.90,85.11,47.64,14.76,141.03,-39.13]
```