ECE M148 Homework 1
Introduction to Data Science Due: April 12, 12:00 PM
Instructor: Lara Dolecek TAs: Harish GV, Jayanth Shreekumar

Please upload your homework to Gradescope by April 12, 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. Consider the following data set $A = \{1, 1, 5, 9, 9\}$. What are the mean and median of A? Now, consider $B = \{1, 1, 5, 9, 9, 11\}$. What are the mean and median of B? Using the mean and median, compare A and B.

A: mean: $\frac{1+1+5+9+9}{5} = \frac{25}{5} = 5$

median: 5

B: mean: $\frac{1+1+5+9+9+11}{6} = \frac{36}{6} = 6$

median: $\frac{5+9}{2} = \frac{14}{2} = 7$

Set A has the same mean & median.
Set B has different mean & median.
Then set A has a perfectly symmetrical distribution while set B does not.
Looking at A and B, we see adding 11 to set A increases the mean by 1 and increases the median by 2.

2. In class, we discussed different ways to sample data. Explain in 1-2 sentences each the advantages and disadvantages of:

   (a) Random sampling
   (b) Stratified sampling
   (c) Systematic sampling
   (d) Cluster sampling

	Advantages	Disadvantages
Random Sampling	This is a uniform sample, it is simple and convenient.	May not equally represent certain subgroups of the sample. It is difficult to get data for the entire population.
Stratified Sampling	This allows for more balanced representation of pre-determined groups.	Pre-grouping may be expensive & complicated.
Systematic Sampling	There is randomness at the beginning and is simple, cheap.	May skip entire groups of individuals, resulting in over/ under representation.
Cluster sampling	After clustering, choosing random samples can be cheap & convenient.	The creation of clusters may result in bias, requires size equality to be effective.

3. As discussed in class, many real-world datasets will contain missing or null values in the data. List four different strategies you could reasonably use to address null values. For each, clarify what the advantages and disadvantages to it are.

	Advantages	Disadvantages
Delete data	Most simple and straightforward when removing incomplete data.	May delete relevant data and/or result in too little data.
Map to the closest point using some feature	Allows you to fill in incomplete data to avoid deletion.	More complicated than deleting incomplete data. May repeat data, leading to over/under representation.
populate using a random choice from the dataset	Allows you to fill in incomplete data to avoid deletion. Less complicated than closest point / statistics.	May repeat data, leading to over/under representation. May lead to inaccurate results.
populate based on statistics such as mean, median, mode, etc.	Allows you to fill in incomplete data to avoid deletion.	More costly to calculate statistics. Depending on what statistic is used, may be prone to misrepresentation due to outliers if using mean.

4. Consider the following sampling scenarios and determine which type of sampling bias is being demonstrated and explain your answer.

   (a) Bob is a wealthy CEO who thinks taxes are too high. To confirm this hypothesis, he asks all his wealthy CEO friends their opinion.

   (b) Sally is a teacher who wants to know how her class is performing. She sends out a survey with the following question: "Do you feel like you will get an A in the course or are you failing?"

   (c) Constantine wants to know people's opinion about his website. He posts a survey link on his website asking for responses.

   You may choose among the following options for the type of bias:

   i) Response Bias : answer affected by how the question is worded
   ii) Voluntary Bias : only those who volunteer are represented
   iii) Convenience Bias : select people who are convenient/available
   iv) Under-coverage Bias : don't include all types of people
   v) Over-coverage Bias : ask too many people, ask people more than once
   vi) Non-response bias : certain people don't participate

a) under-coverage bias, since only wealthy CEO friends are represented, his survey doesn't include all types of people.

b) Response bias, since the question presented is at the two extremes, thus the results are biased.

c) Voluntary bias, since not everyone is required to respond to the survey & it relies on people to volunteer to take the survey. This may result in bias.

14 18 $\frac{18}{6} = 3$
1+4+6+3+2+2

5. Perform KNN Regression on the following data set for different values of K: $(x, y) = \{(1,1), (2,4), (3.2,6), (4,3), (5,2), (6,2)\}$. Start by plotting the given points on a 2-D grid and then fitting a KNN regressor for the different values of K:

   Make sure to draw the regression plot from 0 to 7.

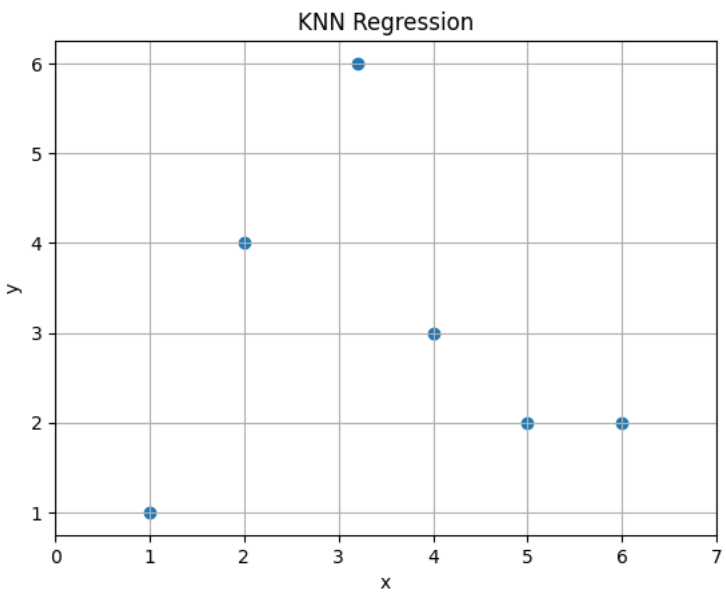
   - K = 1
   - K = 2
   - K = 3
   - K = 6

   Contrast and compare your findings over various choices of K. Is a larger K always better? Is K = 1 always better? Why or why not? Comment on what you think about the KNN performing regression on all $x < 1$.

Larger K is not always better, evident by K=6 case. Observe underfitting in K=6 case.
K=1 is not always the best, as it is more prone to overfitting. With K=1, it is less resilient to noise.
KNN performance on all $x < 1$ will be constant at the value of KNN's result at X=1. KNN cannot capture data trends beyond the given domain of points.

See the next page for code & plot.

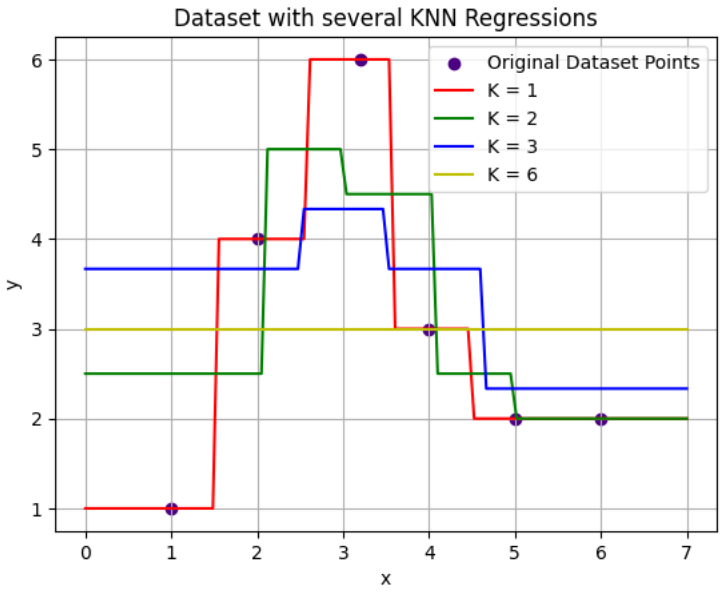
```
from matplotlib import pyplot as plt
from sklearn.neighbors import KNeighborsRegressor
import numpy as np


x = [[1], [2], [3.2], [4], [5], [6]]
y = [1, 4, 6, 3, 2, 2]
plt.scatter(x, y)
plt.grid()
plt.title("KNN Regression")
plt.xlabel("x")
plt.ylabel("y")
plt.xlim([0, 7])
plt.show()
```



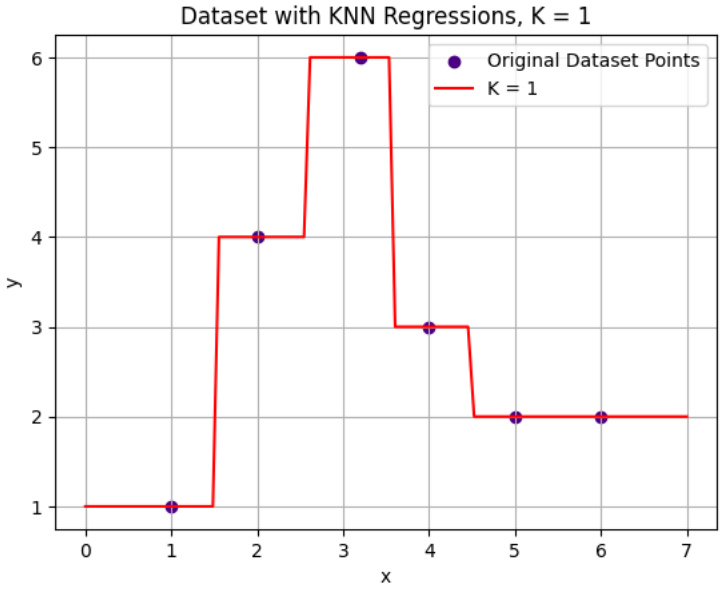

```
k1 = KNeighborsRegressor(n_neighbors = 1)
k1.fit(x, y)
k2 = KNeighborsRegressor(n_neighbors = 2)
k2.fit(x, y)
k3 = KNeighborsRegressor(n_neighbors = 3)
k3.fit(x, y)
k6 = KNeighborsRegressor(n_neighbors =6)
k6.fit(x, y)
```

```
▾        KNeighborsRegressor
KNeighborsRegressor(n_neighbors=6)
```

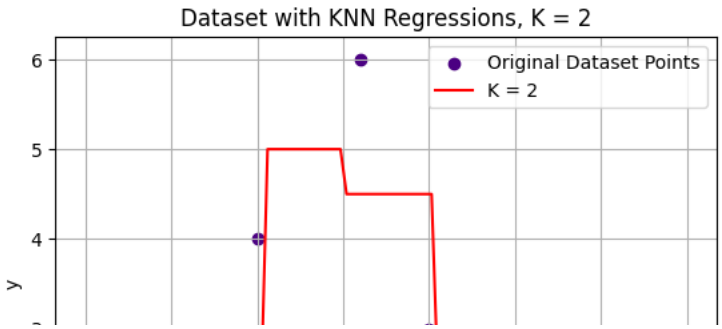
```
pred_y = np.linspace(0, 7, 100)
k1list = []
k2list = []
k3list = []
k6list = []

for i in range (len(pred_y)):
  k1list.append(k1.predict([[pred_y[i]]]))
  k2list.append(k2.predict([[pred_y[i]]]))
  k3list.append(k3.predict([[pred_y[i]]]))
  k6list.append(k6.predict([[pred_y[i]]]))


plt.title("Dataset with several KNN Regressions")
plt.xlabel("x")
plt.ylabel("y")
plt.scatter(x, y, color = "indigo", label = "Original Dataset Points")
plt.plot(pred_y, k1list, label = "K = 1", color = "r")
plt.plot(pred_y, k2list, label = "K = 2", color = "g")
plt.plot(pred_y, k3list, label = "K = 3", color = "b")
plt.plot(pred_y, k6list, label = "K = 6", color = "y")
plt.legend()
plt.grid()
plt.show()
```



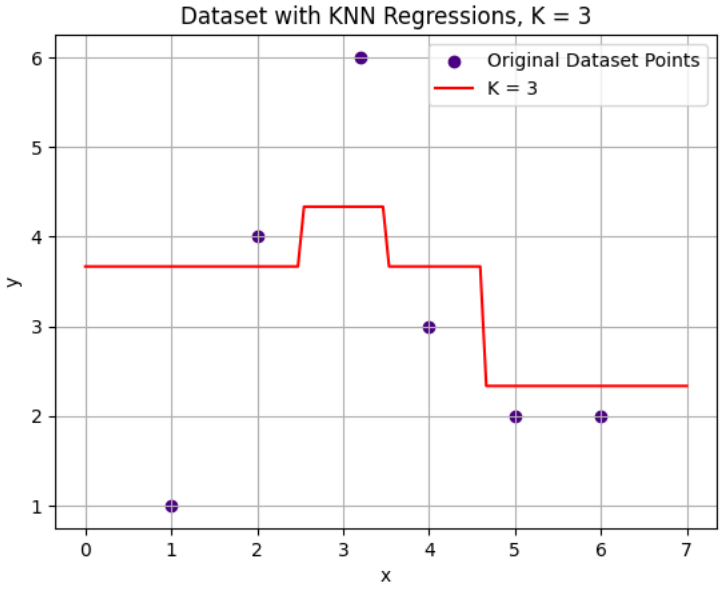

```
plt.title("Dataset with KNN Regressions, K = 1")
plt.xlabel("x")
plt.ylabel("y")
plt.scatter(x, y, color = "indigo", label = "Original Dataset Points")
plt.plot(pred_y, k1list, label = "K = 1", color = "r")
plt.legend()
plt.grid()
plt.show()
```




```
plt.title("Dataset with KNN Regressions, K = 2")
plt.xlabel("x")
plt.ylabel("y")
plt.scatter(x, y, color = "indigo", label = "Original Dataset Points")
plt.plot(pred_y, k2list, label = "K = 2", color = "r")
plt.legend()
plt.grid()
plt.show()
```




```
plt.title("Dataset with KNN Regressions, K = 3")
plt.xlabel("x")
plt.ylabel("y")
plt.scatter(x, y, color = "indigo", label = "Original Dataset Points")
plt.plot(pred_y, k3list, label = "K = 3", color = "r")
plt.legend()
plt.grid()
plt.show()
```




```
plt.title("Dataset with KNN Regressions, K = 6")
plt.xlabel("x")
plt.ylabel("y")
plt.scatter(x, y, color = "indigo", label = "Original Dataset Points")
plt.plot(pred_y, k6list, label = "K = 6", color = "r")
plt.legend()
plt.grid()
plt.show()
```