ECE M148
Introduction to Data Science
Instructor: Lara Dolecek

Homework 1 Solutions
Due: April 12, 12:00 PM
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.

   Solution: Mean of $A = 5$, Median of $A = 5$ Mean of $B = 6$, Median of $B = 7$

   Since the mean and median are the same, we can see that A is possibly symmetric and has few outliers. Conversely, since B has different mean and median, it is more skewed and might contain outliers.

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

   Solution:

   (a) Random sampling: Every member of the population is equally likely to be sampled.

   Advantage: Simple to understand and reduces the impact biases

   Disadvantage: Has a chance to not select a diverse group of samples if only the majority is sampled

   (b) Stratified sampling: Divide the population into groups that may differ in significant ways and then sample from these groups

   Advantage: Ensures diversity of samples and allows analyses of each group separately

   Disadvantage: Requires categorizing every member of the population and determining the characteristics to classify them by. Additionally, some people may fit in multiple groups which isn't captured by the current modeling

   (c) Systematic sampling: Sampling members of a population in regular intervals according to some ordering

   Advantage: If population ordering is random, then systematic sampling can create data samples that are representative of the population

   Disadvantage: If population ordering is not random, then systematic sampling may create a bias in the members sampled

   (d) Cluster sampling: Divide the population into groups that have similar distribution to the whole population and then pick a whole group as a sample.

   Advantage: Can be less costly if members within a group can be sampled with the same method. For example, it is easier to interview everyone in ucla than it is to interview a random sample of people across LA. Clusters can be chosen to be very large and can thus allow for a larger sample size

   Disadvantage: Can have large sampling error if groups are not representative of the whole population

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.

   Solution:

   There are many different solutions to this problem. We simply list a few examples here.

   (a) Delete rows/samples

   Advantage:Useful if the target/label data is missing and we are trying to perform a supervised learning task. Also, useful if the row has many missing elements across columns and thus would not provide much information.

   Disadvantage: Problematic to remove samples if null/missing values are dependent on the sample. Thus, we would be losing information. Additionally, if a large portion of the dataset has null values, deleting them would significantly reduce the number of samples.

   (b) Removing columns with null values

   Advantage:Useful if the null values in the column are a majority which makes imputation difficult. Additionally, if the column is already highly correlated with another column, then losing it would cause little loss of information.

   Disadvantage: Similar disadvantages as deleting rows.

   (c) Assign null values with mean/median

   Advantage: If data is concentrated around the mean or median like for a Gaussian, then using mean/median would cause minimal loss in information and would allow us to use the samples in our model.

   Disadvantage: Can bias the distribution of the data if data is not concentrated close to the mean or median which can impact the results of the model

   (d) Fill in missing values by randomly sampling the column

   Advantage: If column data is independent or close to independent from other feature data, then this is a fair approximation of the missing value

   Disadvantage: Can bias the distribution of the data if column is not independent from other columns

   (e) Fill in missing values with values in rows/samples that are similar to this row/sample

   Advantage: Tries to create a sample data that can close map the original data distribution

   Disadvantage: Difficult to define an appropriate measure of closeness and relies on the dataset to have sufficient number of samples such that a similar one does exist

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

   ii) Voluntary Bias

   iii) Convenience Bias

   iv) Under-coverage Bias

   v) Over-coverage Bias

   vi) Non-response bias

   Solution:

   Each example can exemplify many types of biases. If you provide a sufficient justification for your designation, you will get full points. The following are possible interpretations of the examples:

   (a) This can be an example of either under-coverage bias or convenience bias since Bob is not going out of his way to sample a diverse portion of the populace and is focusing on individuals that he has easy access to.

   (b) This is an example of response bias since the question clearly misses the middle ground of people that are passing with a B or C.

   (c) This can be seen as a voluntary bias since only people with strong opinions will perform the survey. Additionally, this is also an example of under-coverage bias since only people that will visit the website can respond to the survey.

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:

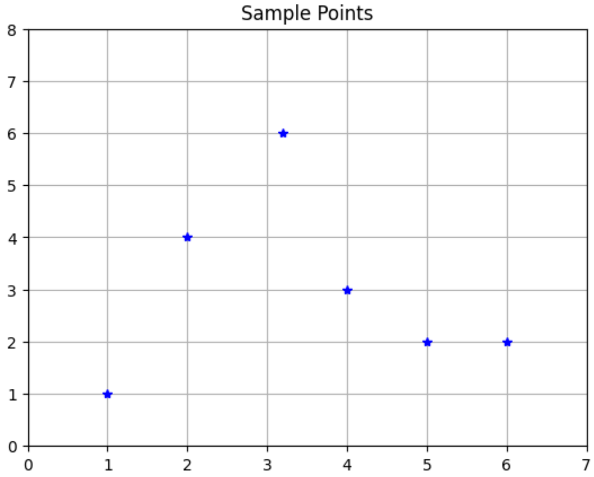
   Make sure to draw the regression plot from 0 to 7.

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

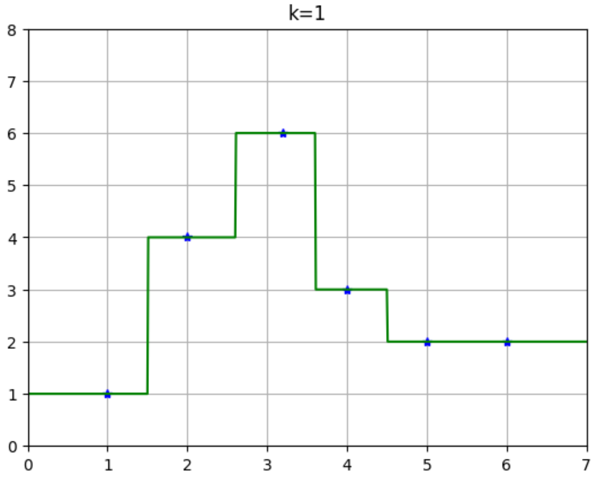
   Contrast and compare your findings over various choices of K. Is a larger K always better? Is K = 1 always better? Why or why not? Comment on what you think about the KNN performing regression on all $x < 1$.

   Solution:

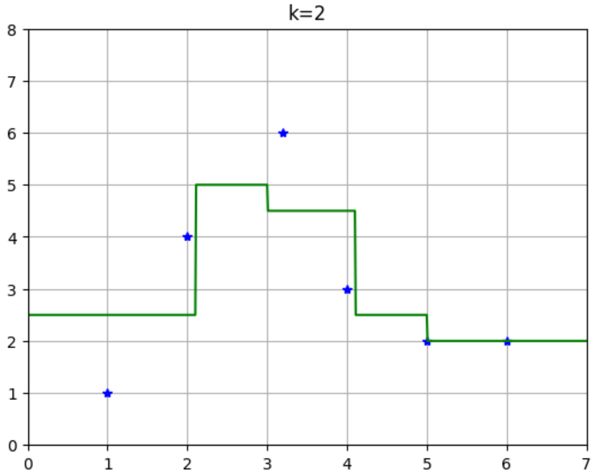
   - Plot of samples



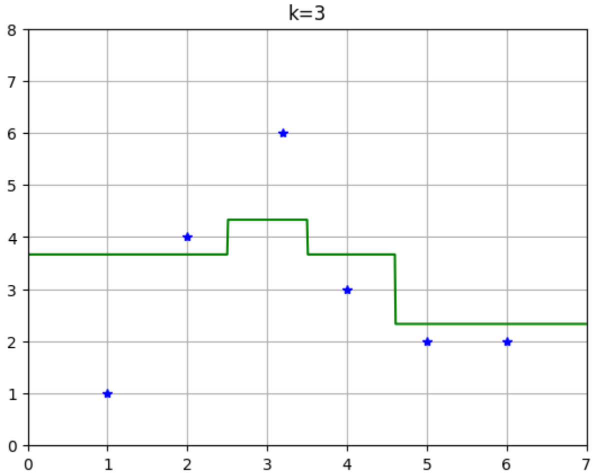
   - K = 1



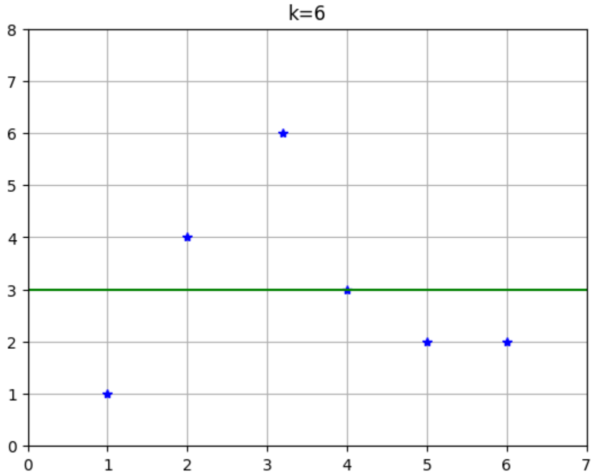
- K = 2



- K = 3



- K = 6



We can see that as K increases, the curve gets smoother and flatter. This is because as K increases, the regressor converges to the mean of the target data. Yet, for small values of K like K = 1, we see that the regressor can change frequently which makes it susceptible to noise. The KNN regressor almost always incorrectly classifies points in $x < 1$ as it does not have points before it to utilize for classification.

CODE

In[1]:

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

In[2]:

```
x = np.array([[1],
                [2],
                [3.2],
                [4],
                [5],
                [6]])
y = np.array([1,4,6,3,2,2])
plt.plot(x,y, "b*")
plt.xlim(0,7)
plt.ylim(0,8)
plt.grid()
plt.title("Sample Points")
plt.savefig("knn_base_ex.png")
```

In[3]:

```
p = np.arange(0,7,0.01)
for k in [1,2,3,6]:
    plt.figure()
    plt.plot(x,y, "b*")
    plt.xlim(0,7)
    plt.ylim(0,8)
    plt.grid()

    q =
     ↪ KNeighborsRegressor(n_neighbors=k).fit(x,y).predict(p.reshape(-1,
     ↪ 1))
    plt.plot(p,q, "g")

    plt.title(f"k={k}")
    plt.savefig(f"knn_k_{k}.png")
```