CMSE 381, Fundamental Data Science Methods

September 17, 2025

Homework 3

Lowell Monis

Instructor: Dr. Mengsen Zhang

Question 1: ISLP § 3.7.3

Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

The coefficients $\hat{\beta}_i$ give us the following model:

$$\text{salary} = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5$$

where $X_4 = \text{GPA} \times \text{IQ}$ and $X_5 = \text{GPA} \times \text{Level}$.

(a) Which answer is correct, and why?

1. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.
2. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.
3. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.
4. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

Let's consider case 1. We assume a fixed IQ and GPA. First, for a high school graduate, the level is set at 0. Thus, the model is updated to:

$$\text{salary}_\text{H} = 50 + 20X_1 + 0.07X_2 + 0.01X_4$$

The X_5 term is eliminated since it's an interaction term between the GPA and level, and level is set at 0.

Second, we consider a college graduate, for whom the level is set at 1. $X_5 = X_3 \times X_1 = 1 \times X_1 = X_1$. Thus the model is updated to:

$$\text{salary}_\text{C} = 50 + 20X_1 + 0.07X_2 + 35 + 0.01X_4 - 10X_1 = 85 + 10X_1 + 0.07X_2 + 0.01X_4$$

Upon considering the difference between the two models:

$$\begin{aligned}\text{salary}_\text{C} - \text{salary}_\text{H} &= (85 + 10X_1 + 0.07X_2 + 0.01X_4) - (50 + 20X_1 + 0.07X_2 + 0.01X_4) \\ &= 35 - 10X_1\end{aligned}$$

Thus, the sign of the difference, which implies who earns more, depends on the GPA.

If $35 - 10\text{GPA} > 0$, or if the difference is positive, GPA < 3.5, and the college graduate earns more.

If $35 - 10\text{GPA} < 0$, or if the difference is negative, GPA > 3.5, and the high school graduate earns more.

If $35 - 10\text{GPA} = 0$, or if the difference is zero, GPA $= 3.5$, and both earn the same.

Thus, for a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough. The third alternative is correct.

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0. Consider the model:

$$\text{salary} = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5$$

When $\text{IQ} = X_2 = 110, \text{GPA} = X_1 = 4.0, \text{Level} = X_3 = 1$, $X_4 = 110 \times 4.0 = 440$ and $X_5 = 4.0 \times 1 = 4$. Therefore:

$$\text{salary} = 50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(440) - 10(4.0) = 137.1$$

Therefore, for the given conditions, the salary of this individual is estimated to be around \$137,100.00 after graduation.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer. This statement is *false*. I can say this because while the coefficient is very small, that does not imply that there is no interaction effect at all. The coefficient is not enough information to determine the statistical significance of the interaction effect. The statement confuses between the existence or significance of the interaction effect, and strength/impact of the effect, which can be sort of determined by the coefficient. I can say that the effect of the coefficient is not dominant in the model when compared to the other predictors. I will need the p-value and the hypotheses, however, to determine whether there is evidence or the lack thereof of an interaction effect between GPA and IQ.

Question 2: ISLP § 2.4.7

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$. The Euclidean distances between each training observation and the test point are given as follows:

$$
\begin{aligned}
d_1 &= \sqrt{(0-0)^2 + (0-3)^2 + (0-0)^2} = \sqrt{9} = 3 \\
d_2 &= \sqrt{(0-2)^2 + (0-0)^2 + (0-0)^2} = \sqrt{4} = 2 \\
d_3 &= \sqrt{(0-0)^2 + (0-1)^2 + (0-3)^2} = \sqrt{10} = 3.162 \\
d_4 &= \sqrt{(0-0)^2 + (0-1)^2 + (0-2)^2} = \sqrt{5} = 2.236 \\
d_5 &= \sqrt{(0+1)^2 + (0-0)^2 + (0-1)^2} = \sqrt{2} = 1.414 \\
d_6 &= \sqrt{(0-1)^2 + (0-1)^2 + (0-1)^2} = \sqrt{3} = 1.732
\end{aligned}
$$

(b) What is our prediction with $K = 1$? Why? If $K = 1$, the one closest training observation to the test point, which will be the most common one too due to its singleton nature, will be observation number 5, $(-1, 0, 1)$, with Euclidean distance 1.414. Since this is also the most common nearest observation, the prediction with $K = 1$ will be Green.

(c) What is our prediction with $K = 3$? Why? If $K = 3$, the three closest training observations to the test point, will be observation numbers 5: $(-1, 0, 1)$, 6: $(1, 1, 1)$, and 2: $(2, 0, 0)$. With two reds and one green, red is the most common nearest observation. Thus the prediction with $K = 3$ will be Red.

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why? I would expect K to be small. This is because when the decision boundary is highly non-linear, we need more flexibility, which is provided by a small K value, allowing the K-NN decision boundary to adapt to the true Bayes decision boundary more easily. A large value of K will lead to a smooth decision boundary, and a sort of averaging-out effect to the non-linearities (K-NN uses majority voting and has less emphasis

on each individual observation). Essentially, the smaller changes in classes within an area get buried within the majority. But when we reduce the value of K, these smaller number of minority classes are highlighted more around their neighbors, and we will be able to see those specific intricacies in the decision boundary due to the ability of the boundary to capture localized non-linearities, because the majority can vary significantly when fewer neighboring observations are considered.

Question 3: ISLP § 4.8.14

Only selected parts need to be completed.

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

First, we load and clean the `Auto` data set.

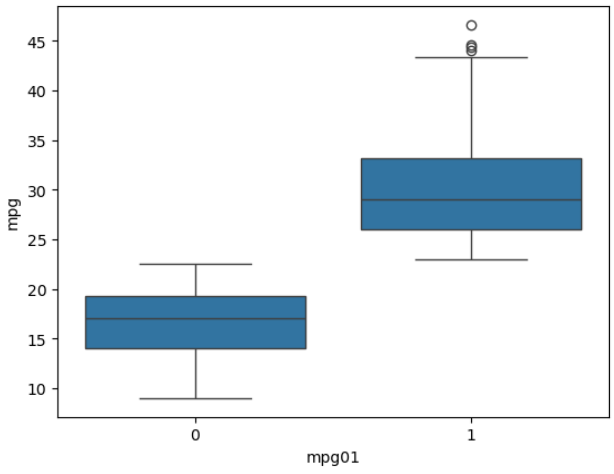
[2]:
```
auto = pd.read_csv('../data/Auto.csv')
auto=auto.replace('?', np.nan)
auto=auto.dropna()
auto['horsepower']=auto['horsepower'].astype('int')
auto=auto.reset_index(drop=True)
```

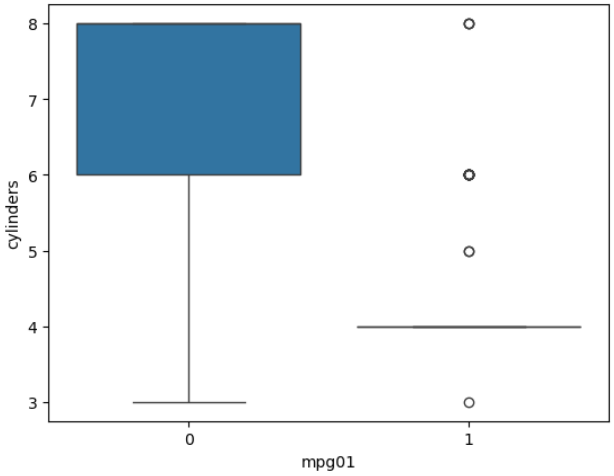
(a) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median.

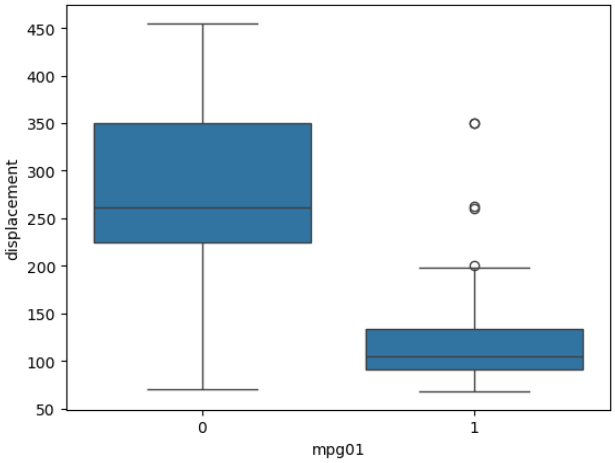
[3]:
```
auto['mpg01']=np.array([1 if i>auto.mpg.median() else 0 for i in auto.mpg])
```

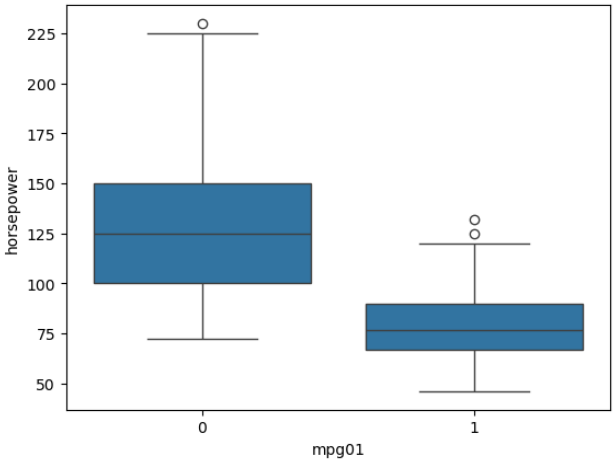
(b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

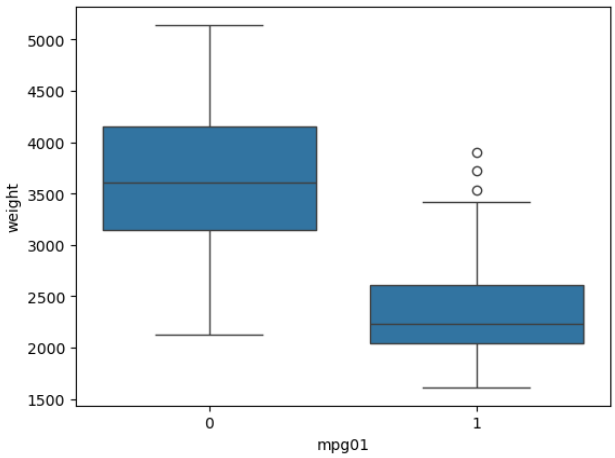
[4]:
```
for column in auto.columns.drop(['mpg01', 'name']):
    sns.boxplot(x=auto['mpg01'], y=auto[column])
    plt.show()
```

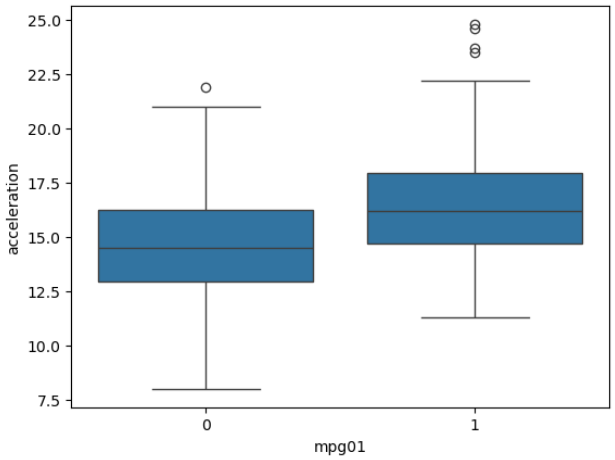


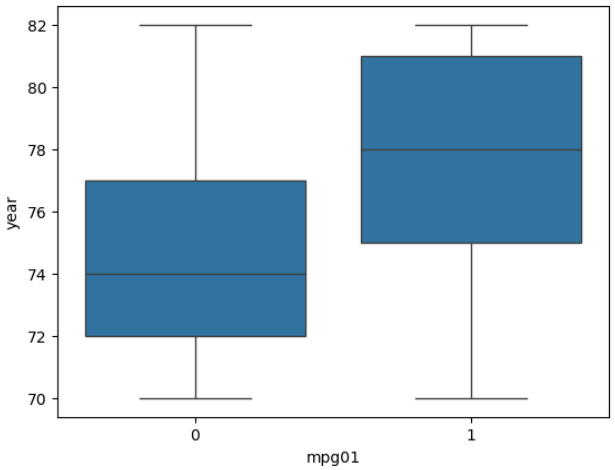

cylinders
8
7
6
5
4
3
0
1
mpg01

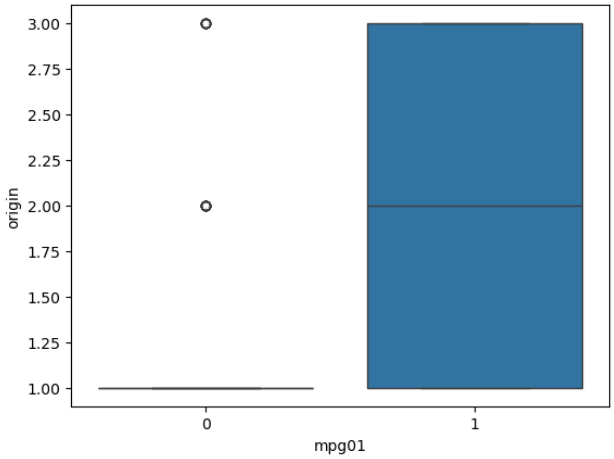

displacement
450
400
350
300
250
200
150
100
50
0
1
mpg01


horsepower
225
200
175
150
125
100
75
50
0
1
mpg01


weight
5000
4500
4000
3500
3000
2500
2000
1500
0
1
mpg01

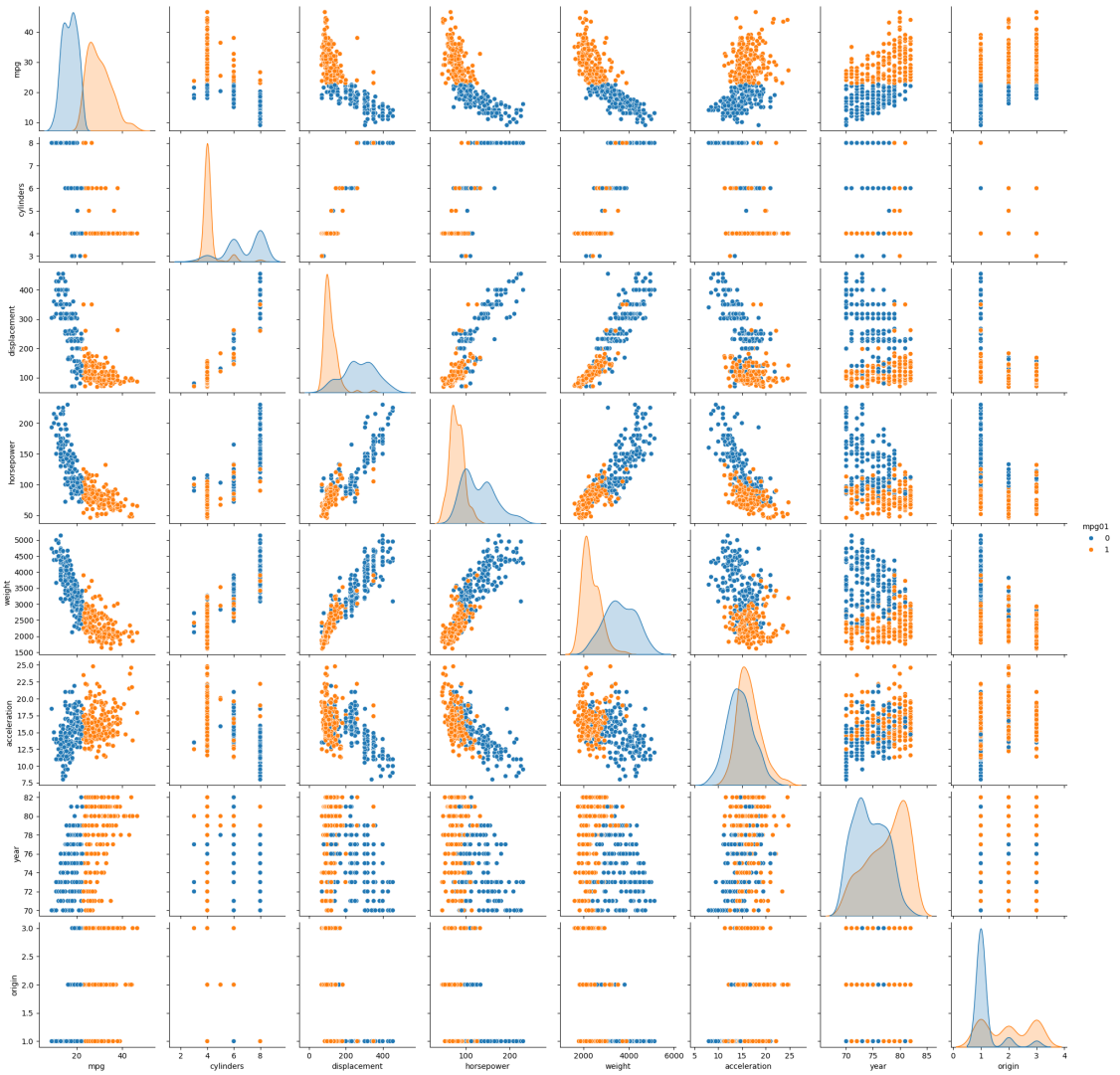

acceleration
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
0
1
mpg01


year
82
80
78
76
74
72
70
0
1
mpg01



[5]:
```
sns.pairplot(data=auto, vars=auto.columns.drop(['mpg01', 'name']), hue='mpg01')
```

[5]: <seaborn.axisgrid.PairGrid at 0x773370510740>



Considering we are working on a K-NN model in this question, ideally we need the classes to be distinctly distinguishable within the scope of the predictor being considered. Thus, we will look into predictors that closely imitate this ideal scenario of distinguishable classes. In other words, if there is a close-to-ideal boundary between the two classes, the predictor associated would be a great fit for the model.

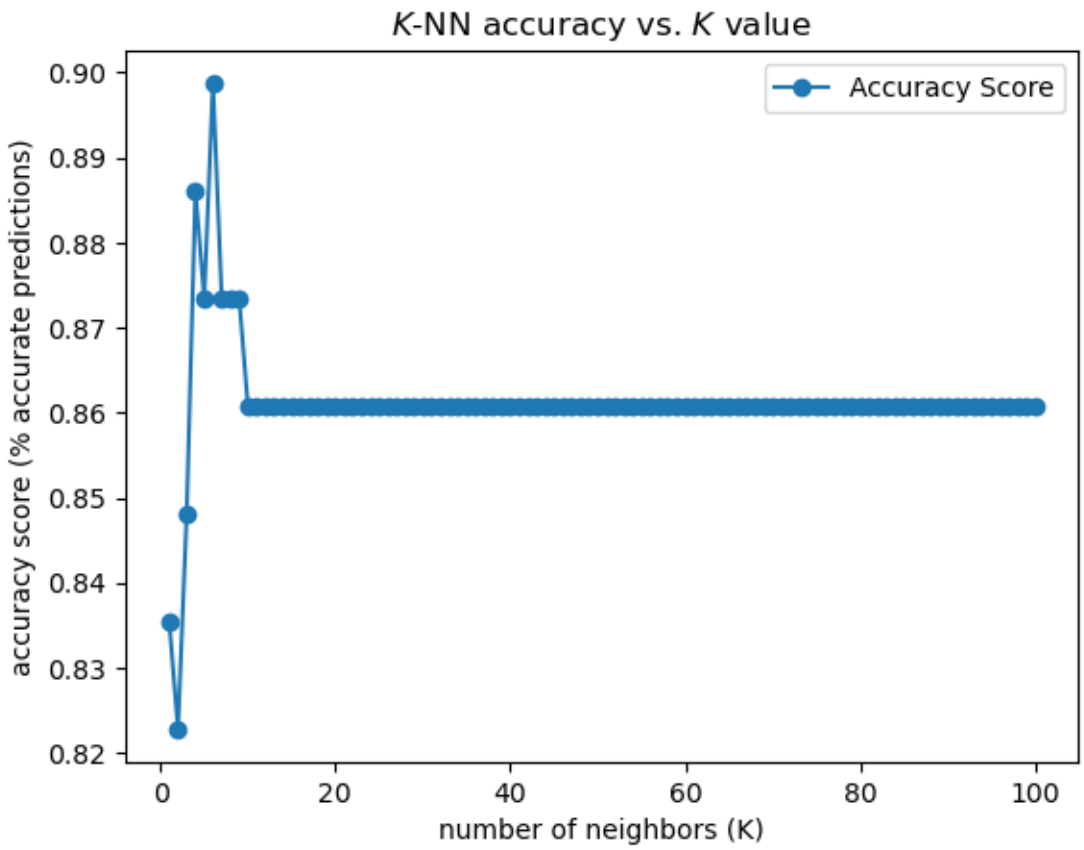
Upon shortlisting from the the scatter plots, and validating my choices using the box plots, I can conclude that `weight`, `displacement`, `cylinders`, and `horsepower` are great predictors. My method for validation was to verify if the interquartile ranges of both the 0 and 1 levels did not overlap. I did not consider outliers and the complete ranges of the data, since I am looking for ideal or close-to-ideal, and not perfect predictors.

(c) Split the data into a training set and a test set. For this model, I am splitting the data 80-20, with the seed set at 381. I am also normalizing the data.

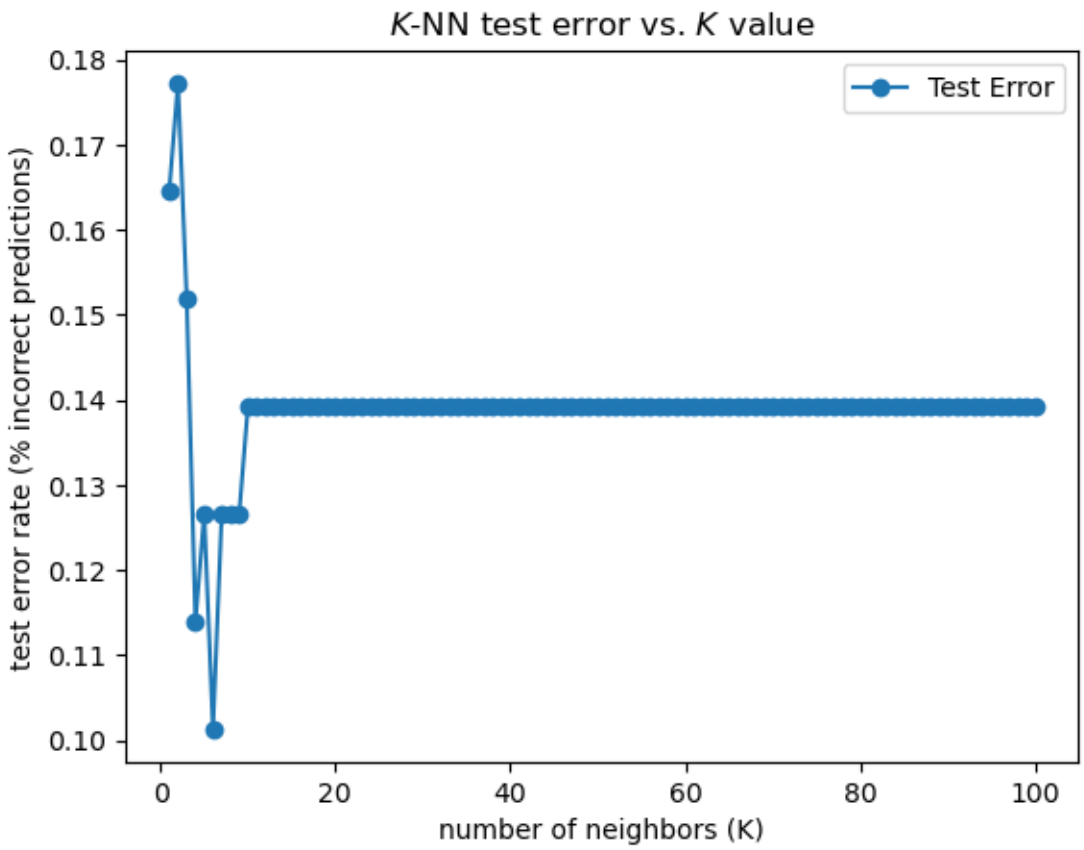
[6]:
```
X=auto[['weight', 'displacement', 'cylinders', 'horsepower']]
y=auto.mpg01
scaler = StandardScaler(with_mean=True, with_std=True, copy=True)
scaler.fit(X)
X_std = scaler.transform(X)
X_scaled = pd.DataFrame(X_std, columns=X.columns, index=X.index)
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled,
    y,
    test_size=0.2,
    random_state=381
)
```

(h) Perform KNN on the training data, with several values of K, in order to predict `mpg01`. Use only the variables that seemed most associated with `mpg01` in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set? I will be using a version of the code provided in ISLP § 4.7 to find the ideal value of K. However, I have modified the code to create a plot to compare accuracy across a range of K values from 1 to 100.

[7]:
```
accuracy_scores = []
for K in range(1, 101):
    knn = KNeighborsClassifier(n_neighbors=K)
    knn.fit(X_train, y_train)
    knn_pred = knn.predict(X_test)
    accuracy_scores.append(accuracy_score(y_test, knn_pred))
plt.plot(range(1, 101), accuracy_scores, label='Accuracy Score', marker='o')
plt.title('$K$-NN accuracy vs. $K$ value')
plt.xlabel('number of neighbors (K)')
plt.ylabel('accuracy score (% accurate predictions)')
plt.legend()
plt.show()
```



[8]:
```
test_errors=[]
for i in accuracy_scores:
    test_errors.append(1-i)
plt.plot(range(1, 101), test_errors, label='Test Error', marker='o')
plt.title('$K$-NN test error vs. $K$ value')
plt.xlabel('number of neighbors (K)')
plt.ylabel('test error rate (% incorrect predictions)')
plt.legend()
plt.show()
```



The test errors are demonstrated in the above line plot. They range from a little above 10%, all the way to a little under 18%. These fluctuations happen when the values of K are small, before the error stays constant at about 14% after reaching a sort of threshold for K. The lowest test error and the highest accuracy is associated with $K = 6$, leading me to conclude that $K = 6$ performs the best for this data.

Question 4: ISLP § 4.8.6

Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6$, $\beta_1 = 0.05$, $\beta_2 = 1$.

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class. Consider the model from ISLP § 4.3.4, Equation 4.7:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$$

We rewrite this model to incorporate the given coefficients and calculate the probability estimator:

$$\hat{p}(X) = \frac{e^{-6+0.05X_1+X_2}}{1 + e^{-6+0.05X_1+X_2}}$$

Finally, we plug in the given values for the student to estimate the probability. $X_1 = 40$, $X_2 = 3.5$:

$$\hat{p}(X) = \frac{e^{-6+0.05(40)+3.5}}{1 + e^{-6+0.05(40)+3.5}} = \frac{e^{-6+2+3.5}}{1 + e^{-6+2+3.5}} = \frac{e^{-0.5}}{1 + e^{-0.5}} \approx 0.37754$$

The probability of this student getting an A in this class is about 0.378, or approximately 38%.

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class? I will work backwards for this questions. Assuming all else is left at the same value, with the exception of X_1, let's set the probability $\hat{p}(x) = 0.5$. To make this easier, I will use the the logarithmic form of the model as in Equation 4.6 in ISLP (§ 4.3.4):

$$\log\left(\frac{\hat{p}(X)}{1 - \hat{p}(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

This can be rewritten with all coefficients and known values plugged in:

$$\log\left(\frac{0.5}{1 - 0.5}\right) = -6 + 0.05X_1 + 3.5$$

This can further be simplified to:

$$\log(1) = 0.05X_1 - 2.5$$

Since $\ln 1 = 0$:

$$0 = 0.05X_1 - 2.5 \implies 0.05X_1 = 2.5 \implies X_1 = \frac{2.5}{0.05} = 50$$

Thus, it can be conclude that for this student to have a 50% chance of getting an A in the class, they must study for 50 hours.

Collaborations and Acknowledgments

In many questions, I have made a direct reference to the equations and code available in the Introduction to Statistical Learning in Python text. I have made exact attributions to the section when I have made those references. In cases with code, all the code has been modified. No code in this text has been sourced verbatim from the text, and has been modified to my personal liking, or to my personal comfort and preference when it comes to the use of packages.