Jeffery Watson **DS_740**
Assignment Lesson_1 due 01/28/2023 at 11:59pm CST

Problem 1. **(1 point)**

For each of the following scenarios, choose which type of data mining method you would expect to perform better: a flexible method (such as polynomial regression or KNN with a small value of k), or an inflexible method (such as simple linear regression or KNN with a large value of k).

a. The relationship between the predictors and the response is highly non-linear.

- A. Flexible
- B. Inflexible

Hint: If the relationship is non-linear, then we expect a method that allows for a curve, such as polynomial regression, to fit the data better than a method that only allows for linear models.

b. The variance of the error terms, i.e., $Var(\varepsilon) = \sigma^2$, is extremely high.

- A. Flexible
- B. Inflexible

Hint: When the data are noisy, we need to guard against the risk of overfitting the data and not being able to estimate new data points effectively. A relatively inflexible model will help avoid this risk. It will have much lower variance at the expense of slightly higher bias.

c. The number of predictors p is extremely large, and the number of observations n is small.

- A. Flexible
- B. Inflexible

Hint: With a small data set, we generally don't have enough information to estimate accurately all of the parameters of a flexible method. Changing a single data point can create a large change in the model, resulting in model variance that's too high.

d. The number of observations n is extremely large, and the number of predictors p is small.

- A. Flexible
- B. Inflexible

Hint: This is the opposite situation as part c: With a large data set, we have enough information to estimate the parameters of a flexible model effectively.

Answer(s) submitted:

- A
- B
- B
- A

submitted: (correct)
recorded: (correct)
Correct Answers:

- A
- B
- B
- A

Problem 2. **(1 point)**

For each of the following situations, decide which type of data mining approach is most appropriate.

a. We have a data set on the largest 500 firms in the US. For each firm, we know the profit, number of employees, industry, and the CEO's compensation. We are interested in understanding which factors are associated with the CEO's compensation.

- A. Regression
- B. Classification
- C. Unsupervised learning

b. We are interested in identifying groups of stocks that tend to increase and decrease in value together (i.e., when stock A goes up, stock B also goes up). We have data on the change in value of 500 companies' stocks from every trading day in 2015.

- A. Regression
- B. Classification
- C. Unsupervised learning

c. We are interested in predicting which customers will be dissatisfied with their shopping experience. We have a data set of 1000 customers, including what they bought, how long they spent in the store, their ages, and whether they were satisfied or dissatisfied with their shopping experience.

- A. Regression
- B. Classification
- C. Unsupervised learning

Hint: For part b, note that our main interest is identifying groupings of similar stocks. We're not trying to use other variables (like the industry or size of the company) to predict whether a company's stock will increase or decrease in value.

Answer(s) submitted:

- A
- C
- B

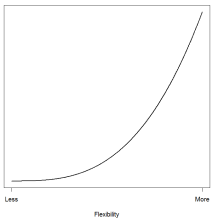
submitted: (correct)
recorded: (correct)
Correct Answers:

- A
- C
- B

Problem 3. **(1 point)**

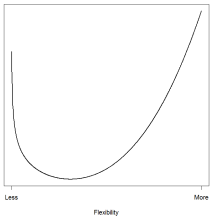
For each of the following graphs, decide whether it represents the squared bias, the variance, or the mean squared error on a test (or validation) data set.

a.



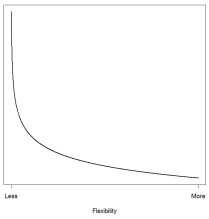
- A. Squared bias
- B. Variance
- C. MSE on a validation set

b.



- A. Squared bias
- B. Variance
- C. MSE on a validation set

c.



- A. Squared bias
- B. Variance
- C. MSE on a validation set

Hint: Look at figure 2.12 on p. 36 of *Introduction to Statistical Learning*.

Answer(s) submitted:

- B
- C
- A

submitted: (correct)
recorded: (correct)
Correct Answers:

- B
- C

Problem 4. **(1 point)**

Suppose you want to predict how much money customers will spend in your store. You have a data set containing the following columns:

- How much each customer spent
- Their income (in dollars)
- Their age (in years)
- Their favorite color (red, yellow, blue, or "other")
- The day of the week when they visit the store (Monday, Tuesday, ..., Sunday)

a. If you are using KNN, how many predictor variables are there?

- A. 11
- B. 4
- C. 13
- D. 6

b. If you are using linear regression, how many predictor variables are there?

- A. 11
- B. 4
- C. 6
- D. 13

For each of the following, decide whether you would treat the discrete, quantitative predictor as categorical or quantitative.

c. You want to use linear regression to predict the population size of a species of fish using Year as a predictor variable, as well as variables about the lakes you will be sampling. The population size of this species varies dramatically from year to year. You are most interested in making predictions about future years for the lakes in your sample.

- A. Categorical
- B. Quantitative

d. In a particular data set, people's shoe sizes range from 4-13. You want to predict people's heights.

- A. Categorical
- B. Quantitative

e. In fashion, the most popular length of sleeve can change rapidly and unpredictably from year to year. You want to predict what length of sleeve will be most preferred by customers in different stores, using Year as a predictor variable, as well as variables about the stores. You are most interested in making predictions about new stores in the same years as are available in the data.

- A. Categorical
- B. Quantitative

f. You want to predict how much money people will spend on travel next year, using the number of US states they have visited in their lifetimes. In your data set of 150 people, the number of US states they have visited ranges from 1 to 48.

- A. Categorical
- B. Quantitative

Answer(s) submitted:

- C
- A
- B
- B
- A
- B

submitted: (correct)
recorded: (correct)
Correct Answers:

- C
- A
- B
- B
- A
- B

Problem 5. **(1 point)**

The table below provides a training data set containing 6 observations, 3 predictors, and 1 qualitative response variable.

Observation	X¡sub¿1	X¡sub¿2	X¡sub¿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 want to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$, using K-nearest neighbors.

a. Find the Euclidean distance between the point $(0,0,0)$ and each observation in the training set.

Observation 1: ____
Observation 2: ____
Observation 3: ____
Observation 4: ____
Observation 5: ____
Observation 6: ____

Hint:

- Recall that the Euclidean distance between two points (x_1,x_2,x_3) and (y_1,y_2,y_3) is $\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2}$
- The "validation set" in this problem is the single point for which we want to make a prediction, which is (0, 0, 0).
- It's fine to use R as a scientific calculator for this problem (e.g., to square numbers and take square roots), but the goal of this problem is to do it without using the *knn* function, to help you get a better sense of what that function is doing behind the scenes.
- WeBWorK expects answers to be within a certain percentage tolerance of the correct answer (so the smaller the correct answer is, the more decimal places of accuracy you need to include in your answer). Usually 4 digits after the decimal point is sufficient.
- If you think you are doing a problem correctly but WeBWorK says it's incorrect, try including more digits after the decimal point **or** entering the exact answer in mathematical notation. For example, you could type
  `sqrt(3)`
  to get $\sqrt{3}$.

b. What is our prediction with $K = 1$?

- A. Red
- B. Green

c. What is our prediction with $K = 3$?

- A. Red
- B. Green

d. If the Bayes decision boundary (which divides red from green) in this problem is highly non-linear, would we expect the best value of K to be large or small?

- A. Small, because we want an inflexible method.
- B. Large, because we want a flexible method.
- C. Large, because we want an inflexible method.
- D. Small, because we want a flexible method.

Hint:

- K-nearest neighbors becomes less flexible as K increases.
- If the boundary between red and green points is very nonlinear and "squiggly", does it make sense to look at neighbors in a large region? It may help to review problem 1 in this WeBWorK set.

Answer(s) submitted:

- 3
- 2
- $\sqrt{10}$
- $\sqrt{5}$
- $\sqrt{2}$
- $\sqrt{3}$
- B
- A
- D

submitted: (correct)
recorded: (correct)

Correct Answers:

- 3
- 2
- $\sqrt{10}$
- $\sqrt{5}$
- $\sqrt{2}$
- $\sqrt{3}$
- B
- A
- D

Problem 6. **(1 point)**

In this problem, we will use K-nearest neighbors to analyze a data set about neighborhoods in Boston.

a. The **Boston** data set is in the **MASS** package. Working in R, install the **MASS** package (if you have not previously done so), load the package, and look at the first 6 rows of data.

- A. Show me how
- B. OK, got it!

Hint:

```
install.packages("MASS")
```

(You can choose any mirror to download the package from. I usually choose the non-https version of Cloud-0.)

```
library(MASS)
head(Boston)
```

b. In this data set, each row represents 1 neighborhood. How many neighborhoods are in the **Boston** data set?

____

Hint:

```
dim(Boston)
```

c. We can view the documentation for the data set by typing

```
?Boston
```

If we want to predict a neighborhood's crime rate (*crim*) based on its *age* and its distance from highways (*rad*), which of the following methods would be more appropriate?

- A. knn
- B. knn.reg

Install the **FNN** package and load the package.

Hint:

```
install.packages("FNN")
library(FNN)
```

Answer(s) submitted:

- B
- 506
- B

submitted: (correct)
recorded: (correct)
Correct Answers:

- B
- 506
- B

Problem 7. **(1 point)**

We'll start by splitting the data into a training set and a validation set.

a. What's the advantage of using the function *set.seed()*?

- A. It initializes R's random number generator, so that the sample() function does not produce an error.
- B. It specifies the value of the first random number R will generate, so the error rates look better.
- C. It allows you or others to reproduce an analysis using the same random numbers.

b. Set the seed to 100 and take a random sample of 350 neighborhoods to be in the training data set. You may click on the hint (visible after your first attempt) to see the code I used.

- A. OK, got it!
- B. Show me how

Hint:

```
set.seed(100)

groups = c(rep(1, 350), rep(2, 156)) # 1 represents the training set
random_groups = sample(groups, 506)

in_train = (random_groups == 1)
```

Solution:

```
set.seed(100)

groups = c(rep(1, 350), rep(2, 156)) # 1 represents the training set
random_groups = sample(groups, 506)

in_train = (random_groups == 1)
```

c. Create a new matrix, *x_train*, containing the predictor variables *age* and *rad* for the training data. Also create a new matrix, *x_test*, containing *age* and *rad* for the validation data.

- For consistency with the autograder on later questions, put the columns in the order *age, rad*.

When doing this with functions from the *dplyr* package, the functions to use are

- A. mutate and filter
- B. arrange and mutate
- C. arrange and summarise
- D. select and filter

Hint:

```
x_train <- Boston %>%
  dplyr::select(c(age, rad)) %>%
  filter(in_train)

x_test <- Boston %>%
  dplyr::select(c(age, rad)) %>%
  filter(!in_train)
```

Note that the MASS package (which we loaded to give us access to the Boston data set) also contains a function called *select*. Using the code

```
dplyr::select
```

specifies that we want the *select* function from *dplyr*.

Solution:

```
x_train <- Boston %>%
  dplyr::select(c(age, rad)) %>%
  filter(in_train)

x_test <- Boston %>%
  dplyr::select(c(age, rad)) %>%
  filter(!in_train)
```

Note that the MASS package (which we loaded to give us access to the Boston data set) also contains a function called *select*. Using the code

```
dplyr::select
```

specifies that we want the *select* function from *dplyr*.

Answer(s) submitted:

- C
- A
- D

submitted: (correct)
recorded: (correct)
Correct Answers:

- C
- A
- D

Problem 8. **(1 point)**

a. Why is it important to standardize *age* and *rad* before using K-nearest neighbors?

- A. The two variables have different standard deviations. A difference of 1 year (in age) is a smaller distance between points than a difference of 1 mile (in rad), but KNN would treat both of these differences the same. Standardizing the variables means that a distance of 1 represents 1 standard deviation, regardless of which variable it is.
- B. The distribution of age is left-skewed, while the distribution of rad is right-skewed. Standardizing the variables will remove any outliers, preventing them from introducing bias into the model.

b. Suppose you were *only* allowed to standardize the data set to have a mean of 0 *or* a standard deviation of 1. Which one should you pick?

- A. Standard deviation
- B. Mean

Fortunately, we *are* allowed to standardize both the mean and standard deviation of the data. Use the code

```
x_train = scale(x_train)
```

to create a new matrix containing standardized versions of *age* and *rad* for the training data.

Then use

```
attr(x_train, "scaled:center")
attr(x_train, "scaled:scale")
```

to view the means and standard deviations for each predictor variable. (The function *attr* extracts the "attributes" of the object *x_train*.)

- There are also many other other ways to compute the means and standard deviation.

c. What was the mean of *rad* in the training data, before scaling?

___

Next, we want to use the same means and standard deviations to standardize the validation data. There are many ways to do this. One way is to use the *scale()* function again, but set additional arguments to specify the means and standard deviations you found in the training data:

```
x_test = scale(x_test, center = attr(x_train, "scaled:center"),
                       scale = attr(x_train, "scaled:scale"))
```

d. Why should we standardize the validation data using the same means and standard deviations as the training data?

- A. Because we want to use the validation data to assess the model we're building on the training data, and the method of standardization is part of that model.
- B. To ensure that any particular standardized value represents the same thing in the validation set as in the training set. E.g., age = 1 represents the same age of neighborhood in both data sets.
- C. Both of the above.

Answer(s) submitted:

- A
- A
- 9.79
- C

submitted: (correct)
recorded: (correct)
Correct Answers:

- A
- A
- 9.79143
- C

Problem 9. **(1 point)**

a. Perform 25-nearest neighbors to predict *crim* for the Boston data. What is the mean squared error on the validation set?

- A. 16.59002
- B. 15.58647
- C. 16.44127
- D. 16.42873

Hint:

```
predictions = knn.reg(train = x_train,
                      test  = x_test,
                      y = Boston$crim[in_train],
                      k = 25)

MSE = mean( (predictions$pred - Boston$crim[!in_train])^2 )
MSE
```

- If you got one of the incorrect answer options for your MSE, double-check how you scaled the training and test data.
- If your MSE is slightly different from the correct multiple-choice answer, it could be because you generated additional random numbers between setting the seed and generating the training data. (Generating random numbers changes the seed.) Go back to problem 6 and run your code straight through until this problem. Do not re-set your random seed or generate extraneous random numbers.
- Another possibility is that you are using an old version of R (pre-3.6). Please update your version of R.
- If your MSE still doesn't match, please post a question on Piazza. Include your complete code for problems 6-8. (Because WeBWorK questions are intended for practice, and often include example code in the hints, it's fine to make this a public post.)

Next, we will focus on interpreting the model. Make histograms of *age* and *rad* in the training data.

b. For what range of values of *age* will be it most appropriate/useful/reliable to make predictions?

- A. seq(50, 200, by = 1)
- B. seq(20, 100, by = 1)
- C. seq(0, 50, by = 1)

c. For what range of values of *rad* will be it most appropriate/useful/reliable to make predictions?

- A. seq(1, 8, by = 1)
- B. seq(1, 24, by = 1)
- C. c(seq(1, 8, by = 1), 24)
- D. c(1, seq(8, 24, by = 1))

Hint:

```
Boston %>%
  filter(in_train) %>%
  gf_histogram(~rad)

Boston %>%
  filter(in_train) %>%
  filter(rad < 24) %>%
  summarise(max(rad))

Boston %>%
  filter(in_train) %>%
  gf_histogram(~age)
```

d. Make a grid of example points at the values of *age* and *rad* you selected. Rename the columns *age* and *rad*.

- A. Show me how
- B. OK, got it!

Hint:

```
age_to_check = seq(20, 100, by = 1)
rad_to_check = c(seq(1, 8, by = 1), 24)

example_data = expand.grid(age_to_check,
                           rad_to_check)
example_data <- example_data %>%
  rename(age = Var1,
         rad = Var2)
```

Answer(s) submitted:

- C
- B
- C
- B

submitted: (correct)
recorded: (correct)
Correct Answers:

- C
- B
- C
- B

Problem 10. **(1 point)**

a. When scaling the example data points, what mean should you use?

- A. The mean age and rad in the training set.
- B. The mean age and rad in the example points.
- C. The mean age and rad in the validation set.
- D. Any of the above.
- E. It is not necessary to scale the example points.

b. Scale the example points and use 25-nearest neighbors to predict *crim* for the example points. Use the same training set as in the previous problem.

- A. OK, got it!
- B. Show me how

Hint:

```
x_example = scale(example_data,
                  center = attr(x_train, "scaled:center"),
                  scale = attr(x_train, "scaled:scale"))

predictions = knn.reg(train = x_train,
                  test  = x_example,
                  y = Boston$crim[in_train],
                  k = 25)
```

c. Make a graph showing points for the values of *rad* and *age*, with different colors to denote the predicted value of *crim*.

- A. OK, got it!
- B. Show me how

Hint:

```
example_data <- example_data %>%
  mutate(pred = predictions$pred)

example_data %>%
  gf_point(rad ~ age, color =~ pred)
```

In this case, this graph is pretty difficult to read, due to the large gap in realistic values of *rad* and the large number of points with low predicted values of *crim*. An alternative approach is to make a line graph of the relationship between *pred* and *age*, for selected values of *rad*:

```
example_data %>%
  filter(rad %in% c(1,8,24)) %>%
  gf_line(pred ~ age, color =~ factor(rad))
```

d. Interpret the graph.

Neighborhoods with [/low/high] accessibility of radial highways tend to have low rates of crime, perhaps because

- `they're easy for police to reach`
- criminals want a fast getaway

. Among neighborhoods with [/low/high] accessibility of radial highways, older neighborhoods tend to have [/more/less] crime. This could be because they're more worn-down and

- `don't have things worth stealing`
- easy targets

, or perhaps because they're well-preserved and

- have things worth stealing
- have good alarm systems

. More investigation is warranted into the relationship between age and affluence of a neighborhood.

e. You are a time traveller who wants to advise city planners about how to reduce crime rates. To which of the following cities would the interpretation above be most relevant? (It may help to refer to the documentation for the Boston data set.)

- A. Chicago, 1980
- B. Boston, 1878
- C. Tilt Cove, Canada, 1976

Hint: Consider similarities between both the city and the year in which the data were gathered.

Answer(s) submitted:

- `A`
- `A`
- `A`
- low
- criminals want a fast getaway
- high
- more
- easy targets
- have things worth stealing
- `A`

submitted: (correct)
recorded: (correct)
Correct Answers:

- A
- A
- A
- low
- criminals want a fast getaway
- high
- more
- easy targets
- have things worth stealing
- A