Midterm Quiz 1 Verified

Bookmark this page

Midterm Quiz 1 due Mar 6, 2023 15:00 +08 Completed

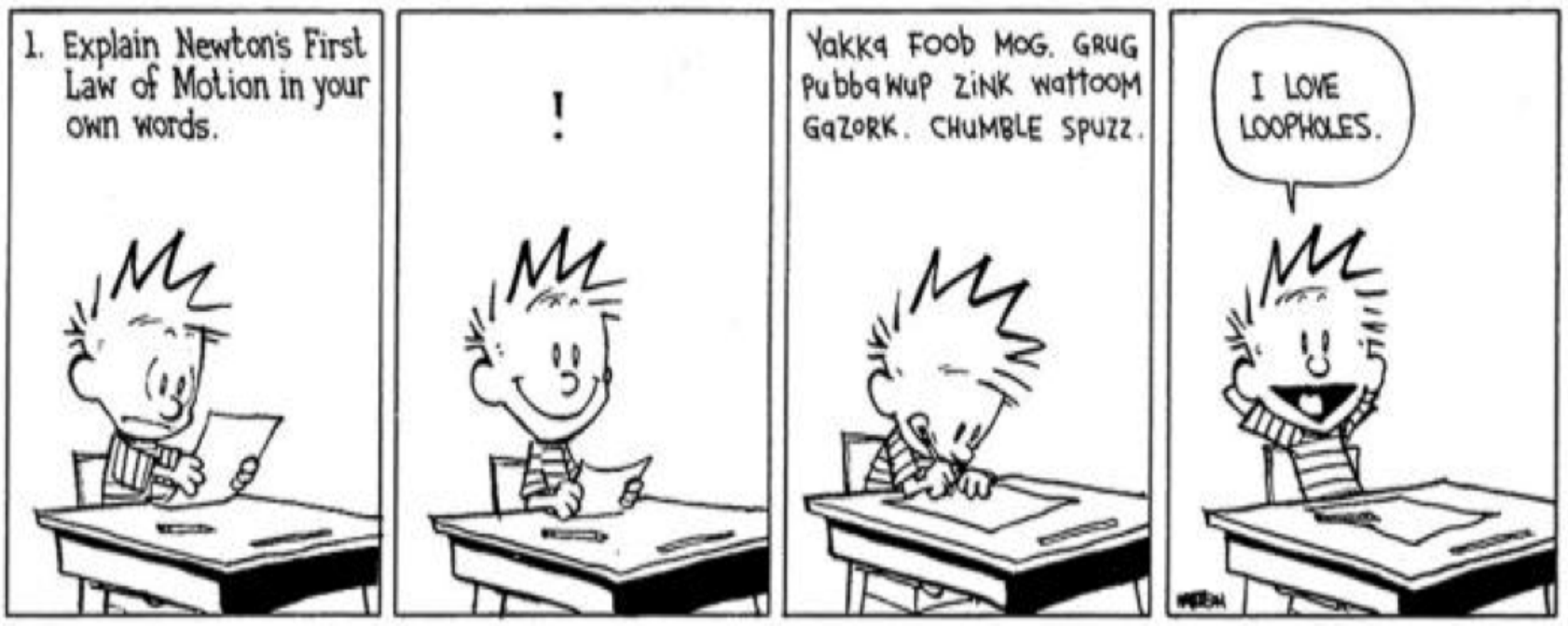
View the Proctoring System Requirements to ensure that your set-up will work. Note that proctoring is only supported on MacOS and Windows machines. We recommend 2 GB of free space on your machine, and a functioning Webcam is required. Your space should be clean, no writing visible on walls or surfaces, and you should be alone in the room. Please make sure that you have verified your ID before taking the exam.

---

95 Minute Time Limit

Instructions

- **Work alone.** Do not collaborate with or copy from anyone else.
- Work the problems in any order you wish, but submit your answer to each before ending the exam.
- You may use any of the following resources:
  - One sheet (both sides) of handwritten (not photocopied or scanned) notes
- If any question seems ambiguous, use the most reasonable interpretation (i.e. don't be like Calvin):



- **If you experience any technical issues (i.e. Math Processing Error), please save your current selected answers and refresh the page. If the issue persists, then please finish the exam and let the Instructors know about the issue in a private Piazza post afterwards.**
- Good Luck!

Question 0 -- Practice with Drag & Drop

0 point possible (ungraded)

⌨ Keyboard Help

Some of the quiz questions are Drag-and-Drop. You'll need to drag one or more answers to a location.

Some answers might not be used at all, and some answers will be used once. To get full credit you might need to drag more than one answer to some locations, just one answer to other locations, and some locations might not have any correct answers.

Please do this quick practice question. The question will give you feedback to make sure you've done it correctly, but the real quiz questions will not.

x=1,y=7

x + y = 5 (x plus y equals 5)	x=1,y=4
x + y = 2 (x plus y equals 2)	
xy = 6 (x times y equals 6)	x=2,y=3 x=1,y=6

Submit You have used 3 of 10 attempts. Reset Show Answer

FEEDBACK

✔ Correctly placed 3 items

ℹ Good work! You have completed this drag and drop problem. Note that: (1) There are two places you could've put (x=2,y=3); either one would be correct. (2) One location (x+y=2) had nothing dragged to it. Another location had two answers dragged to it. (3) One choice (x=1,y=7) was not dragged anywhere, since it wasn't correct for anything.

Question 1

10/13 point (graded)

⌨ Keyboard Help

Drag ***each*** of the 13 models/methods to one of the 5 categories of question it is commonly used for, unless no correct category is listed for it. For models/methods that have more than one correct category, choose any one correct category; for models/methods that have no correct category listed, do not drag them.

CUSUM Principal component analysis

Classification	CART k-nearest-neighbor Support vector machine
Clustering	k-means
Response prediction	ARIMA Exponential smoothing Linear regression Logistic regression Random forest
[illegible]	Cross validation

Validation	
Variance estimation	GARCH

Submit You have used 1 of 1 attempts. Reset Show Answer

FEEDBACK

- ✔ Correctly placed 8 items
- ✖ Misplaced 1 item
- ✖ Did not place 2 required items
- ✱ Final attempt was used, highest score is 10.0
- ℹ **Good work! You have completed this drag and drop problem.**

Question 2

2.73/3.0 points (graded)

Select all of the following models that are designed for use with time series data:

- [x] ARIMA ✱
- [x] CUSUM ✱
- [ ] Support vector machine
- [ ] Random forest
- [ ] k-nearest-neighbor
- [ ] GARCH ✔
- [ ] k-means
- [ ] Logistic regression
- [x] Exponential smoothing ✱
- [ ] Principal component analysis
- [ ] Linear regression

✱

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Information for Questions 3a, 3b, 3c

Figures A and B show the training data for a soft classification problem, using two predictors (x_1 and x_2) to separate between black and white points. The dashed lines are the classifiers found using SVM. Figure A uses a linear kernel, and Figure B uses a nonlinear kernel that required fitting 16 parameter values.

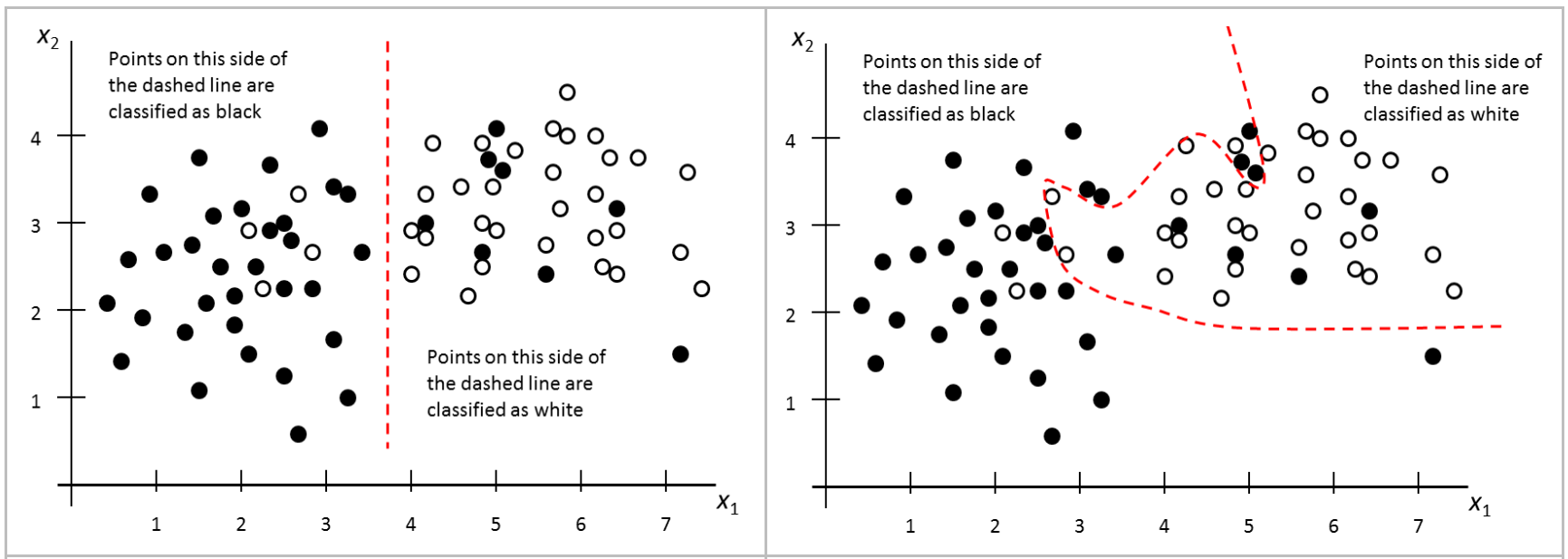


Figure A | **Figure B**

Question 3a

2.4000000000000004/3.0 points (graded)

3a. Select all of the following statements that are true.

- [x] Figure A's classifier is based only on the value of x_1.
- [ ] Figure A's classifier would probably perform worse on test data than on the training data.
- [ ] Figure A's classifier has a narrower margin than Figure B's classifier in the training data.
- [ ] Figure A's classifier incorrectly classifies exactly 4 black points as white in the training data.
- [ ] Figure A shows that the black point (7.2,1.4) is colored incorrectly; it should actually be white.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 3b

2.25/3.0 points (graded)

3b. Select all of the following statements that are true.

- [ ] Figure B's classifier is better than Figure A's classifier, because Figure B's classifier classifies more of the training data correctly.
- [x] Figure B's classifier is more likely to be over-fit than Figure A's classifier.

- [ ] Figure B's classifier incorrectly classifies exactly 5 black points in the training data.

- [ ] Figure B shows that the black point (7.2,1.4) is colored incorrectly; it should actually be white.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 3c

2.25/3.0 points (graded)

3c. Select all of the following statements that are true.

- [x] A new point at (6,4) would be classified as white by Figure A's classifier.
- [x] A new point at (6,4) would be classified as white by Figure B's classifier.
- [x] A new point at (6,4) would be classified as white by a k-nearest-neighbor algorithm for $1 \leq k \leq 10$.
- [x] In Figure A, if the training data had 1000 more black points to the left of the classifier, a 1000-nearest-neighbor algorithm would classify a new point at (6,4) as white.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 3d

2.0100000000000002/3.0 points (graded)

In the soft classification SVM model where we select coefficients $a_0 \ldots a_m$ to minimize

$$\sum_{j=1}^{n} max\{0, 1 - (\sum_{i=1}^{m} a_i x_{ij} + a_0)\, y_j\} + C \sum_{i=1}^{m} a_i^2$$

3d. Select all of the following statements that are correct.

- [ ] Decreasing the value of C could increase the margin.
- [x] Allowing a smaller margin could decrease the number of classification errors in the training set.
- [x] Increasing the value of C could decrease the number of classification errors in the training set.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 3e

3.0/3.0 points (graded)

3e. In the hard classification SVM model, it might be desirable to put the classifier in a location that has equal margin on both sides... (select all correct answers):

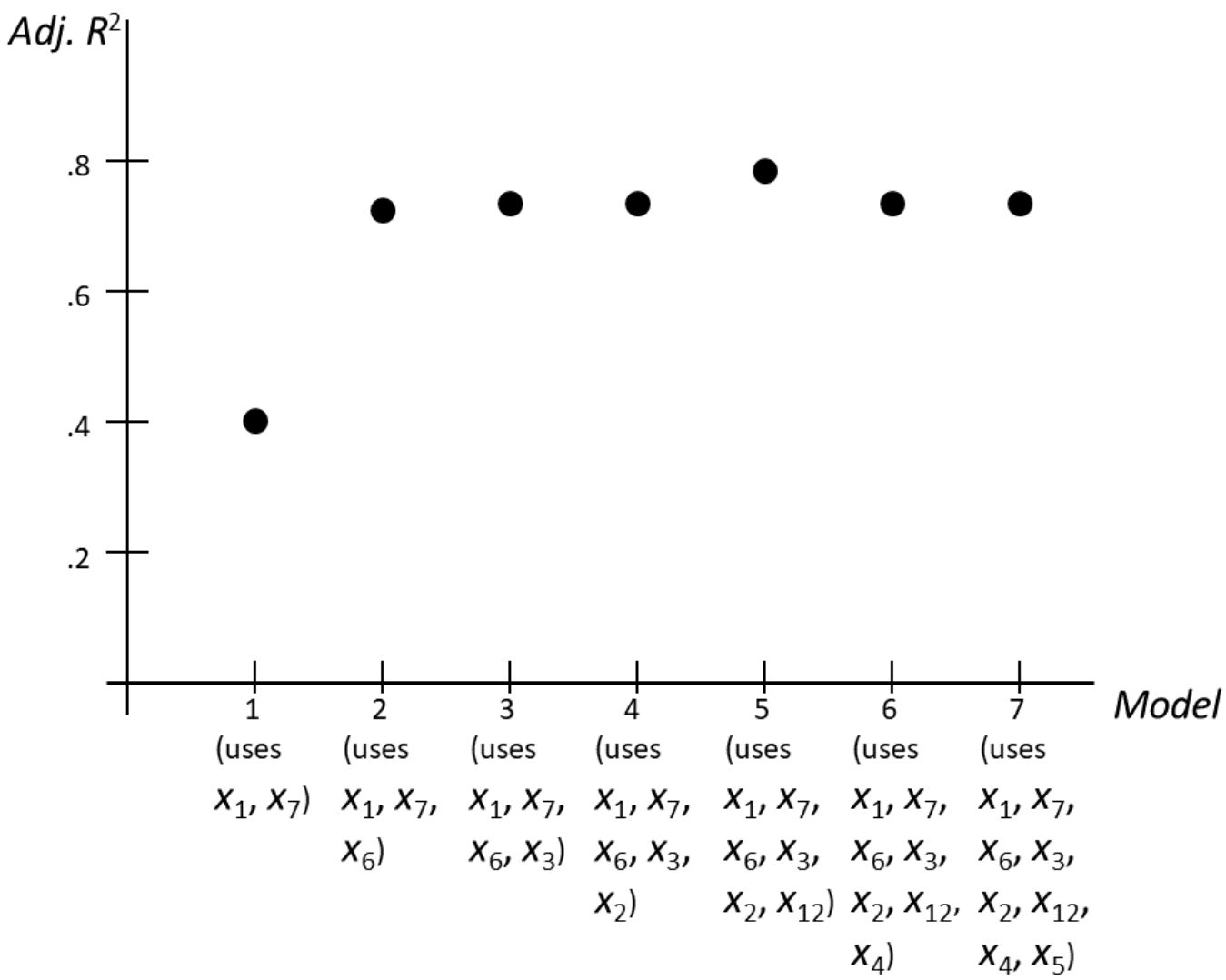
- [x] ...because moving the classifier will usually result in more classification errors in the validation data.
- [x] ...because moving the classifier will usually result in more classification errors in the test data.
- [ ] ...when the costs of misclassifying the two types of points are significantly different.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Questions 4a, 4b, 4c

Seven different regression models have been fitted, using different sets of variables. The figure below shows the resulting adjusted R-squared value for various models, as measured by cross-validation.



Question 4a

3.0/3.0 points (graded)

3.0/3.0 points (graded)

Which of the models would you expect to perform worst on a test data set?

- ○ Model 6, because it has a slightly lower Adjusted R^2 than Model 5 and uses one more predictor.
- ○ Model 2, because it's the simplest of those with a high Adjusted R^2.
- ○ Model 5, because it has the highest Adjusted R^2.
- ◉ Model 1, because it has much lower Adjusted R^2.

✔

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 4b

3.0/3.0 points (graded)

Under which of the following conditions would Model 7 be the most appropriate to use (select all correct answers)?

- [ ] Data collection for x_5 is too expensive for it to be used in the model.
- [x] Government regulations require using x_5 for this sort of model.
- [ ] It is important to find the simplest good model.
- [ ] The value of x_3 is not known in time for use in the model.

✔

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Additional Information for Question 4c

The table below shows the Akaike Information Criterion (AIC), Corrected AIC, and Bayesian Information Criterion (BIC) for each of the models.

Model	AIC	Corrected AIC	BIC
1	-5.58	-5.32	2.07
2	-5.67	-5.15	3.89
3	-6.51	-5.62	4.96
4	-4.77	-3.41	8.61
5	-2.80	-0.85	12.49
6	-1.31	1.35	15.90

7	0.19	3.71	19.31

Question 4c

0.75/3.0 points (graded)

Based on the table above and the figure shown for Question 4a, select all of the following statements that are correct.

- [ ] BIC suggests that Model 1 is very likely to be better than Model 2.
- [x] Among Models 3 and 4, AIC suggests that Model 3 is $e^{(-6.51-(-4.77))/2} = 41.9\%$ as likely as Model 4 to be better.
- [ ] Among Models 3 and 4, AIC suggests that Model 4 is $e^{(-6.51-(-4.77))/2} = 41.9\%$ as likely as Model 3 to be better.
  ✔
- [ ] Adjusted R^2 (see figure above 4a) and BIC (see table above 4c) both agree that Model 5 might be a little better than Model 6.
  ✔

✱

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for all parts of Question 5

Atlanta's main library has collected the following day-by-day data over the past six years (more than 2000 data points):

x_1 = Number of books borrowed from the library on that day
x_2 = Day of the week
x_3 = Temperature
x_4 = Amount of rainfall
x_5 = Whether the library was closed that day
x_6 = Whether public schools were open that day

Question 5a

2.0/2.0 points (graded)

Select all data that are not categorical or binary:

- [x] Number of books borrowed from the library on that day
- [ ] Day of the week
- [x] Temperature
- [x] Amount of rainfall
- [ ] Whether the library was closed that day

- [ ] Whether public schools were open that day

✔

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Questions 5b and 5c

2.0/4.0 points (graded)

The library believes that if it was hotter yesterday, fewer books will be borrowed today (and if it was cooler yesterday, more books will be borrowed today), so they add a new predictor:

x_7 = temperature the day before

b. If the library is correct that on average, if it was hotter yesterday, fewer books will be borrowed today (and if it was cooler yesterday, more books will be borrowed today), what sign (positive or negative) would you expect the new predictor's coefficient a_7 to have?

- ( ) Positive, because the response (books borrowed today) is a positive number
- (•) Negative, because higher values of x_7 decrease the response (books borrowed today)
- ( ) Positive, because higher values of x_7 increase the response (books borrowed today)

✔

c. Does x_7 make the model autoregressive?

- ( ) No, because the model does not use previous response data to predict the day t response. ✔
- ( ) Yes, because the model uses day $t-1$ data to predict day t circulation.
- (•) Yes, because the model uses both day $t-1$ and day t temperature data as predictors.

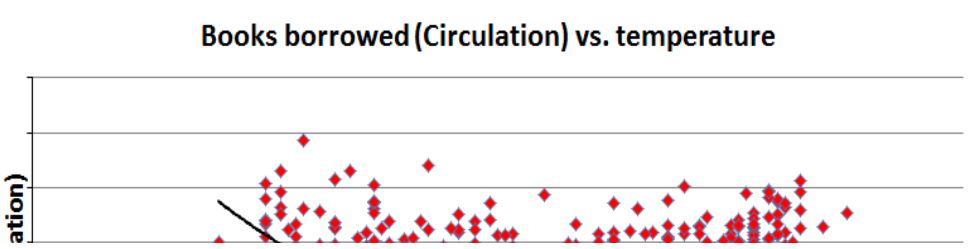
✖

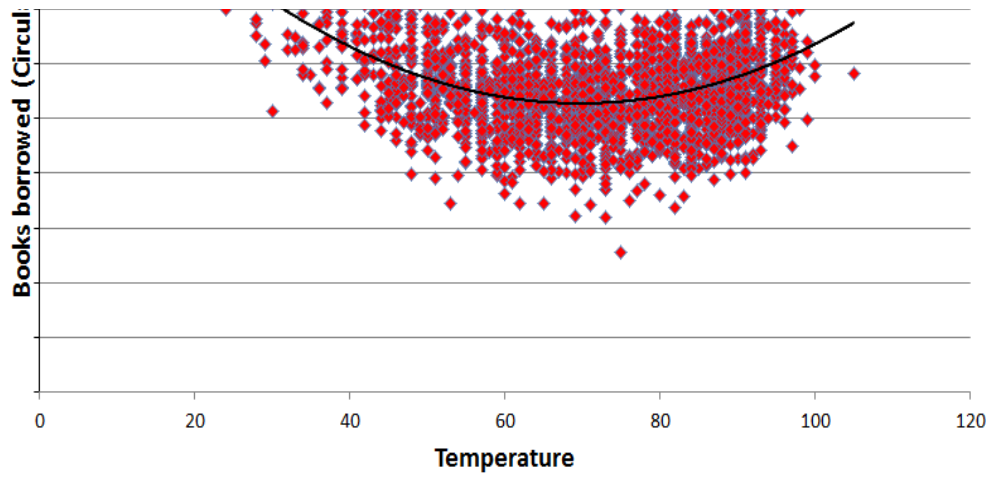
Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Question 5d

The library believes that as the temperature gets either too cold or too hot, more people come indoors to the library to borrow books. They have fit the data to a quadratic function (see the figure below).





Question 5d

4.0/4.0 points (graded)

How would you incorporate the new information above into the library's regression model?

- [x] Add a (temperature)^2 variable to the model.
- [ ] Replace the temperature variable with a (temperature)^2 variable in the model.
- [ ] Change the model to estimate the square root of the books borrowed, as a function of temperature, day of the week, inches of rainfall, whether the day is a holiday, and whether schools were open.

✔

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5e-i,ii

6.0/6.0 points (graded)

The library has built a triple exponential smoothing (Holt-Winters) model of the number of books borrowed each day, using a multiplicative weekly cycle of seasonality (i.e., L=7).

i. Every year on July 4, the library shoots off fireworks in its parking lot, so nobody is allowed to borrow books that day. The model only has a weekly seasonality, not an annual one. Is the model likely to over-predict or under-predict books borrowed on July 4?

- [x] Over-predict
- [ ] Under-predict
- [ ] Neither

✔

ii. Is the model likely to over-predict or under-predict books borrowed on July 5? [Assume the library is open and allows borrowing on July 5.]

- [ ] Over-predict
- [x] Under-predict
- [ ] Neither

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5e-iii

3.0/3.0 points (graded)

iii. Aside from seasonal and trend effects, the library believes that the random variation in books borrowed each day is large. Should they expect the best value of α (the baseline smoothing constant) to be:

- ○ $\alpha < 0$
- ◉ $0 < \alpha < \frac{1}{2}$
- ○ $\frac{1}{2} < \alpha < 1$
- ○ $\alpha > 1$

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Questions 5f, 5g, 5h

The library would like to compare the regression and exponential smoothing models to determine which is a better predictor, using the mean absolute error |(books borrowed) – (model's estimate)|/n as a measure of prediction quality.

Question 5f

4.0/4.0 points (graded)

Select the best of the following four options for splitting the data:

- ○ 15% for training, 70% for validation, 15% for test
- ◉ 70% for training, 15% for validation, 15% for test
- ○ 15% for training, 15% for validation, 70% for test
- ○ 55% for training, 15% for cross-validation, 15% for validation, 15% for test

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5g

4/4 point (graded)

Keyboard Help

Match each data set with its purpose. Drag the purpose next to the appropriate data set.

Test set	Estimate quality of selected model
Training set	Fit parameters of all models
Validation set	Compare all models & select best

Submit You have used 1 of 1 attempts.

Reset Show Answer

FEEDBACK

- ✔ Correctly placed 3 items
- ✔ Final attempt was used, highest score is 4.0
- ℹ **Good work! You have completed this drag and drop problem.**

Question 5h

2.0/4.0 points (graded)

The person who built these models discovered that although the exponential smoothing model performed well on the training set, it performed much worse on the validation set:

	Mean absolute error (training set)	**Mean absolute error (validation set)**
Regression model	130	139
Exponential smoothing model	128	167

Select all of the reasonable suggestions below:

- [ ] The exponential smoothing model is probably worse, because it does much worse on the validation set. ✔
- [x] The exponential smoothing model is probably fit too much to random patterns (i.e., it is overfit), because it performs much worse than the regression model on the validation set.
- [x] To choose between the models, we should see which one does better on the validation set. ✱
- [x] If there had been 20 models, the one that performed best on the validation set would probably not perform as well on the test set as it did on the validation set. ✱

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5i

0.99/3.0 points (graded)

Fewer books are borrowed on Fridays than any other day. The library would like to determine whether there has been a change in the Friday effect on borrowing, over the past forty years. Select all of the approaches that might reasonably be correct.

- [ ] Use CUSUM on the number of books borrowed each day over the past forty years.
- [x] Use exponential smoothing (with $L = 365$) to find the daily mulitplier values C_t, and use CUSUM on those values.
- [ ] Build a regression model for each of the forty years, and use CUSUM on the coefficients of the Friday variable.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Questions 6a, 6b

A logistic regression model was built to model the probability that a retailer's inventory of a popular product will run out before the next delivery from the manufacturer, based on a number of factors (amount of current inventory, past demand, promotions, etc.).

If the logistic regression's output is greater than a threshold value p, the retailer pays an additional amount D for a quick delivery to avoid running out.

There are three confusion matrices below, for three different threshold values of p:

		Model result	
	p=0.3	Run out*	Okay
True	Run out*	91	9
	Okay	49	51

*Run out *unless* retailer pays for early delivery

		Model result	
	p=0.5	Run out*	Okay
True	Run out*	76	24
	Okay	27	73

*Run out *unless* retailer pays for early delivery

		Model result	
	p=0.7	Run out*	Okay
True	Run out*	53	47
	Okay	8	92

*Run out *unless* retailer pays for early delivery

Question 6a

2.0100000000000002/3.0 points (graded)

Let D be the cost of paying for a quick delivery (if the model's output is above p). Let C be the cost of running out of inventory. Select all of the statements that are correct:

- [ ] When p=0.7, the total cost is $(53D + 47D + 8C)$.

When p=0.7, the total cost is $(53D + 47C + 8D)$.

The fewest extra deliveries are made when p=0.7.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 6b

3.0/3.0 points (graded)

Early delivery is expensive for this retailer; it estimates the cost C of running out to be equal to the cost D of paying for an early delivery (i.e., C = D). Which threshold value of p would you suggest?

- p = 0.3
- p = 0.5
- p = 0.7

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 7

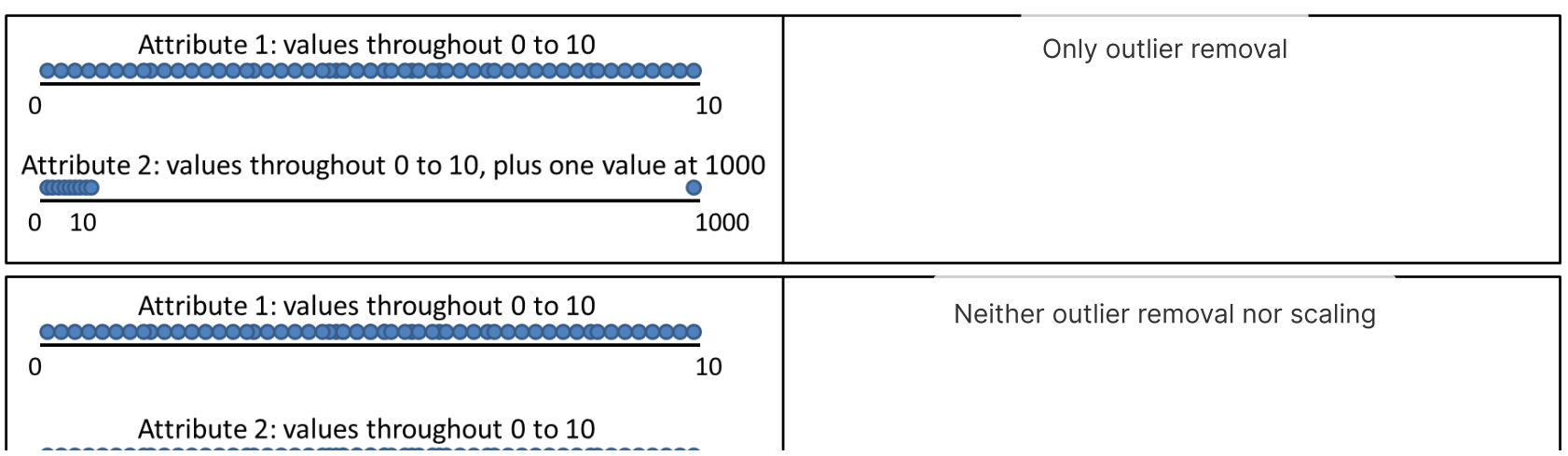
3.2/8 point (graded)

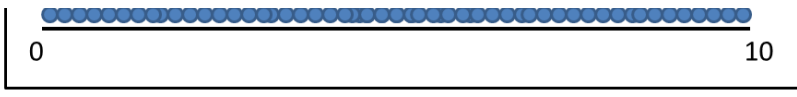
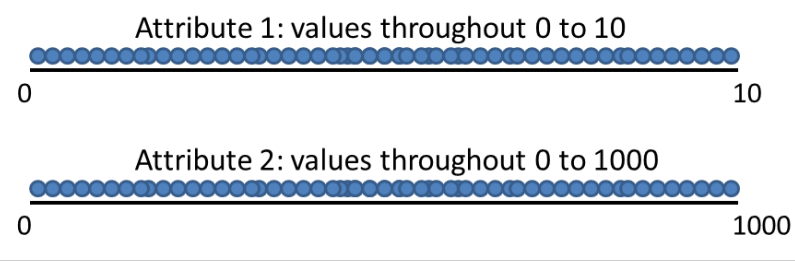
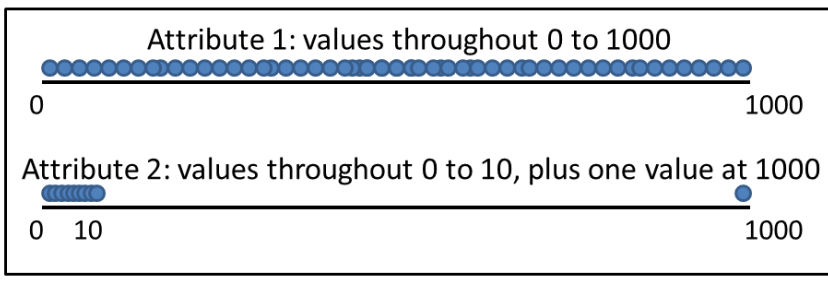
Keyboard Help

The figures below each show a data set that will be used in k-means clustering algorithms (where distance between values is important).

Each data set has two attributes. For each data set, drag to it the data preparations that are needed for k-means to work well on the data set.

First scaling and then outlier removal



	
	Only scaling
	First outlier removal and then scaling

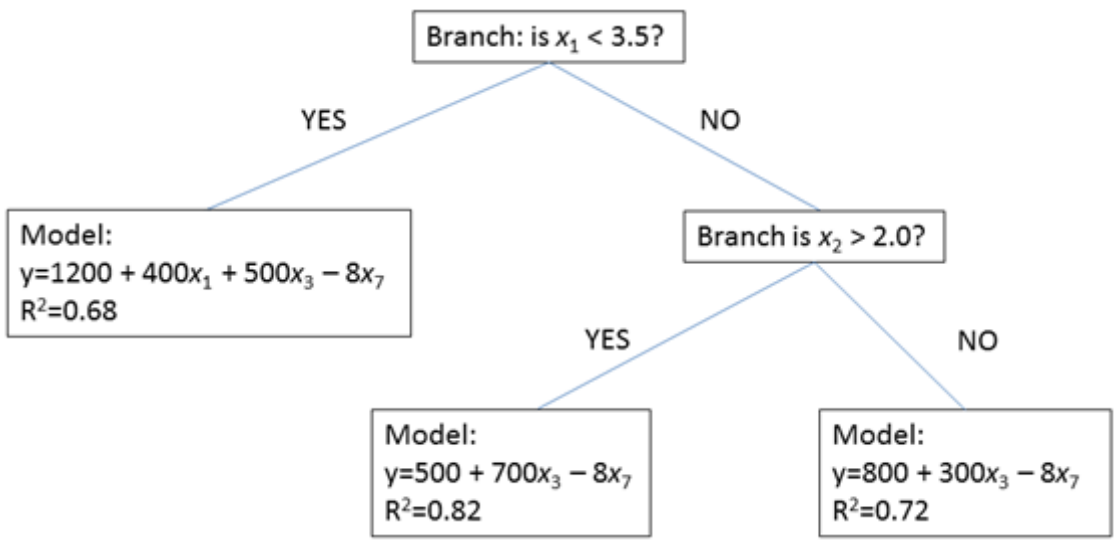
Submit You have used 1 of 1 attempts. Reset Show Answer

FEEDBACK

- ✔ Correctly placed 2 items
- ✖ Misplaced 2 items
- ✖ Did not place 1 required item
- ✱ Final attempt was used, highest score is 3.2
- ℹ **Good work! You have completed this drag and drop problem.**

Information for Questions 8a, 8b

A regression tree approach was used to describe the effect of 7 different covariates (x1 through x7) on monthly sales. The tree is shown below. In each model, only the significant covariates are shown.



Question 8a

5.0/5.0 points (graded)

Select all of the following statements that are true according to this regression tree:

- [ ] The effect of x_7 depends on the values of other variables.
- [ ] x_2 is irrelevant when predicting monthly sales.
- [x] x_6 is irrelevant when predicting monthly sales.
- [x] The model's predictions are best when both x_1 and x_2 are large ($x_1 \geq 3.5$ and $x_2 > 2.0$).

- [ ] The effect of x_3 on sales is smallest when x_1 is small and x_2 is large ($x_1 < 3.5$ and $x_2 > 2.0$).

✔

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 8b

2.0100000000000002/3.0 points (graded)

A random forest model was built for the same purpose, using the same 7 covariates. Which of the following statements are true?

- [x] The random forest model does not return a single tree solution that can be analyzed. *
- [x] The random forest model uses many trees with different branchings. *
- [x] The random forest model cannot report the relative importance of each variable.

*

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Question 8c

A data scientist has run principal component analysis on the 7 covariates, with the following results:

Component	Eigenvalue
1	2.20
2	0.12
3	0.10
4	0.09
5	0.08

Legal

Terms of Service & Honor Code
Privacy Policy
Accessibility Policy
Trademark Policy
Sitemap
Cookie Policy
Do Not Sell My Personal Information

Connect

Blog
Contact Us
Help Center
Security
Media Kit
