Midterm Quiz 2 - Spring 2019 - Verified Learners

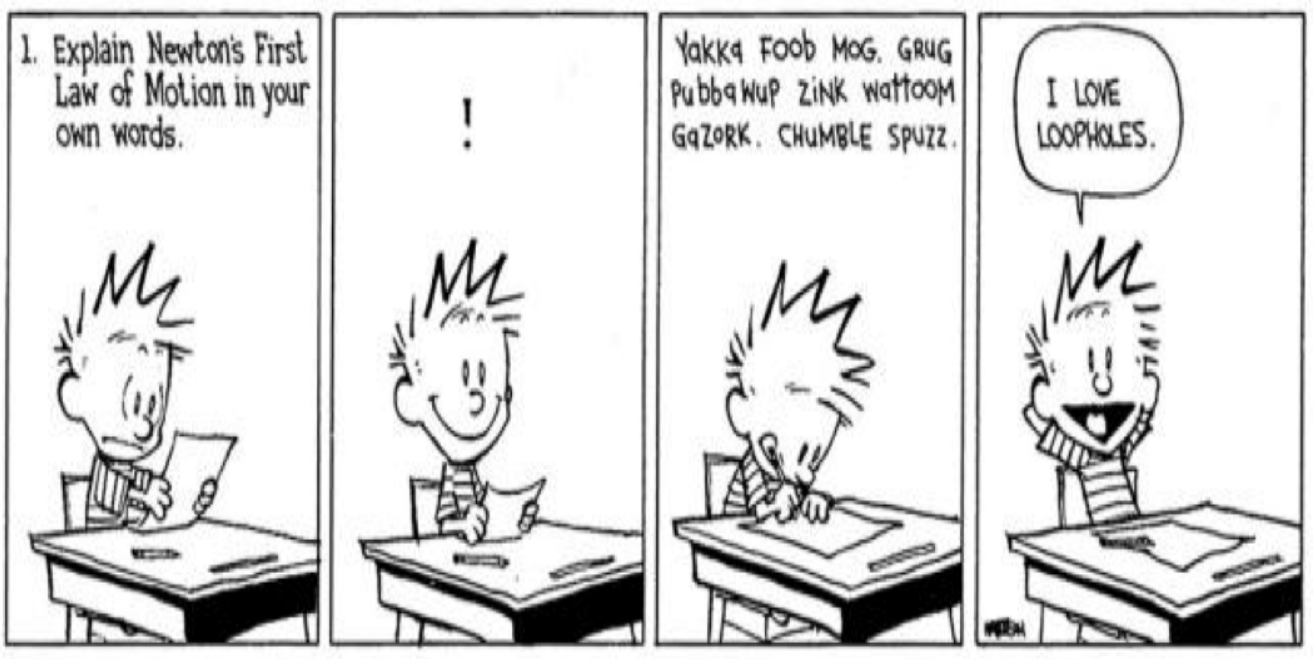
View the Proctoring System Requirements to ensure that your set-up will work. Note that proctoring is only supported on MacOS and Windows machines. We recommend 1GB of free space on your machine, and a functioning Webcam is required. Your space should be clean, no writing visible on walls or surfaces, and you should be alone in the room. Please make sure that you have verified your ID before taking the exam.

The proctored exams are offered by a 3rd party vendor called Software Secure. If you get stuck during the proctoring process you are welcome to contact Software Secure's chat which is available 24/7 or call them at 1-844-224-9759

90 Minute Time Limit

Instructions

- **Work alone.** Do not collaborate with or copy from anyone else.
- 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):



- Good Luck!

This the beginning of Midterm Quiz 2. Please make sure that you submit all your answers before the time runs out. Once you submit an answer to a question, you cannot change it. There is no overall Submit button.

Information for Question 1

There are five questions labeled "Question 1." Answer all five questions. For each of the following five questions, select the probability distribution that could best be used to model the described scenario. Each distribution might be used, zero, one, or more than one time in the five questions.

Question 1

1.4/1.4 points (graded)

Time between people entering a grocery store

Exponential ▼ ✔ **Answer:** Exponential

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 1

1.4/1.4 points (graded)

Number of faces in a 500-person auditorium that are correctly identified by deep learning (DL) software

Binomial ▼ ✔ **Answer:** Binomial

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 1

1.4/1.4 points (graded)

Time from when a generator is turned on until it fails

Weibull ▼ ✔ **Answer:** Weibull

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 1

0.0/1.4 points (graded)

Time from the beginning of Fall until the first snowflake is seen

Exponential ▼ ✖ **Answer:** Weibull

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 1

1.4/1.4 points (graded)
Number of people entering a grocery store each minute

Poisson ▼ ✔ **Answer:** Poisson

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Questions 2a, 2b

5.0/10.0 points (graded)
Five classification models were built for predicting whether a neighborhood will soon see a large rise in home prices, based on public elementary school ratings and other factors. The training data set was missing the school rating variable for every new school (3% of the data points).

Because ratings are unavailable for newly-opened schools, it is believed that locations that have recently experienced high population growth are more likely to have missing school rating data.

- Model 1 used imputation, filling in the missing data with the average school rating from the rest of the data.
- Model 2 used imputation, building a regression model to fill in the missing school rating data based on other variables.
- Model 3 used imputation, first building a classification model to estimate (based on other variables) whether a new school is likely to have been built as a result of recent population growth (or whether it has been built for another purpose, e.g. to replace a very old school), and then using that classification to select one of two regression models to fill in an estimate of the school rating; there are two different regression models (based on other variables), one for neighborhoods with new schools built due to population growth, and one for neighborhoods with new schools built for other reasons.
- Model 4 used a binary variable to identify locations with missing information.

- Model 5 used a categorical variable: first, a classification model was used to estimate whether a new school is likely to have been built as a result of recent population growth; and then each neighborhood was categorized as "data available", "missing, population growth", or "missing, other reason".

a. If school ratings can be reasonably well-predicted from the other factors, and new schools built due to recent population growth cannot be reasonably well-classified using the other factors, which model would you recommend?

- Model 1
- Model 2 ✔
- Model 3
- Model 4 ✖
- Model 5

b. In which of the following situations would you recommend using Model 3? [All predictions and classifications below are using the other factors.]

- Ratings can be well-predicted, and reasons for building schools can be well-classified. ✔
- Ratings can be well-predicted, and reasons for building schools cannot be well-classified.
- Ratings cannot be well-predicted, and reasons for building schools can be well-classified.
- Ratings cannot be well-predicted, and reasons for building schools cannot be well-classified.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Question 3

In a diet problem (like we saw in the lessons and homework), let x_i be the amount of food `i` in the solution (x_i `>= 0`), and let `M` be the maximum amount that can be eaten of any food.

Suppose we added new variables y_i that are binary (i.e., they must be either 0 or 1): if food `i` is eaten in the solution, then it is part of the solution (y_i `= 1`); otherwise y_i `= 0`.

There are five questions labeled "Question 3." Answer all five questions. For each of the following five questions, select the mathematical constraint that best corresponds to the English sentence. Each constraint might be used, zero, one, or more than one time in the five questions.

Question 3

1.4/1.4 points (graded)
Select the mathematical constraint that corresponds to the following English sentence:

No more than two of broccoli, cheese sauce, and peanut butter may be eaten.

- $y_{peanutbutter} + y_{cheesesauce} = 0$
- $y_{peanutbutter} = 1 - y_{cheesesauce}$
- $y_{broccoli} \leq y_{cheesesauce} + y_{peanutbutter}$
- $y_{broccoli} + y_{cheesesauce} + y_{peanutbutter} \leq 2$ ✔
- $x_{cheesesauce} \leq M y_{cheesesauce}$

- $y_{cheesesauce} = 1$
- $x_{broccoli} \leq M y_{peanutbutter}$
- $x_{broccoli} \geq M y_{peanutbutter}$

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 3

1.4/1.4 points (graded)

Select the mathematical constraint that corresponds to the following English sentence:

Unless peanut butter is eaten, no amount of broccoli can be eaten.

- $y_{peanutbutter} + y_{cheesesauce} = 0$
- $y_{peanutbutter} = 1 - y_{cheesesauce}$
- $y_{broccoli} \leq y_{cheesesauce} + y_{peanutbutter}$
- $y_{broccoli} + y_{cheesesauce} + y_{peanutbutter} \leq 2$
- $x_{cheesesauce} \leq M y_{cheesesauce}$
- $y_{cheesesauce} = 1$
- $x_{broccoli} \leq M y_{peanutbutter}$ ✔

- $x_{broccoli} \geq My_{peanutbutter}$

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 3

1.4/1.4 points (graded)

Select the mathematical constraint that corresponds to the following English sentence:

If broccoli is eaten, then either cheese sauce or peanut butter (or both) must also be eaten.

- $y_{peanutbutter} + y_{cheesesauce} = 0$
- $y_{peanutbutter} = 1 - y_{cheesesauce}$
- $y_{broccoli} \leq y_{cheesesauce} + y_{peanutbutter}$ ✔
- $y_{broccoli} + y_{cheesesauce} + y_{peanutbutter} \leq 2$
- $x_{cheesesauce} \leq My_{cheesesauce}$
- $y_{cheesesauce} = 1$
- $x_{broccoli} \leq My_{peanutbutter}$
- $x_{broccoli} \geq My_{peanutbutter}$

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 3

1.4/1.4 points (graded)

Select the mathematical constraint that corresponds to the following English sentence:

If peanut butter is not eaten, then no amount of broccoli may be eaten.

- $y_{peanutbutter} + y_{cheesesauce} = 0$
- $y_{peanutbutter} = 1 - y_{cheesesauce}$
- $y_{broccoli} \leq y_{cheesesauce} + y_{peanutbutter}$
- $y_{broccoli} + y_{cheesesauce} + y_{peanutbutter} \leq 2$
- $x_{cheesesauce} \leq M y_{cheesesauce}$
- $y_{cheesesauce} = 1$
- $x_{broccoli} \leq M y_{peanutbutter}$ ✔
- $x_{broccoli} \geq M y_{peanutbutter}$

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 3

1.4/1.4 points (graded)

Select the mathematical constraint that corresponds to the following English sentence:

Either cheese sauce or peanut butter (or both) must be eaten with broccoli.

- $y_{peanutbutter} + y_{cheesesauce} = 0$
- $y_{peanutbutter} = 1 - y_{cheesesauce}$
- $y_{broccoli} \leq y_{cheesesauce} + y_{peanutbutter}$ ✔
- $y_{broccoli} + y_{cheesesauce} + y_{peanutbutter} \leq 2$
- $x_{cheesesauce} \leq M y_{cheesesauce}$
- $y_{cheesesauce} = 1$
- $x_{broccoli} \leq M y_{peanutbutter}$
- $x_{broccoli} \geq M y_{peanutbutter}$

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 4a

5.0/5.0 points (graded)

A company has created a stochastic discrete-event simulation model of its customer service call center, including call arrivals, resource usage (workers who specialize in answering each type of calls, supervisors, etc.), and call duration.

The call center is not first-come-first-served; a call from a major client will be answered first, ahead of even long-waiting callers with smaller accounts.

When a new call comes in, the call center will run the simulation to quickly give the caller an estimate of the expected wait time before being helped.

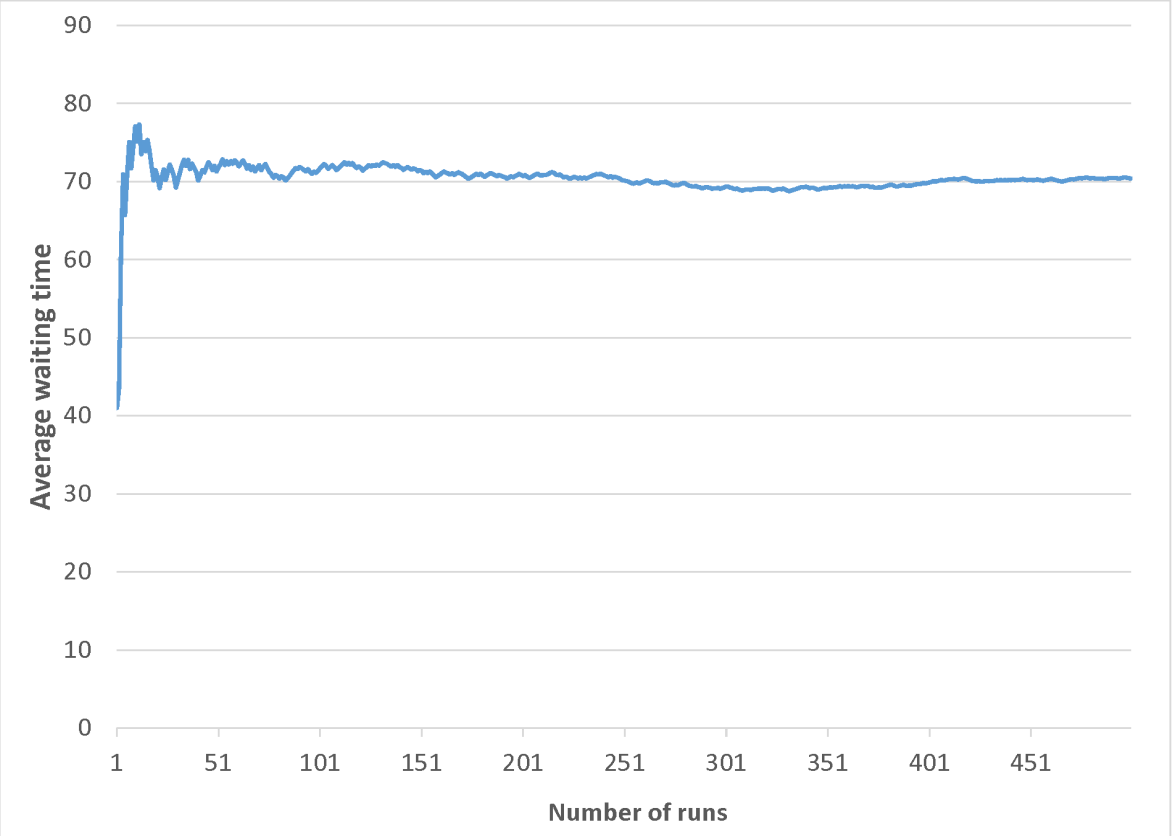
How many times does the company need to run the simulation for each new caller (i.e., how many replications are needed)?

- Once, because the outcome will be the same each time
- Many times, because of the variability and randomness ✔
- Once, because each patient is unique

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Question 4b



The figure above shows the average of the first *x* simulated wait times, as new replications ("runs") are run and added into the overall average. It is not showing the wait time just for each replication. For example, after *x*=101 replications, the wait time of the 101st replication is not necessarily 72, but the average of those 101 replications is about 72.

Question 4b

0.0/5.0 points (graded)

If the goal is to report the expected wait time to within +/- 2 minutes, what can you conclude from the figure above? Select all of the answers that are correct.

- [x] The simulation could have been stopped after 400 runs (replications). ✔
- [x] The simulation could even have been stopped after 300 runs (replications). ✔
- [x] The simulated wait time was 50 or less just once out of all the runs (replications).

- [ ] The expected wait time of simulated runs (replications) is likely to be between 75 and 85.
- [ ] There is significant variability in the simulated wait time of the runs (replications). ✔

✖

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 4c

6.0/6.0 points (graded)

Suppose it is discovered that simulated wait times are 50% higher than actual wait times, on average. What would you recommend that they do?

- ( ) Scale down all estimates by a factor of 1/1.50 to get the average simulation estimates to match the average actual wait times.
- (•) Investigate to see what's wrong with the simulation, because it's a poor match to reality. ✔
- ( ) Use the 50%-higher estimates, because that's what the simulation output is.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Question 5

For each of the optimization problems below, select its most precise classification. In each model, x are the variables, all other letters (a,b,c) refer to known data, and the values of c are all positive.

There are seven questions labeled "Question 5". Answer all seven questions. Each classification might be used, zero, one, or more than one time in the seven questions.

Question 5

1.0/1.0 point (graded)

Minimize $\sum_i c_i x_i$

subject to $\sum_i a_{ij} x_i \geq b_j$ for all j

all $x_i \geq 0$

Linear program ▼ ✔ **Answer:** Linear program

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5

1.0/1.0 point (graded)

Minimize $\sum_i (\log c_i) x_i$

subject to $\sum_i a_{ij} x_i \geq b_j$ for all j

all $x_i \geq 0$

Linear program ▼ ✔ **Answer:** Linear program

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5

1.0/1.0 point (graded)
Minimize $\sum_i c_i x_i^2$

subject to $\sum_i a_{ij} x_i \geq b_j$ for all j

all $x_i \geq 0$

Convex quadratic program ▼ ✔ **Answer:** Convex quadratic program

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5

1.0/1.0 point (graded)
Minimize $\sum_i c_i x_i$

subject to $\sum_i a_{ij} x_i \geq b_j$ for all j

all $x_i \in \{0, 1\}$

Integer program ▼ ✔ **Answer:** Integer program

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5

1.0/1.0 point (graded)
Minimize $\sum_i c_i \sin x_i$

subject to $\sum_i a_{ij}x_i \geq b_j$ for all j

all $x_i \geq 0$

General non-convex program ▼ ✔ **Answer:** General non-convex program

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5

0.0/1.0 point (graded)
Minimize $\sum_i c_i|x_i - 6|$

subject to $\sum_i a_{ij}x_i \geq b_j$ for all j

all $x_i \geq 0$

Linear program ▼ ✖ **Answer:** Convex program

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 5

0.0/1.0 point (graded)
Minimize $\sum_i c_i x_i$

subject to $\sum_i \sum_k a_{ikj}x_i x_k \geq b_j$ for all j

all $x_i \geq 0$

Linear program ▼ ✖ **Answer:** General non-convex program

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Questions 6a,6b,6c

6.68/12.0 points (graded)

A supermarket is analyzing its checkout lines, to determine how many checkout lines to have open at each time.

At busy times (about 10% of the times), the arrival rate is 5 shoppers/minute. At other times, the arrival rate is 2 shoppers/minute. Once a shopper starts checking out (at any time), it takes an average of 3 minutes to complete the checkout.

[NOTE: This is a simplified version of the checkout system. If you have deeper knowledge of how supermarket checkout systems work, please do not use it for this question; you would end up making the question more complex than it is designed to be.]

a. The first model the supermarket tries is a queuing model with 4 lines open at all times. What would you expect the queuing model to show?

- Wait times are low at both busy and non-busy times.
- Wait times are low at busy times and high at non-busy times.
- Wait times are low at non-busy times and high at busy times. ✖
- Wait times are high at both busy and non-busy times. ✔

b. The second model the supermarket tries is a queuing model with 20 lines open during busy times and 10 lines open during non-busy times. What would you expect the queuing model to show?

- Wait times are low at both busy and non-busy times. ✔
- Wait times are low at busy times and high at non-busy times.

- Wait times are low at non-busy times and high at busy times.
- Wait times are high at both busy and non-busy times.

c. The third model the supermarket tries is a Markov chain, where each state is the number of people waiitng (e.g., 0 people waiting, 1 person waiting, etc.). When there are at least 10 total people waiting (across all lines), the supermarket will open a new checkout line, which stays open until nobody is left waiting.

Select all of the following statements about the model and the memoryless property (previous states don't affect the probability of moving from one state to another) that are true.

- [ ] The process is memoryless.
- [x] If the arrivals follow the Poisson distribution and the checkout times follow the Exponential distribution, then the process is memoryless.
- [x] We can't say the process is memoryless because the model is not well-defined. For example, the transition probabilities from the state "3 people are waiting" depend on how many lines are currently open, which isn't captured as part of the states of the Markov chain. ✔

✱

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Questions 7a,7b

5.0/10.0 points (graded)

A charity is testing two different mailings to see whether one generates more donations than another. The charity is using A/B testing: For each person on the charity's mailing list, the charity randomly selects one mailing or the other to send. The results after 2000 trials are shown below.

	Trials	Donation rate	95% confidence interval
Option A	1036	4.8%	3.6%-6.2%
Option B	964	5.2%	3.8%-6.6%

Note: Higher donation rates are better.

a. What should the retailer do?

- Switch to exploitation (utilize Option A only; A is clearly better)
- Switch to exploitation (utilize Option B only; B is clearly better)
- More exploration (test both options; it is unclear yet which is better) ✔

Later, the charity developed 7 new options, so they used a multi-armed bandit approach where each option is chosen with probability proportional to its likelihood of being the best. The results after 2000 total trials are shown below.

	Donation rate	95% confidence interval
Option #1	3.2%	0.8%-6.3%
Option #2	4.2%	1.2%-7.2%
Option #3	5.2%	2.4%-8.5%
Option #4	5.5%	2.7%-8.6%
Option #5	6.5%	3.8%-9.5%
Option #6	10.8%	8.0%-13.8%
Option #7	15.0%	12.1%-17.9%

Note: Higher donation rates are better.

b. What should the retailer do?

- Move to pure exploitation: use only Option 7 ✖
- Continue the multi-armed bandit approach using only Options 6 and 7 ✔

- Continue the multi-armed bandit approach using only Options 5, 6, and 7
- Continue the multi-armed bandit approach using all seven options

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Information for Question 8a

For each of the mathematical optimization models, select the variable-selection/regularization method it most-precisely represents (or select "none of the above" if none of the other choices are appropriate). In each model, x is the data, y is the response, a are the coefficients, n is the number of data points, m is the number of predictors, and T and λ are appropriate constants.

There are four questions labeled "Question 8a". Answer all four questions. Each of the choices might be used zero, one, or more than one time in the four questions.

Question 8a

1.0/1.0 point (graded)

Minimize $\sum_{i=1}^{n} (y_i - (a_0 + \sum_{j=1}^{m} a_j x_i j))^2$

subject to $\lambda \sum_{j=1}^{m} |a_j| + (1-\lambda) \sum_{j=1}^{m} (a_j)^2 \leq T$

Elastic net ▼ ✔ **Answer:** Elastic net

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 8a

1.0/1.0 point (graded)

Minimize $\sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{j=1}^{m} a_j x_i j\right)\right)^2$

subject to $\sum_{j=1}^{m} |a_j| \leq T$

Lasso regression ▼ ✔ **Answer:** Lasso regression

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 8a

1.0/1.0 point (graded)

Minimize $\sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{j=1}^{m} a_j x_i j\right)\right)^2$

None of the above ▼ ✔ **Answer:** None of the above

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 8a

1.0/1.0 point (graded)

Minimize $\sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{j=1}^{m} a_j x_i j\right)\right)^2$

subject to $\sum_{j=1}^{m} (a_j)^2 \leq T$

Ridge regression ▼ ✔ **Answer:** Ridge regression

Submit You have used 1 of 1 attempt

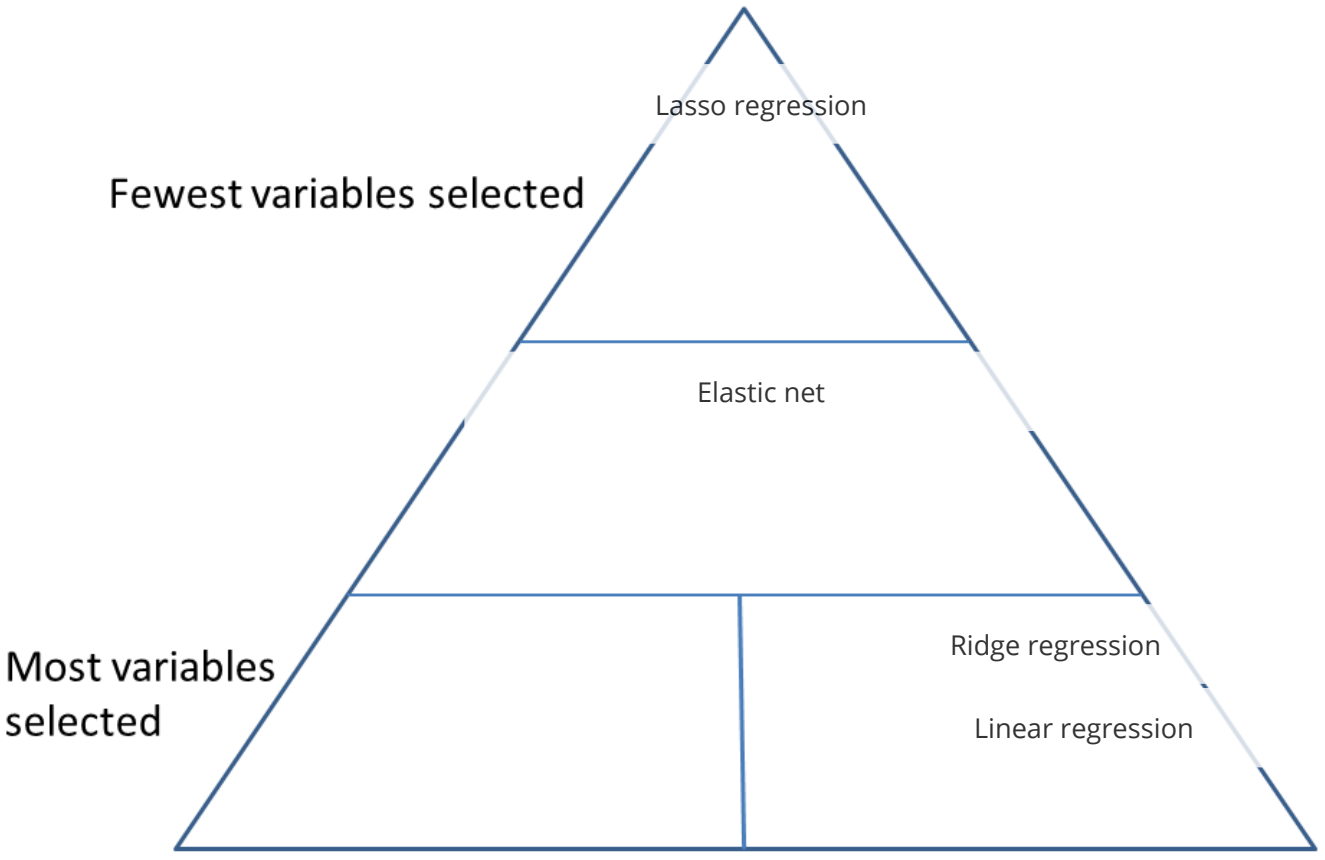
Answers are displayed within the problem

Question 8b

4/4 points (graded)

Keyboard Help

Rank the following regression and variable-selection/regularization methods from fewest variables selected to most variables selected. All four methods will be used (the bottom contains two equivalent spaces).



Submit You have used 1 of 1 attempts.

Reset Show Answer

FEEDBACK

✔ Correctly placed 4 items.

✔ Good work! You have completed this drag and drop problem.

✔ Final attempt was used, highest score is 4.0

Question 8c

6.0/6.0 points (graded)

Select all of the following reasons that you might want to use stepwise regression, lasso, etc. to limit the number of factors in a model.

- ☐ To find a more-complex model
- ☑ To find a simpler model ✔
- ☑ Because there isn't enough data to avoid overfitting a model with many factors ✔

✔

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Question 8d

3.0/3.0 points (graded)

In the simple linear regression model

$$minimize \sum_{i=1}^{n} (y_i - (a_0 + \sum_{j=1}^{m} a_j x_{ij}))^2$$

i. What are the variables from a regression perspective?

- (●) Only x_{ij} ✔
- ( ) Both x_{ij} and a_j
- ( ) Both x_{ij} and y_i ✔
- ( ) Only a_j
- ( ) Only y_i

ii. What are the variables from an optimization perspective?

- ( ) Only y_i
- ( ) Both x_{ij} and a_j
- (●) Only a_j ✔
- ( ) Only x_{ij}
- ( ) Both x_{ij} and y_i

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 8e

3/7 points (graded)

Keyboard Help

Put the following seven steps in order, from what is done first to what is done last.

Step 1	Remove outliers
Step 2	
Step 3	Impute missing data values Scale data
Step 4	Fit lasso regression model on all variables
Step 5	Fit linear regression, regression tree, and random forest models using variables chosen by lasso regression
Step 6	Pick model to use based on performance on a different data set
Step 7	Test model on another different set of data to estimate quality

Submit You have used 1 of 1 attempts.

Reset | Show Answer

FEEDBACK

✔ Correctly placed 3 items.

✖ Misplaced 4 items.

ℹ Good work! You have completed this drag and drop problem.

✱ Final attempt was used, highest score is 3.0

There are five questions labeled "Question 9". Answer all five questions. For each question, select the most appropriate model/approach to answer the question/analyze the situation described. Each model/approach might be used zero, one, or more than one time in the five questions.

Question 9

1.4/1.4 points (graded)

Which groups of genetic markers often appear together in people?

Louvain algorithm ▼ ✔ **Answer:** Louvain algorithm

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Question 9

1.4/1.4 points (graded)

Is the median age of MS Analytics students different between the on-campus and online cohorts?

Non-parametric test ▼ ✔ **Answer:** Non-parametric test

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 9

1.4/1.4 points (graded)

How many checkout lanes should a supermarket open to avoid long wait times?

Queuing ▼ ✔ **Answer:** Queuing

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 9

1.4/1.4 points (graded)

In the MS Analytics program, which groups of electives are often taken by the same students?

Louvain algorithm ▼ ✔ **Answer:** Louvain algorithm

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Question 9

0.0/1.4 points (graded)

Estimate the number of workers required to work at a call center based on call arrivals and lengths.

Stochastic optimization ▼ ✖ **Answer:** Queuing

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

This the end of Midterm Quiz 2. Please make sure that you submit all your answers before the time runs out. There is no overall Submit button.