ISYE 6501x Introduction to Analytics Modeling

Sample Quiz #1 Questions

NOTES

1. The real quiz will have more questions, and cover more material; these questions are just meant to give you an idea of the question style and depth.

2. Because of the online format, I will try to make some of the answers more structured than the purely-free-answer format in two of the questions below.

3. This is being posted early, because a bunch of you asked for it. Some of the topics covered below are things you'll see in the weeks between now and when you take the quiz, so if they don't look familiar yet, don't worry!

NAME_____

ISYE 6501x, Introduction to Analytics Modeling

Quiz #1 – 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!

1. Figure A below shows a linear classifier (dashed line) for a classification problem, using two predictors (x_1 and x_2) to separate between black and white points. Figure B shows a CART (classification tree) approach to the same problem.

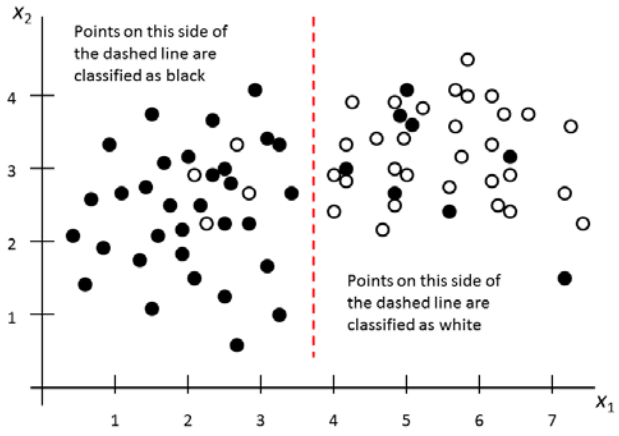


Figure A

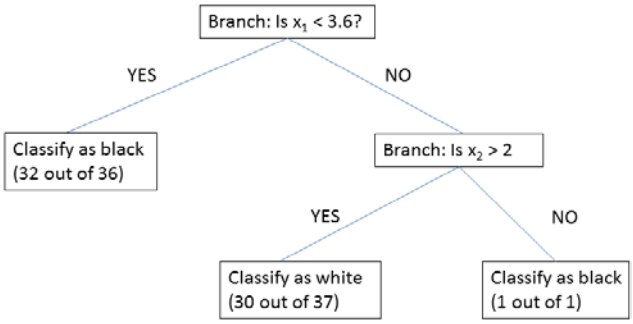


Figure B

In each leaf, "a out of b" means that there are b data points in the leaf, and a of them are classified correctly using the leaf's answer.

a. In Figure A, which predictor (x_1 or x_2) is <u>not</u> important for separating between the black and white points in this model?

(CORRECT ANSWER: x_2. The classifier is a vertical line, so all that matters is whether x_1 is larger or smaller than 3.5.)

b. In Figure B, both x_1 and x_2 are used to classify the points (even though one was unimportant in Figure A). Which classification model do you think is better (Figure A or Figure B), and why?

<u>CHOICES</u>

i. Figure A, because Figure B overfits the lower-rightmost leaf.
ii. Figure B, because it misclassifies 11 points, and Figure A misclassifies 12 points.
i. Figure B, because it uses both predictors for classification.
ii. Figure A, because it is a simpler model.

(CORRECT ANSWER: i. The lower-rightmost leaf has just one data point in it, a clear example of overfitting. Although ii is a true answer, it is not correct: B is not a better model even though it misclassifies one fewer point, because the apparent better fit is due to overfitting. iv might be a reasonable answer in general, but in this case the overfitting of B overrides having or not having a slightly simpler model.

<u>NOTE:</u> As we saw in the lessons, as a rule of thumb each leaf should have at least 5% of the data points, and a common rule of thumb for a factor-based model is to have at least 10 times as many data points as factors selected.)

2. A geologist would like to build a model to predict the probability that a volcano will erupt in a given week. The geologist has previous eruption data, as well as several factors that can be used as predictors.

   a. Which of the following models would be most appropriate for the geologist to use to predict the probability of an eruption?

      CHOICES

      a. ARIMA
      b. CART
      c. Cross-validation
      d. CUSUM
      e. Exponential smoothing
      f. GARCH
      g. k-means clustering
      h. k-nearest-neighbor classification
      i. Linear regression
      j. Logistic regression
      k. Support vector machine

      (CORRECT ANSWER: j. Logistic regression is the model we've seen (o will see before the quiz for predicting probabilities from a set of factors.)

   b. Select all of the following models that would be appropriate for the geologist to use to classify data points into "eruption" or "not eruption".

      CHOICES

      a. ARIMA
      b. CART
      c. CUSUM
      d. k-nearest-neighbor
      e. Support vector machine

      (CORRECT ANSWERS: b,d,e (all three must be selected for full credit). All three of these are methods we've seen (or will see before the quiz) for classifying based on a set of factors.)

   b. The geologist used a simple 50% threshold: if the model predicts a probability of 50% or more that the volcano will erupt, the geologist recommends that the nearby towns be evacuated. Do you think 50% is the right threshold, or should it be higher or lower? Explain why.

      (CORRECT ANSWER: I accepted any answer that made sense. Many people suggested using a lower threshold, since lives would be at stake: even if the probability of eruption is small, the towns should be evacuated. Others suggested a higher threshold, also to save lives, because of the crying-wolf effect: if the towns were evacuated a couple of times and there was no eruption, people would be less likely to evacuate again, even if the model suggested a higher probability. And there were other answers I accepted too. Basically, if your answer used analytics reasoning correctly, you got credit; if it didn't use analytics reasoning correctly, you didn't get credit.)

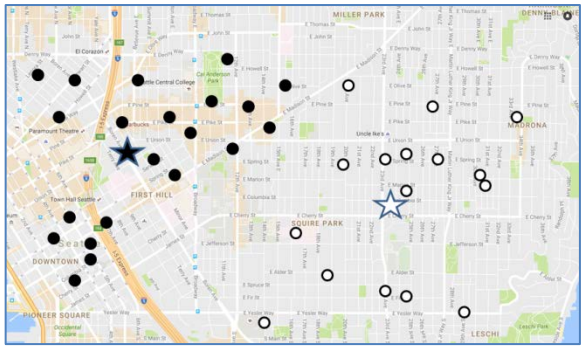
c. Instead of building a model to predict probability, suppose the geologist measures the magnitude of tremors in the area over time. When the magnitude of the tremors changes (gets much larger), the geologist will recommend evacuating nearby towns. Which of the following models would be most appropriate for the geologist to use?

CHOICES

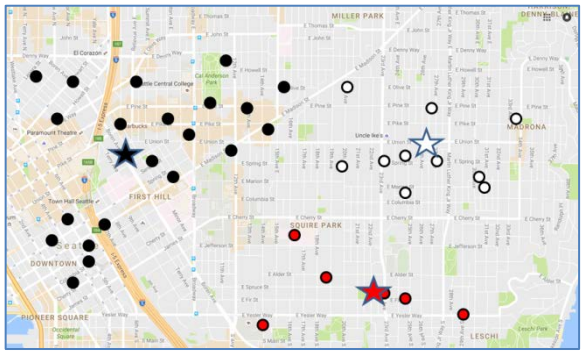
a. ARIMA
b. CART
c. Cross-validation
d. CUSUM
e. Exponential smoothing
f. GARCH
g. k-means clustering
h. k-nearest-neighbor classification
i. Linear regression
j. Logistic regression
k. Support vector machine

(CORRECT ANSWER: d. CUSUM is the model we've seen for directly detecting changes.)

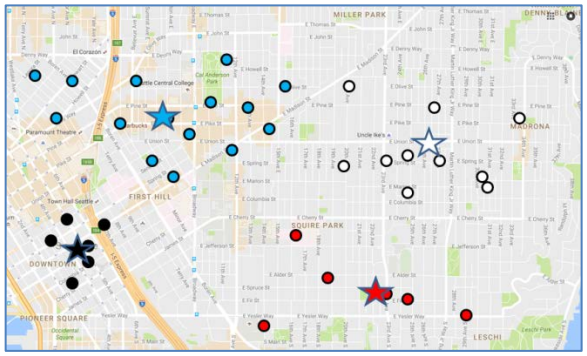
3. A police department has used a *k*-means approach to find geographic clusters of instances of crime in their city. Their goal is to find locations for new police patrols to reduce crime in high-crime areas. The figures below show the solutions for *k*=2, *k*=3, and *k*=4; circles are crime instances and stars are cluster centers.



k=2



k=3



k=4

Based on how far officers can effectively patrol, the police department initially selected the *k*=4 solution. However, they then realized that some of the crimes were committed during the day and others at night, and they might have two sets of new patrols, one set during the day and set one at night. How would you suggest the police department redo or change its analysis?

(CORRECT ANSWER: I accepted any answer that made good analytics sense. The most common one (and the one I was expecting) is that the department could do two clusterings: one using only daytime data points and one using only nighttime data points, and create daytime patrols and nighttime patrols based on those two solutions.)

4. Recall the equations for triple exponential smoothing (Winters'/Holt-Winters method):

$$S_t = \alpha \frac{x_t}{C_{t-L}} + (1-\alpha)(S_{t-1} + T_{t-1})$$
$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$
$$C_t = \gamma \frac{x_t}{S_t} + (1-\gamma)C_{t-L}$$

A construction vehicle manufacturer wants to use this model to analyze a production process where construction vehicles are produced in batches of exactly 170, and a batch takes an average of 9 days to be completed (usually between 8 and 10). Our data includes the day each vehicle's production is completed, its sequence in the batch (e.g., 57th out of 170), the day within the batch that it was completed (e.g., completed on the 3rd day the batch was being produced), and the number of hours the vehicle operated before its first breakdown.

Vehicle ID	Sequence in batch	Day within batch that vehicle was produced	Hours of operation before first breakdown
047-92-1HA	56	3	1570
091-46-7ZQ	57	3	2349
854-A9-21B	58	3	3016
620-88-4GA	59	4	2201
Etc.			

Based on this data, the manufacturer wants to use a triple exponential smoothing model to determine whether any patterns exist in the number of hours before the first breakdown, based on a vehicle's sequence number in its batch.

For each of the mathematical terms on the left, pick the appropriate number or description from the right.

a. x_t
b. L

i. 170
ii. 9
iii. Sequence in batch
iv. Day within batch that vehicle was produced
v. Hours of operation before first breakdown

(CORRECT ANSWERS:

a. v. x_t is the observed value of the response, the hours of operation before the first breakdown.

b. i. L is the length of the cycle. Since the question specifies "a vehicle's sequence number in its batch", the cycle length is the 170 vehicles in each batch.)

c. If the manufacturer observes that the values of C are generally close to 1, except that they are significantly lower than 1 for vehicles built near the beginning of batches, what can be concluded?

CHOICES

i. There is no effect of sequence in batch on the number of hours before the first breakdown.
ii. Vehicles built early in a batch tend to break down more quickly.
iii. Vehicles built early in a batch tend to break down more quickly, because workers are adjusting to the different specifications in a each new batch.
iv. Vehicles built early in a batch tend to take longer to break down.
v. Vehicles built early in a batch tend to take longer to break down, because workers are paying closer attention to their work early in each new batch.

(CORRECT ANSWER: ii. Values of C less than 1 mean that the response (hours before first breakdown) is lower, so those vehicles tend to break down sooner. However, all this model can do is make the observation; there's nothing in it to explain *why* the effect is observed – so although iii. might sound like it makes sense, the model does not say anything about causality.)

d. If the values of T tend to be slightly positive, what can be concluded?

CHOICES
i. Vehicles built more recently tend to take longer to break down.
ii. Vehicles built more recently tend to break down more quickly.

(CORRECT ANSWER: i. Positive values of T mean that the response is getting higher over time, so newer vehicles' responses (time until first breakdown) tend to be higher. So, vehicles built more recently tend to take longer to break down for the first time.

e. Suppose the manufacturer wanted to use a regression model to answer the same question, using the same data: two predictors (sequence in batch and day within batch) and one response (hours of operation before first breakdown).

If the manufacturer first used principal component analysis on the data, what would you expect?

CHOICES
i. The first component would be much more important than the second.
ii. The second component would be much more important than the first.
iii. The two components would have approximately the same importance.

(CORRECT ANSWER: i. With 170 vehicles produced in 8-10 days, the two predictors (sequence in batch, and day within batch) will be highly correlated. So, the second component will have much less importance, because its effect will only be whatever is uncorrelated with the first component. Consequently, in the PCA results the first eigenvalue will be much larger than the second.)