Math 154, Exam 2 - sample questions

no solutions available

Fall 2021

1. You have just been hired as a data scientist for a tech startup. You and your new colleague, Kathy, have been given a data set X, consisting of n observations of p explanatory variables, and a binary response vector y of length n. [You may assume that n is several thousand and p is several dozen.] Kathy reports that her classification model, f, achieves an error rate of 11% on in-sample testing. That is, Kathy used X and y to build f, then computed $f(X) = \hat{y}$, compared $\hat{y}$ to y, and found that they agreed 89% of the time.

a. Explain to Kathy why 11% is not necessarily an accurate estimate of her model's true error rate.
b. Carefully explain how a 3-fold cross-validation scheme could be used to estimate the true error rate of f. Be as specific as you can.

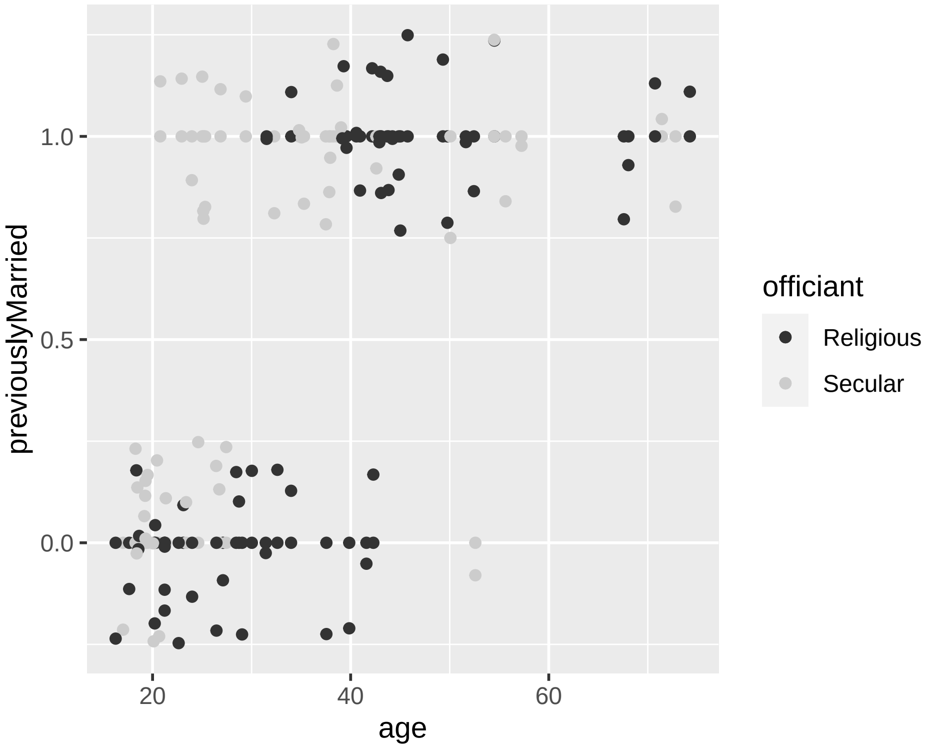
2. Consider supervised vs unsupervised learning.

a. What is the difference between the two?
b. Name a few methods that we have covered which fall into each category.
c. Name a few methods that we have covered which don't belong to either supervised or unsupervised learning.

3. The data set `Marriage` contains data 98 marriage records from the Mobile, AL probate court. Binary variables are available about whether the marriage was presided over by a religious or secular official, the bride's age, and whether she had been previously married. Consider the following (decision) classification tree for whether the officiant was religious or secular.

```
## n= 98
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 98 46 Religious (0.530612 0.469388)
##    2) age>=39.093 35  8 Religious (0.771429 0.228571) *
##    3) age< 39.093 63 25 Secular (0.396825 0.603175)
##      6) previouslyMarried< 0.5 42 19 Religious (0.547619 0.452381)
##       12) age>=27.942 8  0 Religious (1.000000 0.000000) *
##       13) age< 27.942 34 15 Secular (0.441176 0.558824)
##         26) age>=20.163 18  8 Religious (0.555556 0.444444) *
##         27) age< 20.163 16  5 Secular (0.312500 0.687500) *
##      7) previouslyMarried>=0.5 21  2 Secular (0.095238 0.904762) *
```

a. Draw the decision boundaries on the plot below, labeling each box clearly with either RELIGIOUS or SECULAR. [Note: Some jitter has been applied to the y values to make them easier to see.]



b. Betsy is a 24-year-old first-time bride. Do you predict that she will have a religious or secular ceremony? How confident are you in that prediction?
c. Briefly summarize how the classification model informs your understanding of the real-world question. That is, describe the model in your own words.

4. Consider a classification problem using kNN. We have n training points $x_1, x_2, \ldots, x_n$ corresponding to labels $y_1, y_2, \ldots, y_n$.

Is it possible to build a classification tree (which splits on the i^{th} variable at each node) which behaves exactly the same as $k = 1$ kNN classifier?

5. Consider training a decision tree on n two dimensional vectors $x = (x_1, x_2)$.

a. Assume we have two equal vectors x and x' in our training set (that is, all attributes of x and x' are exactly the same). Can removing x' from our training data change the decision tree we learn for the data set? Explain.
b. Assume that the training instances are linearly separable. That is, there exists a $\{\mathbf{w}, b\}$ such that

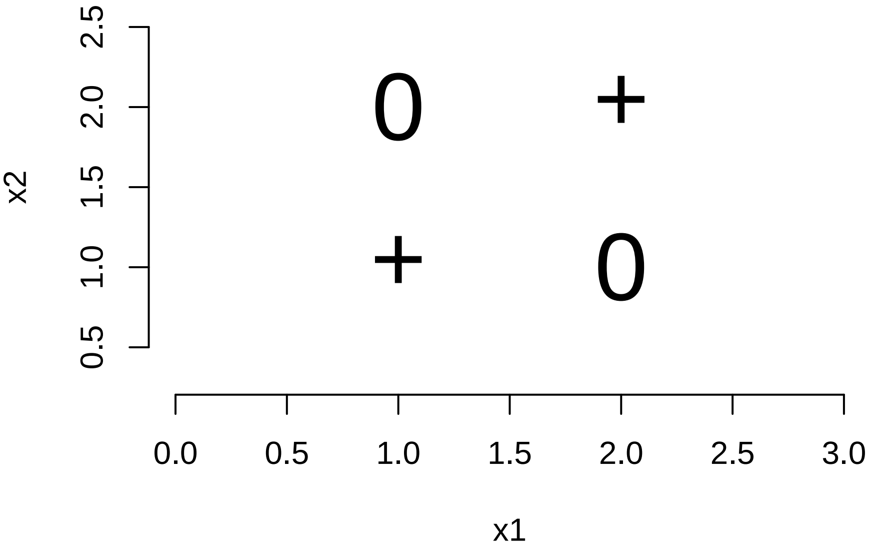
$$y_i = \begin{cases} +1 & \mathbf{w} \cdot \mathbf{x}_i + b > 0 \\ -1 & \mathbf{w} \cdot \mathbf{x}_i + b \leq 0 \end{cases}$$

Can a decision tree correctly classify these vectors? If so, what is an upper bound on the number of nodes corresponding to the perfect classification tree? If not, why not?

c. Now assume that the n inputs are not linearly separable (that is, no $\{w, b\}$ such that the points can be correctly classified using the above rule). Can a decision tree correctly classify these vectors? If so, what is an upper bound on the number of splits corresponding to the perfect classification tree? If not, why not?

6. TRUE/FALSE: The training error for 1-NN classifier is zero.

7. Consider the following data set:



Which of the following classifiers will achieve zero training error on this data set?

a. SVM with quadratic kernel.
b. Tree with depth of 2.
c. 3-NN classifier

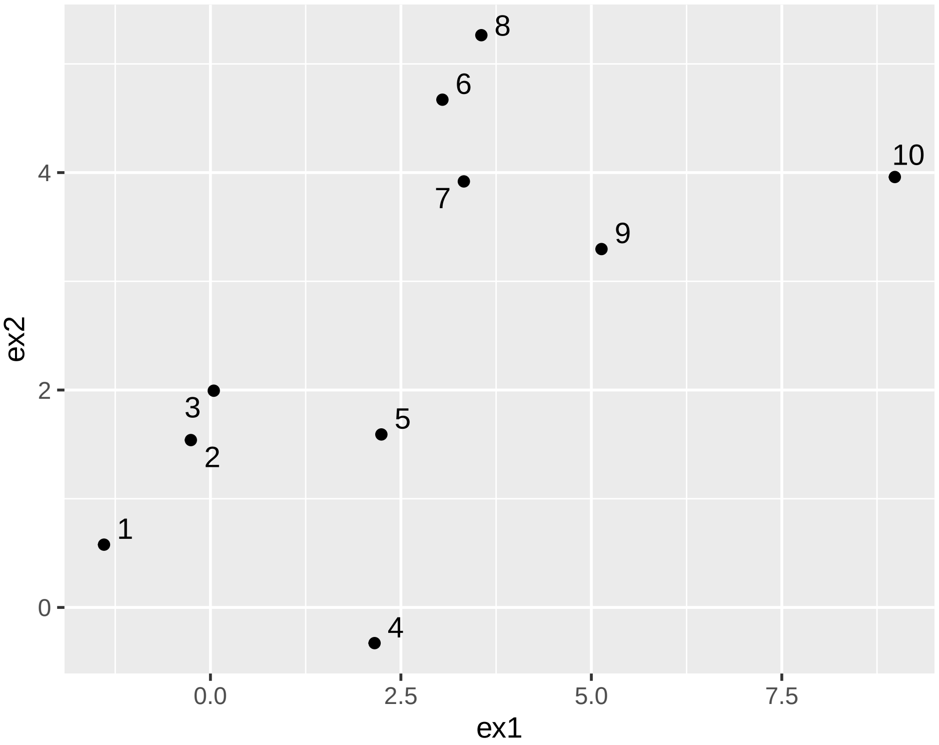
8. Why does the kernel trick allow us to solve SVMs with high dimensional feature spaces, without substantially increasing the running time?

9. Suppose we get some training data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$, and a SVM with the radial basis function is fit to the data with fixed parameters $\gamma = 1$ and $C = 1$.

a. Suppose we observe a good fit to the training data (training error is low). How do you think the trained SVM will predict test data from the same source?
b. Now suppose we do not observe a good fit for $\gamma = 1$, but by playing around, we see that the fit is quite good for $\gamma = 0.02$. We decide to predict future data using $\gamma = 0.02$. What can we expect from the test error?
c. Is there a better way to optimize γ? Explain.
d. Suppose again that the fit for $\gamma = 1$ is not very good. Rather than playing with γ, we decide to use a different kernel. People have tried many different kernels in the literature. Suppose we try 100 different kernels on the training data (without tuning the parameters). Suppose that one of them fits the data really well. So we decided to predict the test data using that kernel. What do you expect from the test error?

10. Consider the following dataset.

```
## 'data.frame':    10 obs. of  3 variables:
##  $ point: int  1 2 3 4 5 6 7 8 9 10
##  $ ex1  : num  -1.3973 -0.2572 0.0436 2.1547 2.2439 ...
##  $ ex2  : num  0.578 1.54 1.994 -0.328 1.591 ...
```



```
##        1     2     3     4     5     6     7     8     9
## 2   1.80
## 3   2.84  1.14
## 4   4.74  3.65  3.29
## 5   5.50  3.91  3.00  2.17
## 6   7.84  6.06  5.02  5.46  3.34
## 7   8.34  6.60  5.52  5.33  3.25  1.28
## 8   9.77  8.03  6.93  7.02  4.92  2.15  1.69
## 9  10.68  9.01  7.98  6.86  5.22  3.90  2.77  2.71
## 10 14.15 12.46 11.52 10.05  8.72  7.20  6.40  5.93  4.03
```

Suppose you are given initial assignment cluster center as Cluster1: point #1, Cluster2: point #10.

a. Using the (diagonal) distance matrix provided, assign the points to the two clusters, given the initial cluster centers.
b. Find the two updated cluster centers. Use the (diagonal) distance matrix, not using Euclidean distance. That is, find the observation which is closest to the rest of the points (as measured by the sum of the distances) in the given cluster.

11. How are the individual trees in a random forest different from:

a. a classification / regression tree?
b. a bagged tree?

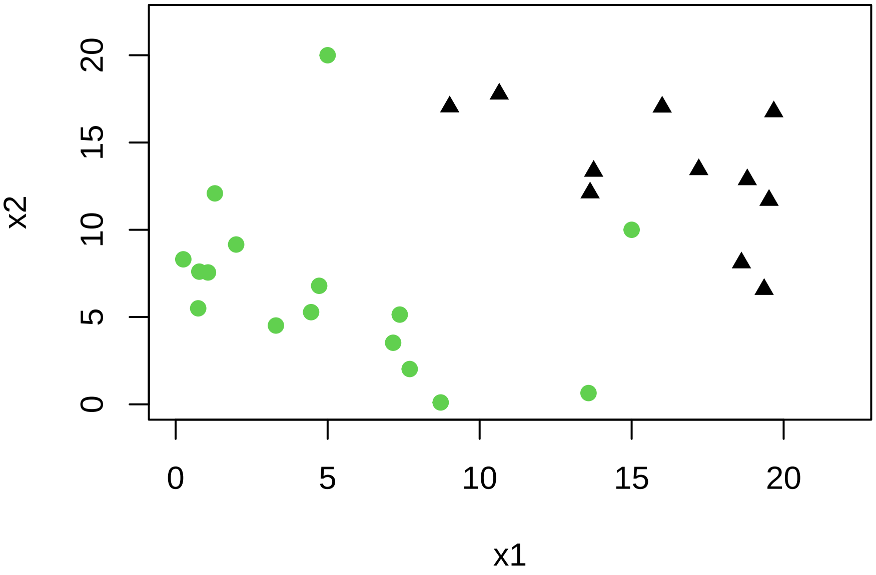
12. How can out of bag samples be used to obtain unbiased error estimates on future data? Why are Random Forests especially suitable for getting accurate error estimates using out of bag data?

13. If I use N-Fold cross validation to select the tuning parameters, is the error rate that is returned by the N-fold cross validation procedure a good indication of how well the model built by my algorithm will do on future data? If not, what is? (in other words, how can I compare how well two different learning algorithms will do on a specific data set?)

14. Consider each of the classification methods we have covered.

a. For each one, what are the associated tuning parameters that need to be set (/optimized/cross validated) by the user?
b. Some of the tuning parameters are given to make the model fit better, and some are given to make the model more constrained. Which are which?
c. The α parameter in CART, and the C parameter in SVM are both specifically designed to keep the model from perfectly fitting the training data.
   i. Why do we need to tune α and C to keep the training data from being perfectly fit?
   ii. How do we tune α and C to keep the training data from perfectly fitting? That is, how are α and C estimated?

15. Consider the following data.



a. What would be the decision boundary in this problem for a very large value of C ($C \to \infty$)? (Where C is the cost parameter associated with SVMs.) Sketch the boundary on the plot and give a one sentence justification in your own words.
b. For $C \approx 0$, where would the decision boundary be? Sketch the boundary on the plot and give a one sentence justification in your own words.
c. Which of the two decision boundaries is likely to work better on the test data? Justify in your own words.
d. Add a point to the figure which will not change the decision boundary learned for a very large value of C. Explain your reasoning.
e. Add a point to the figure which will not change the decision boundary learned for a very small value of C. Explain your reasoning.

16. In random forests, the variable importance is calculated separately for

a. each class (response variable): TRUE or FALSE
b. each observation: TRUE or FALSE

17. For an individual variable, the quantity of "reduction in sum of squares" (or "reduction in node purity" for classification) would be straightforward to calculate for any tree model. However, variable importance is more meaningful in Random Forests than in CART. Why?

18. A friend says to you "I know that to tune a classification model, I need to cross validate. So I broke up the data into 10 subsets, found the CV error (model: 9/10 of the data, test: 1/10 of the data - repeated 10 times) for each value of $\gamma = (0.1, 0.5, 1, 10, 100)$. I chose the γ value with the lowest CV error rate ($\gamma = 1$), and I plan to use that γ value to build the final model from the entire dataset. I know the final model will have an approximate error rate (on future test data) of approximately 0.047, because that is

the error rate from cross validating."

Which of the following is true?

a. Your friend is correct.
b. Your friend is correct to use $\gamma = 1$ but not correct to consider the error rate of 0.047 as representative of future test data.
c. Your friend is wrong about $\gamma = 1$ but correct that the test data error rate will be close to 0.047.
d. Your friend is wrong about both γ and the test data error rate.

Explain.

19. Consider the following similarities (note the word similarity, not distance!) between five points:

	A	B	C	D	E
A	1.00	0.92	0.35	0.22	0.21
B	0.92	1.00	0.61	0.44	0.16
C	0.35	0.61	1.00	0.37	0.10
D	0.22	0.44	0.37	1.00	0.33
E	0.21	0.16	0.10	0.33	1.00

a. Cluster the five points using complete linkage. Draw a dendrogram to depict the clustering you obtain.

b. Some clustering algorithms only work on distance matrices. A standard way to convert similarity matrices S_{ij} to distances D_{ij} is as follows:

$$D_{ij} = 1 - S_{ij}.$$

What if you had used the conversion

$$D_{ij} = \sqrt{S_{ii} - 2S_{ij} + S_{jj}}?$$

Will this result in the same clustering?

c. Repeat (b) when using group average linkage instead of complete linkage.

20. Consider the Amazon recommendations of the form: "People who bought this, also bought…" Those recommendations are most likely a result of which learning method we have covered? Explain.

21. We usually use standardize the explanatory variables before using the radial basis kernel in SVM. What is true about standardizing the explanatory variables?

i. We do variable standardization so that no variable will dominate the others.
ii. Sometimes, variable standardization is not feasible.
iii. Variable standardization always helps when we use radial basis kernel in SVM.

a. i.

b. i. and ii.

c. i. and iii.

d. ii. and iii.