CS 4646 Exam 1

June 21, 2018

• This is a closed book exam.
• No electronics are permitted.
• Once you exit the room you will not be allowed to continue the exam.
• You must fill-in the bubble(s) corresponding to your answer(s) in the provided bubble sheet.
• You must turn in all pages, otherwise you will receive a 0 on the exam.

Good Luck.

Last, First Name:__________________________________________
<table>
<thead>
<tr>
<th>Choice</th>
<th>Fill in oval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>2</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>3</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>4</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>5</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>6</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>7</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>8</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>9</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>10</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>11</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>12</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>13</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>14</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>15</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>16</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>17</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>18</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>19</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>20</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>21</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>22</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>23</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>24</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>25</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>26</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>27</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>28</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>29</td>
<td>a b c d e f g h</td>
</tr>
<tr>
<td>30</td>
<td>a b c d e f g h</td>
</tr>
</tbody>
</table>
Part I: Multiple Choice (Fill bubble on bubble sheet)

1. a b c d Consider kNN, linear regression (LR), and decision tree (DT) learning (using correlation for splitting). Which option correctly lists the methods from fastest to slowest in query time?
   a. kNN, DT, LR
   b. kNN, LR, DT
   c. DT, kNN, LR
   d. LR, kNN, DT
   e. DT, LR, kNN
   f. LR, DT, kNN

2. a b c d Consider kNN, linear regression (LR), and decision tree (DT) learning (using correlation for splitting). Which option correctly lists the methods from fastest to slowest in training time?
   a. kNN, DT, LR
   b. kNN, LR, DT
   c. DT, kNN, LR
   d. LR, kNN, DT
   e. DT, LR, kNN
   f. LR, DT, kNN

3. a b c d Consider two decision trees trained on the exact same data. DT was trained using correlation for splitting, RT was trained using splits determined randomly. Both trees were trained with leaf_size = 1. Which option below correctly describes (in order): The fastest to train, the fastest to query, the highest accuracy on in-sample data?
   a. DT, DT, DT
   b. RT, about the same, about the same
   c. RT, RT, about the same
   d. DT, RT, about the same

4. a b c d Consider overfitting when using kNN and decision trees. When overfitting occurs with these two methods, in which “direction” does it occur?
   a. As k increases; As leaf_size increases
   b. As k increases; As leaf_size decreases
   c. As k decreases; As leaf_size decreases
   d. As k decreases; As leaf_size increases

5. a b c d Which of the following is NOT an advantage of ensemble learners?
   a. Less susceptibility to overfitting.
   b. Reduction in RMSE when using weak learners.
   c. Improved computational speed compared to individual weak learners.
   d. Can work with many types of weak learners.

6. a b c d Which of the following is NOT typically used as a means of assessing a learning algorithm?
   a. The standard deviation of predicted values.
   b. Compute time for training.
   c. Compute time for querying.
   d. Correlation of predicted values with known correct values.
7. a b c d Consider a data set composed of 1000 samples where \( X \) is drawn randomly uniformly from \( -2\pi \) to \( +2\pi \), and \( Y = \sin(X) \) (two full sine wave cycles). Consider kNN, decision trees, random trees and linear regression. Which statement is true regarding in-sample RMSE?
   a. Linear regression will perform best.
   b. kNN, decision trees and random trees will all do better than linear regression.
   c. kNN and linear regression will perform about the same.
   d. Decision trees will perform significantly better than random trees.

8. a b c d Suppose you are using one of the minimizing optimizers from ScikitLearn. You are using it to optimize your portfolio for MAXIMUM cumulative return, and \( port \_val \) are the daily total values of the portfolio for a particular allocation. Which of the following would be the best way to compute the objective function for the optimizer?
   a. objective = \( port \_vals \_max() \)
   b. objective = - \( port \_vals \_max() \)
   c. objective = \( port \_vals[-1] / port \_vals[0] \)
   d. objective = - \( port \_vals[-1] / port \_vals[0] \)

9. a b c d What distinguishes supervised from unsupervised learning?
   a. Supervised learning requires input data and expected outcomes for the input data, while unsupervised learning is not passed expected outcomes.
   b. Supervised learning uses unlabeled data for training while unsupervised learning uses labeled data.
   c. Supervised learning uses k-nearest neighbor while unsupervised learning uses decision forests.
   d. Supervised learning can achieve 0 RMSE with 1 for correlation while it is not possible to achieve the same results using unsupervised learning.

10. a b c d Which one of the following is a regression problem?
    a. An unknown animal has four legs, a long neck, 10 feet high and only eats grass. What animal is it?
    b. I am considering to purchase a laptop computer with \( X \) gigs of disc space, with a \( Y \) inch screen, from \( Z \) manufacture. How much should it pay?
    c. I have a fruit that is round shaped, red color and starts with an alphabet a. What fruit is it?
    d. This country celebrates Christmas in summer season, speaks English, has multiple time zones. What is this country?

11. a b c d For which of the following problems will a regression learning algorithm be applicable as opposed to a classification algorithm?
    a. Predicting percentage of breast cancer occurrences according to age-groups.
    b. Determining if a stock price will go up or down the next trading day using previous day stock prices.
    c. Identifying the breed of dog in a picture.
    d. Predicting the outcome of 2016 US presidential elections - who will be the next US president?
12. Which of the following statements is TRUE in regards to overfitting of linear regression models? Note: D = number of dimensions
   a. For linear regression models, overfitting is more likely to occur as D INCREASES
   b. For linear regression models, overfitting is more likely to occur as D DECREASES
   c. For linear regression models, overfitting does not occur
   d. For linear regression models, overfitting occurs, but is equally as likely for any value of D

13. Which of the following methods will benefit least from bagging. (Assume perfectly uniform random sampling and sufficiently high number of bags)
   a. kNN
   b. Decision Trees
   c. Random Forest
   d. Linear Regression
   e. Depends on the nature of the data.

14. Which decision tree algorithm is better in terms of performance for use in an ensemble learner
   a. JR Quinlan’s decision tree algorithm
   b. Adele Cutler’s random tree algorithm
   c. kNN algorithm
   d. Both A and B combined

15. The most likely reason for overfitting in a kNN classifier can be:
   a. Using low value of k . (It means considering less neighboring points)
   b. Using high value of k . (It means considering more neighboring points)
   c. Using Euclidean distance as a distance metric to measure distance between two points
   d. Using Manhattan distance as a distance metric to measure distance between two points.

16. Which of the following statements are true:
   1. Quinlan Trees select feature which has lowest correlation with output values
   2. Cutler Trees uses correlation/information gain to select feature to split on
   a. Statement 1
   b. Statement 2
   c. Both
   d. None

17. Which of the following three models, kNN, decision trees, and linear regression, can be guaranteed to achieve perfect in-sample prediction accuracy (assuming there are no one-to-many relationships between the X to Y values)?
   a. kNN, if we set k (the number of nearest neighbors to use), to equal the number of training observations.
   b. Decision Trees, if we set the leaf size (the maximum number of observations to be used at a leaf) to equal the number of training observations.
   c. Linear Regression, if we use a second order model to capture any curvature of the data.
   d. Both kNN and Decision Trees if we set k =1 and leaf_size = 1
18. a b c d  You have been tasked with building a learning model that must be trained with incoming sample data in real-time. Which model would be the most suitable and why?
   a. Regression, since we only need to plug the sample data into an equation to train it.
   b. kNN, since there is no training needed, we just add the new data point.
   c. Decision tree, since we can just add a leaf to the tree for each new sample.
   d. None of these models are suitable, it is better to use deep learning for real time tasks.
   e. It depends on how well-defined the problem is, and if we can make a guess as to the shape of the model.

19. a b c d  What are some reasons that could explain a high out-of sample error in your model?
   a. Overfitting and high RMSE on your in-sample data.
   b. Low RMSE and high correlation between your Ypredict and Ytest on your test set.
   c. High correlation and low RMSE on your training set. Overfitting on your out-of sample data.
   d. Overfitting on your training set and low correlation on your test set.

20. a b c d  Which of the following statements is true with respect to parameterized and instance-based models?
   a. Parameterized models train quicker but take up more memory as compared to instance-based models.
   b. Parameterized models take more time to train as well as take up more memory as compared to instance-based models.
   c. Instance-based models make stronger assumptions about the underlying functions mapping the data as compared to parameterized models.
   d. Instance-based models have a higher memory footprint as compared to parameterized models.
21. Given 100 days of data (sampled daily), which is the proper python formula to calculate sharpe ratio? Assume sr is Sharpe Ratio, dr is daily return, ar is annual return, and rfr is the risk-free-rate. np is the numpy library.
   a. sr = np.sqrt(252) * (dr - rfr).mean()/dr.std()
   b. sr = np.sqrt(252) * np.std(dr - rfr)/np.mean(dr -rfr)
   c. sr = np.sqrt(252) * np.mean(dr - rfr)/np.std(dr)
   d. sr = np.sqrt(100) * (dr - rfr).mean()/dr.std()
   e. sr = np.sqrt(252) * (dr - rfr).mean()/ar.std()

22. In order to optimize sharpe ratio, we define an objective function and a function to calculate statistics (set rfr=0):
   ```python
   def statistics(allocs):
       ...
       k = np.sqrt(252)
       ...
       return ___
   
   def min_func_sharpe_ratio(allocs):
       return -statistics(allocs)[3]
   ```

   What will be the possible code in the blank line
   a. return cummulative_return, average_daily_return, sharpe_ratio
   b. return cummulative_return, average_daily_return, std_daily_return,
      k*cummulative_return/std_daily_return
   c. return cummulative_return, average_daily_return, std_daily_return,
      k*average_daily_return/std_daily_return
   d. return cummulative_return, average_daily_return, std_daily_return,
      k*cummulative_return/average_daily_return

23. What is the output of the following code?
   ```python
   import numpy as np
   x = np.array([[1,2,3],[5,6,7]], dtype='int')
   y = np.array(x/2)
   x[1,0] = 0
   print y.sum(axis=1)[-1]
   ```
   a. 2
   b. 4
   c. 8
   d. 9

24. a b c d text ammendment xxx 24 xxx
25. a b c d text ammendment xxx 25 xxx
26. a b c d text ammendment xxx 26 xxx
27. a b c d text ammendment xxx 27 xxx
28. a b c d text ammendment xxx 28 xxx
29. a b c d text ammendment xxx 29 xxx
30. a b c d text ammendment xxx 30 xxx
What is the output of this code:

```python
import pandas as pd
import numpy as np
a = [[22,33,44,55],
     [1,2,3,np.nan],
     [111,222,np.nan,444],
     [11,22,33,44]]

df = pd.DataFrame(a,columns=['COL1','COL2','COL3','COL4'])
df = df.fillna(method='ffill')
df = df/df.ix[0,0:]
print df
```

(a)

<table>
<thead>
<tr>
<th></th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>22.000000</td>
<td>33.000000</td>
<td>44.000000</td>
<td>55.000000</td>
</tr>
<tr>
<td>1</td>
<td>0.045455</td>
<td>0.060606</td>
<td>0.068182</td>
<td>1.000000</td>
</tr>
<tr>
<td>2</td>
<td>5.045455</td>
<td>6.727273</td>
<td>0.068182</td>
<td>8.072727</td>
</tr>
<tr>
<td>3</td>
<td>0.500000</td>
<td>0.666667</td>
<td>0.750000</td>
<td>0.800000</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th></th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.000000</td>
<td>0.060606</td>
<td>0.068182</td>
<td>1.000000</td>
</tr>
<tr>
<td>1</td>
<td>0.045455</td>
<td>1.000000</td>
<td>0.068182</td>
<td>0.800000</td>
</tr>
<tr>
<td>2</td>
<td>5.045455</td>
<td>6.727273</td>
<td>1.000000</td>
<td>8.072727</td>
</tr>
<tr>
<td>3</td>
<td>0.500000</td>
<td>0.666667</td>
<td>0.750000</td>
<td>1.000000</td>
</tr>
</tbody>
</table>

(c)

<table>
<thead>
<tr>
<th></th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td>1</td>
<td>0.045455</td>
<td>0.060606</td>
<td>0.068182</td>
<td>NaN</td>
</tr>
<tr>
<td>2</td>
<td>5.045455</td>
<td>6.727273</td>
<td>NaN</td>
<td>8.072727</td>
</tr>
<tr>
<td>3</td>
<td>0.500000</td>
<td>0.666667</td>
<td>0.750000</td>
<td>0.800000</td>
</tr>
</tbody>
</table>

(d)

<table>
<thead>
<tr>
<th></th>
<th>COL1</th>
<th>COL2</th>
<th>COL3</th>
<th>COL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td>1</td>
<td>0.045455</td>
<td>0.060606</td>
<td>0.068182</td>
<td>1.000000</td>
</tr>
<tr>
<td>2</td>
<td>5.045455</td>
<td>6.727273</td>
<td>0.068182</td>
<td>8.072727</td>
</tr>
<tr>
<td>3</td>
<td>0.500000</td>
<td>0.666667</td>
<td>0.750000</td>
<td>0.800000</td>
</tr>
</tbody>
</table>
What is the output of this python code?

```python
import numpy as np
m = np.matrix([[1,1,0,1],
               [0,0,1,0],
               [1,1,0,1],
               [0,1,1,0]])

print m[1:3,-1] + m[-3:-1,1]
```

Select one answer:

a) [[1 2]]
b) [[1 2 0]]
c) [[0]
    [2]]
d) [[0]
    [2]
    [1]]
How should section A be filled in to complete code that will cause the following output:

```python
import pandas as pd
a = pd.DataFrame([1,2,3], columns=['First'])
print a
b = pd.DataFrame([1,2,3,4], columns=['Second'])
print b
print __A__
```

Output:

```
<table>
<thead>
<tr>
<th></th>
<th>First</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>First</th>
<th>Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>NaN</td>
<td>4</td>
</tr>
</tbody>
</table>
```

Select one answer:
a) a.join(b)
b) a.join(b, how='inner')
c) a.join(b, how='outer')
d) pd.concat([a,b])
What is the output of the following script:

```python
import pandas as pd
df = pd.DataFrame(data={'price':[1,2,3,4,5]},
                   index=pd.date_range('2010-01-01','2010-01-05'))

pd.rolling_mean(df, window=2)
```

Select one answer:

a) 

<table>
<thead>
<tr>
<th>date</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-01-01</td>
<td>1.0</td>
</tr>
<tr>
<td>2010-01-02</td>
<td>1.5</td>
</tr>
<tr>
<td>2010-01-03</td>
<td>3.0</td>
</tr>
<tr>
<td>2010-01-04</td>
<td>4.5</td>
</tr>
<tr>
<td>2010-01-05</td>
<td>6.0</td>
</tr>
</tbody>
</table>

b) 

<table>
<thead>
<tr>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1.5</td>
</tr>
<tr>
<td>2 2.5</td>
</tr>
<tr>
<td>3 3.5</td>
</tr>
<tr>
<td>4 4.5</td>
</tr>
</tbody>
</table>

c) 

ValueError: min_periods must be >= 0

d) 

<table>
<thead>
<tr>
<th>date</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-01-01</td>
<td>NaN</td>
</tr>
<tr>
<td>2010-01-02</td>
<td>1.5</td>
</tr>
<tr>
<td>2010-01-03</td>
<td>2.5</td>
</tr>
<tr>
<td>2010-01-04</td>
<td>3.5</td>
</tr>
<tr>
<td>2010-01-05</td>
<td>4.5</td>
</tr>
</tbody>
</table>
XXX 28 XXX

Fill in the blank for "y" to cause the following output:

```python
import numpy as np
x = np.random.random([2,5])
print x
y = ???
print y
```

Output:
```
[[ 0.05736545  0.66388265  0.90395058  0.9522113   0.92785198]
 [ 0.77764742  0.25293629  0.27912528  0.98815477  0.10810053]]
[[ 0.66388265  0.9522113 ]]
[ 0.25293629  0.98815477]]
```

Select one answer:

a) y = x[:, 1:4:2]
b) y = x[1:4:2, :]
c) y = x[0:1, 1:2:4]
d) y = x[1:2:4, 0:1]
Given the following allocations:
allocs = [0.1,0.2,0.3,0.4]

Which of the following code will reverse the allocations?
a) allocs = allocs[::-1]
b) allocs = allocs * -1
c) allocs = allocs[-1:]
d) allocs = allocs[:-1:

The numpy array below contains closing prices for six securities over a ten day period.

What is the output of this python code?

Code:

```python
import numpy as np
prices = np.array(
    [[ 86.8 ,  81.64,  90.36,  33.95,  74.48,  86.23],
     [ 86.7 ,  81.13,  94.18,  33.82,  74.47,  84.48],
     [ 87.28,  83.38,  92.62,  33.38,  73.26,  85.13],
     [ 84.67,  82.03,  90.62,  32.59,  71.39,  82.75],
     [ 85.01,  81.46,  92.3 ,  31.99,  72.15,  84.46],
     [ 83.19,  79.15,  90.19,  31.69,  70.77,  83.92],
     [ 81.19,  80.09,  88.28,  31.49,  69.83,  80.76],
     [ 81.34,  79.74,  87.34,  31.75,  71.09,  80.88],
     [ 78.78,  77.74,  84.97,  30.65,  68.51,  79.79],
     [ 78.81,  78.6 ,  83.02,  30.67,  69.94,  80.39]])
```

print prices[:, -1] ### what is the output of this???

Select one answer:
a) A list of closing prices of all six securities on the 10th day
b) 80.39 (The closing price of GLD on the 10th day)
c) A list of the closing price of GLD for all ten days
d) It does not actually output any prices; it actually generates an IndexError exception