Midterm 1 – Data and Classifiers

Background Information

For this midterm, we are going to look at a version of the Titanic data set. This data set records some of the details of passengers on the Titanic and whether they survived or not. This is an ongoing competition on Kaggle where they have split the data set into a training set ***train.csv*** (what you train your model on) and a test set (where they evaluate your model.) Feel free to try your hand at it. You can train up a model on Kaggle's versions of notebooks or Jupyter Lab.

Here is a list of the features in the Kaggle dataset
https://www.kaggle.com/c/titanic/data?select=train.csv

1. *PassengerID* – A unique integer identifier
2. *Survival* - Survival (0 = No; 1 = Yes). Not included in test.csv file.
3. *Pclass* - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
4. *Name* - Name
5. *Sex* - Sex
6. *Age* - Age
7. *Sibsp* - Number of Siblings/Spouses Aboard
8. *Parch* - Number of Parents/Children Aboard
9. *Ticket* - Ticket Number
10. *Fare* - Passenger Fare
11. *Cabin* - Cabin
12. *Embarked* - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

The original source of the competition is documented on openML here:
https://www.openml.org/d/40945
With features

1. *survived* - Survival (0 = No; 1 = Yes)
2. *pclass* - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
3. *name* - Name
4. *sex* – Sex (male, female)
5. *age* - Age
6. *sibsp* - Number of Siblings/Spouses Aboard
7. *parch* - Number of Parents/Children Aboard
8. *ticket* - Ticket Number
9. *fare* - Passenger Fare
10. *cabin* - Cabin
11. *embarked* - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
12. *boat* – Number of lifeboat
13. *body* – Number of recovered body

14. *home*.dest – Home location / Destination

Note: The data set on openML has a bunch of numerical features that have been encoded in the CSV file as strings. While you can import the CSV file using pandas, to use it to train a model, you would want to apply type conversions to the data set. I include the csv file, so you can look at it with Excel. If you want to do more with it, drop me an email.

I have downloaded and created links to both of these data sets for your convenience on the Canvas Midterm Assignment.

Question 1 (20 points): A) Make a copy of the mid_starter.ipynb. Fill in the code cells as indicated. Read in the ***train*** dataset. Give a screen shot showing the first 10 instances of the data set. (This is the Kaggle version.)
B) Use hist() to look at the distributions of the data values in the **train** dataset and point out any anomalies.
C) Go to openML (https://www.openml.org/d/40945) and look at the distributions over the full dataset. Make a reasoned argument about why Kaggle kept the features ***age*** and ***cabin***, but discarded features ***boat***, ***body***, and ***home.dest*** for use in the challenge.

Submission 1 of 3:

- (5) Screen shot of first 10 instances in the train data set.
- (5) Discuss distributions of features in train data set.
- (10) Discuss the feature selection choices between Kaggle and openML.

Question 2 (20 points): A) What feature would you like to add to the ***titanic*** data set that could allow you to make a better prediction of who would survive? Give a story for how this feature could affect survival. Can this feature be created out of the existing features in the data set? Explain.
B) Use the mid_starter.ipynb code to plot age against survival. Try some other combinations and look for patterns. Make an argument for the 2 most important features.

Submission 2 of 3:

- (10) Your desired feature and story.
- (5) Screen shot of age vs survival.
- (5) Your two chosen features and why.

Question 3 (20 points): Train a classifier model from the ***train*** data set. Continue working on mid_starter.ipynb.

(Note: we don't need to split the data set, since that has already been done by Kaggle.

a) *Read in the train dataset (already done)*
b) ***Clean Missing Data*** *on age and replace with the mean.*
c) *Create X (select two input features to train from) – Done for you*
d) *Create y from feature survived – Done for you*
e) ***Choose a classifier and create the model.***
f) ***Train the Model*** *on X and y from (c) and (d).*
g) ***Report the performance of the model on the training set.***
h) ***Look at the performance #s and the graphs and discuss the performance of the model.***

Submission 3 of 3:

- (5) Screen shot of cleaning code
- (5) Screen shot of training code.
- (5) Screen shot confusion matrix and metrics results.
- (5) Discussion of results