Neal Sakash

(Partnered with Patrick Hendley)

CSCI 334

February 13, 2017

Predicting Titanic Survivors Using Bayesian Classifiers

To demonstrate the use of Naive Bayesian classifiers in Python, we used the Titanic data set from Kaggle. The primary method used was the Gaussian algorithm for classification. Three parameters were used in the model. These included passenger class, age, sex, and price of fare.Our first step was to take the training data and put it into a format which would be better suited for analysis.

```
In [20]: import pandas as pd
         import numpy as np
         import csv as csv
         from sklearn.naive_bayes import GaussianNB

In [21]: train_df = pd.read_csv('train.csv', header=0)
```

The “train” and “test” files from Kaggle were made into a dataframe. Missing values for age and fare were replaced by median values. Once the cleanup was completed, we used the sklearn Naive Bayes packages to determine the likelihood of people surviving based on the three parameters.

The following resources were used for our testing and submission.

https://www.kaggle.com/ashwinmoorthy/titanic/naive-bayes

http://scikit-learn.org/stable/modules/naive_bayes.html

https://www.kaggle.com/danylchuk/titanic/gaussian-naive-bayes

http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

The first model uses the parameters of passenger class, sex, and fare, and when used against the testing data showed 244 that didn't survive and 174 that did survive.

```
Test set survival:
[[  0 244]
 [  1 174]]
   PassengerId  Survived
0          892         0
1          893         1
2          894         0
3          895         0
4          896         1
```

The second model added age and added one more person to the survival total

```
trainData = pd.DataFrame.as_matrix(train[['Pclass', 'Sex', 'Fare', 'Age']])
trainTarget = pd.DataFrame.as_matrix(train[['Survived']]).ravel()
testData = pd.DataFrame.as_matrix(test[['Pclass', 'Sex', 'Fare','Age']])
```

```
Test set survival:
[[  0 245]
 [  1 173]]
   PassengerId  Survived
0          892         0
1          893         1
2          894         0
3          895         0
4          896         1
```

The third model added the amount of parents and children aboard and when compared to the test set, showed 247 perishing and 171 surviving

The following resources were used for our testing and submission.

https://www.kaggle.com/ashwinmoorthy/titanic/naive-bayes

http://scikit-learn.org/stable/modules/naive_bayes.html

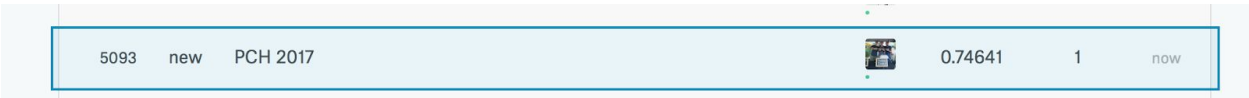
https://www.kaggle.com/danylchuk/titanic/gaussian-naive-bayes

http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

```
In [37]: output = gnb.predict(test_data).astype(int)
         print output

Predicting...
[0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 1 0 0 1 1 1 0 1 1 1 0 1 0 0 0 0 0 1 1 1 0 1
 1 0 0 0 0 0 1 1 0 1 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0 0 1 1 1 1 0 1 1 1 0 1 1
 1 1 0 1 0 1 0 1 0 0 0 0 0 1 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 0 1 0 0 0 0 0 0
 1 1 1 1 1 0 0 1 1 1 1 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0
 0 0 1 0 0 1 0 0 1 1 0 1 1 0 1 0 0 1 1 0 1 1 0 0 0 0 0 0 1 1 1 1 1 1 0 1 1 0 1
 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 1 1 0 1 1 0 1 0 0 0 0 1 0 0 1 1 1 1 0 1 0
 1 0 1 1 0 1 0 0 0 1 0 0 1 0 1 0 1 1 1 1 1 1 0 0 0 1 0 1 1 1 0 1 0 0 0 0 0 0 1
 0 0 0 1 1 0 0 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 0 1 1 1 0 0 1 0 0 0 1 0 1 0 1 0 0
 1 0 0 0 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 1 0 1 1 0 0 1 1 0
 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0 0 1 0 1 0 0 1 0 1 1 0 1 0 0 1 1 0
 0 1 0 0 1 1 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 1 1 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0
 1 1 1 1 1 1 0 1 0 0 0]
```

After submitting our model to Kaggle, we received a score of 74.6% accuracy.



One bit of trouble we ran into was creating a ROC curve. Even after reviewing the scikit-learn examples we ran into repeated errors.

```
NameErrorTraceback (most recent call last)
<ipython-input-8-bd4b6c64ccb9> in <module>()
      1 plt.figure()
      2 lw = 2
----> 3 plt.plot(fpr[2], tpr[2], color='blue',
      4          lw=lw, label='ROC curve (area = %0.2f)' % roc_auc[2])
      5 plt.plot([0, 1], [0, 1], color='navy', lw=lw, linestyle='--')

NameError: name 'fpr' is not defined
```

The following resources were used for our testing and submission.

https://www.kaggle.com/ashwinmoorthy/titanic/naive-bayes

http://scikit-learn.org/stable/modules/naive_bayes.html

https://www.kaggle.com/danylchuk/titanic/gaussian-naive-bayes

http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html