Tashfeen, Ahmad
CSCI 3223: Machine Learning II

Homework 1

Question 1. Read the Naive Bayes section of the Scikit learn's documentation. Name the five Naïve Bayes' models that are described there.

Question 2. In a couple of lines, state the difference between the models implemented by Scikit learn and the one we implemented in homework 3 from Machine Learning I.

Question 3. Keeping the same source for the Titanic data set and data preprocessing under `data.py`, start a Python file (`bayes.py`) with *only* the imports shown in listing 1.

```
1  import numpy as np
2  from matplotlib import pyplot as plt
3  from sklearn.naive_bayes import GaussianNB, MultinomialNB, ComplementNB
4  from sklearn.naive_bayes import BernoulliNB, CategoricalNB
5  from data import X, y, X_, y_
6
7  class NaiveBayes:
8      """
9      Your implementation from Machine Learning I.
10     """
```

Listing 1. Submissions with any other imports will earn zero credit.

The Python class `NaiveBayes` in the code listing 1 is your implementation from Machine Learning I[1]. Using the data split `X, y, X_, y_`, evaluate the accuracy with `X_, y_` after fitting each of the `GaussianNB, MultinomialNB, ComplementNB, BernoulliNB` and `CategoricalNB` on `X, y`.

Note that the categorical model requires a little special attention. It throws an exception if the test set contains data points that were not seen in the training set. Since this is very likely for the fare and the age feature, we may round those to the nearest multiple of 10 like so,

```
X[:,2:4] = 10 * np.round(X[:,2:4] / 10)
X_[:,2:4] = 10 * np.round(X_[:,2:4] / 10)
```

This data adjustment should only be done for the `CategoricalNB` model and it should not affect any other fits.

1) What is the baseline accuracy for how we split the data in training and testing samples?
2) Plot the accuracies of all the models including yours as a bar graph of bars sorted in non-increasing order. The x–ticks should be the names of the models and the y–label should be "Accuracy".

Question 4. Analysing the plot from question 3, which model has the best performance?

Submission Instructions

1) Submit a PDF that answers the questions and contains all the plots that the assignment asks for.
2) Submit your `bayes.py`. If this Python file does not run on the console with `python3 bayes.py` after any necessary path adjustments, the submission will receive no credit.

[1] If you did not take that course here or with the same professor, you'll need to implement a Naïve Bayes classifier as described here.