CSE 572: Lab 6

In this lab, you will practice implementing the probabilistic Naive Bayes classifier.

To execute and make changes to this notebook, click File > Save a copy to save your own version in your Google Drive or Github. Read the step-by-step instructions below carefully. To execute the code, click on each cell below and press the SHIFT-ENTER keys simultaneously or by clicking the Play button.

When you finish executing all code/exercises, save your notebook then download a copy (.ipynb file). Submit the following **three** things:

1. a link to your Colab notebook,
2. the .ipynb file, and
3. a pdf of the executed notebook on Canvas.

To generate a pdf of the notebook, click File > Print > Save as PDF.

Implement Naive Bayes manually

You like to play pickup soccer at the Sun Devil Fitness Center (SDFC). However, you noticed that on some days there are enough people to play a scrimmage but on some days there are not enough people. It's more fun for you to play a scrimmage, and it's a lot of effort for you to pull yourself away from studying Data Mining, so you decide that you only want to go to the SDFC to play soccer when it's likely there will be enough players for a scrimmage. You think players' attendance might be dependent on the weather and proximity to exam weeks, so you collect some observations about these attributes on the days that you've gone to play in the past and whether or not there was a scrimmage on those days. You code that dataset below.

```
import pandas as pd

# Create the dataframe
d = {
     'weather': ['Sunny','Sunny','Overcast','Rainy','Rainy','Rainy','Overcast','Sunny','Sunny','Rainy','Sunny','Overcast','Overcast','Rainy']
     'exam-proximity': ['High','High','High','Medium','Low','Low','Low','Medium','Low','Medium','Medium','Medium','High','Medium'],
     'scrimmage': ['No','No','Yes','Yes','Yes','No','Yes','No','Yes','Yes','Yes','Yes','Yes','No']
    }

df = pd.DataFrame(data=d)

df
```

	weather	exam-proximity	scrimmage
0	Sunny	High	No
1	Sunny	High	No
2	Overcast	High	Yes
3	Rainy	Medium	Yes
4	Rainy	Low	Yes
5	Rainy	Low	No
6	Overcast	Low	Yes
7	Sunny	Medium	No
8	Sunny	Low	Yes
9	Rainy	Medium	Yes
10	Sunny	Medium	Yes
11	Overcast	Medium	Yes
12	Overcast	High	Yes

Today, the weather is Sunny and the proximity to exams is Medium. Implement a Naive Bayes classifier to decide if there is likely to be a scrimmage today and thus you should go to the SDFC.

First, calculate the prior probability of a scrimmage $P(Y = yes)$ and $P(Y = no)$

```
p_y_yes = df[df['scrimmage'] == 'Yes'].shape[0] / df.shape[0]

p_y_yes
```

```
0.6428571428571429
```

```
# YOUR CODE HERE
p_y_no = df[df['scrimmage'] == 'No'].shape[0] / df.shape[0]
p_y_no
```

```
0.35714285714285715
```

Next, we calculate the class-conditional probabilities for the weather and exam-proximity attributes: $P(weather = sunny|no)$, $P(weather = sunny|yes)$, $P(examproximity = medium|no)$, $P(examproximity = medium|yes)$.

Recall that for categorical attributes, $P(X_i = c|y) = \frac{n_c}{n}$ where n_c is number of instances where $X_i = c$ and belongs to class y and n is total number of occurrences of class y.

```
p_sunny_no = df[(df['scrimmage'] == 'No') & (df['weather']=='Sunny')].shape[0] / df[df['scrimmage'] == 'No'].shape[0]

p_sunny_no
```

```
0.6
```

```
# YOUR CODE HERE
p_sunny_yes = df[(df['scrimmage'] == 'Yes') & (df['weather']=='Sunny')].shape[0] / df[df['scrimmage'] == 'Yes'].shape[0]
p_sunny_yes
```

```
0.2222222222222222
```

```
p_medium_no = df[(df['scrimmage'] == 'No') & (df['exam-proximity']=='Medium')].shape[0] / df[df['scrimmage'] == 'No'].shape[0]

p_medium_no
```

```
0.4
```

```
# YOUR CODE HERE
p_medium_yes = df[(df['scrimmage'] == 'Yes') & (df['exam-proximity']=='Medium')].shape[0] / df[df['scrimmage'] == 'Yes'].shape[0]

p_medium_yes
```

```
0.4444444444444444
```

Question 1:

The Naive Bayes assumption is that weather (X_1) and exam proximity (X_2) are conditionally independent given the class value Y. This is true if $P(X_1|X_2, Y) = P(X_1|Y)$, i.e., the value of X_2 has no influence on the value of X_1 given Y. Thus the Naive Bayes assumption is that weather and exam proximity are independent given the variable Y (whether or not there is a scrimmage). Is this a reasonable assumption? Why or why not?

Answer:

YOUR ANSWER HERE

Yes. The weather is not affected by exams and vice versa and thus we can say the weather is independent of the proximity to exams given any value of Y.

Assuming the attributes are conditionally independent given Y allows us to compute $P(X|Y)$ by multiplying the class-conditional probabilities $P(X_1|Y)$ and $P(X_2|Y)$. We compute this below.

```
p_x_yes = p_sunny_yes * p_medium_yes

p_x_yes
```

```
0.09876543209876543
```

```
# YOUR CODE HERE
p_x_no = p_sunny_no * p_medium_no
p_x_no
```

```
0.24
```

Now we are ready to determine our classification. According to Bayes theorem, if $P(X|No)P(No) > P(X|Yes)P(Yes)$, then $P(No|X) > P(Yes|X)$ and we should classify Scrimmage = No and we should not go to the SDFC. If the reverse is true, then we should classify Scrimmage = Yes and we should go to the SDFC.

Below, we calculate $P(X|No)P(No) >$ and $P(X|Yes)P(Yes)$ and check if $P(X|No)P(No) > P(X|Yes)P(Yes)$.

```
p_no_x = p_x_no * p_y_no

p_no_x
```

```
0.08571428571428572
```

```
#YOUR CODE HERE
p_yes_x = p_x_yes * p_y_yes
p_yes_x
```

```
0.06349206349206349
```

```
# Check if P(Y=no|X) is greater than P(Y=yes|X)
p_no_x > p_yes_x
```

```
True
```

Question 2:

Is it likely there will be a scrimmage today, and thus should you go to the SDFC to play soccer? Answer Yes or No.

Answer:

YOUR ANSWER HERE
Scrimmage - No
Go to SDFC - No

Implement Naive Bayes using Scikit-learn

In this section, we will use scikit-learn to implement Gaussian Naive Bayes to predict whether samples in the Wisconsin breast cancer dataset have the class value 'benign' or 'malignant'. Gaussian Naive Bayes estimates the class-conditional probabilities for each attribute by estimating a Gaussian probability density function for each attribute. You can read more about the Gaussian Naive Bayes classifier (and other Naive Bayes classifiers assuming different types of probability distributions) in the sklearn documentation.

```
import pandas as pd
import numpy as np

# Load the Wisconsin breast cancer dataset
data = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data', header=N
data.columns = ['Sample code', 'Clump Thickness', 'Uniformity of Cell Size', 'Uniformity of Cell Shape',
                'Marginal Adhesion', 'Single Epithelial Cell Size', 'Bare Nuclei', 'Bland Chromatin',
                'Normal Nucleoli', 'Mitoses','Class']

data = data.drop(['Sample code'],axis=1)

data = data.replace('?',np.NaN)
data['Bare Nuclei'] = pd.to_numeric(data['Bare Nuclei'])

data
```

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bla Chromat
0	5	1	1	1	2	1.0	
1	5	4	4	5	7	10.0	
2	3	1	1	1	2	2.0	
3	6	8	8	1	3	4.0	
4	4	1	1	3	2	1.0	
...	...	...	...	...	...	...	
694	3	1	1	1	3	2.0	

After loading the dataset, we clean it by removing samples with missing data, duplicates, or outliers using the code from Lab 2.

```
def inds_nans(df):
    inds = df.isna().any(axis=1)
    # print('Found {} rows that had NaN values.'.format(inds.sum()))
    return inds

def inds_dups(df):
    inds = df.duplicated()
    # print('Found {} rows that were duplicates.'.format(inds.sum()))
    return inds

def inds_outliers(df):
    # In this example, we defined outliers as values that are +/- 3 standard deviations
    # from the mean value. To identify such values, we need to compute the Z score for
    # every value by subtracting the feature-wise mean and dividing by the feature-wise
    # standard deviation (also known as standardizing the data).
    df = df[df.columns[:-1]]
    Z = (df-df.mean())/df.std()
    # The below code will give a value of True or False for each row. The row will be
    # True if all of the feature values for that row were within 3 standard deviations of
    # the mean. The row will be False if at leaset one of the feature values for that row
    # was NOT within 3 standard deviations of the mean.
    inlier_inds = ((Z > -3).sum(axis=1)==9) & ((Z <= 3).sum(axis=1)==9)
    # The outliers are the inverse boolean values of the above
    outlier_inds = ~inlier_inds
    # print('Found {} rows that were outliers.'.format(outlier_inds.sum()))
    return outlier_inds


# Select only the rows at index locations that were not nans, duplicates, or outliers
data_clean = data.loc[~((inds_nans(data) | inds_dups(data)) | inds_outliers(data)),:]

data_clean
```

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bla Chromat
0	5	1	1	1	2	1.0	
1	5	4	4	5	7	10.0	
2	3	1	1	1	2	2.0	
3	6	8	8	1	3	4.0	
4	4	1	1	3	2	1.0	
...	...	...	...	...	...	...	
693	3	1	1	1	2	1.0	
694	3	1	1	1	3	2.0	
696	5	10	10	3	7	3.0	
697	4	8	6	4	3	4.0	

Next we normalize the data using the code from Lab 3 so the features will have approximately normal distributions.

```
from sklearn import preprocessing

# Normalize the feature columns
data_clean[data_clean.columns[:-1]] = preprocessing.normalize(data_clean[data_clean.columns[:-1]], norm='l2')
```

```
/usr/local/lib/python3.8/dist-packages/pandas/core/frame.py:3678: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-co
  self[col] = igetitem(value, i)
```

```
data_clean
```

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bl Chroma
0	0.753778	0.150756	0.150756	0.150756	0.301511	0.150756	0.45
1	0.319438	0.255551	0.255551	0.319438	0.447214	0.638877	0.19
2	0.538816	0.179605	0.179605	0.179605	0.359211	0.359211	0.53
3	0.380235	0.506979	0.506979	0.063372	0.190117	0.253490	0.19
4	0.609994	0.152499	0.152499	0.457496	0.304997	0.152499	0.45
...	...	...	...	...	...	...	
693	0.588348	0.196116	0.196116	0.196116	0.392232	0.196116	0.39
694	0.566947	0.188982	0.188982	0.188982	0.566947	0.377964	0.18
696	0.233126	0.466252	0.466252	0.139876	0.326377	0.139876	0.37
697	0.233285	0.466569	0.349927	0.233285	0.174964	0.233285	0.58

Split the data into a training and test set with 70% train and 30% test.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(data_clean[data_clean.columns[:-1]],
                                                    data_clean[data_clean.columns[-1]],
                                                    test_size=0.3,
                                                    random_state=0)
```

Use the GaussianNB object in sklearn to fit a Gaussian Naive Bayes classifier and predict the class labels for the test set based on probabilities estimated from the training set.

```
from sklearn.naive_bayes import GaussianNB

gnb = GaussianNB()


# Fit the model parameters using the training data
gnb = gnb.fit(X_train, y_train)


# Predict the test set classes using the trained model
y_pred = gnb.predict(X_test)
```

Compute the accuracy of this model on the test set.

```
from sklearn.metrics import accuracy_score
# YOUR CODE HERE

print('Test data accuracy: {}'.format((accuracy_score(y_test, y_pred))))
```

```
Test data accuracy: 0.8666666666666667
```

Compute the accuracy of this model on the training set.

```
# YOUR CODE HERE

y_pred_train = gnb.predict(X_train)
print('Train data accuracy {}'.format((accuracy_score(y_train, y_pred_train))))
```

```
Train data accuracy 0.8924731182795699
```