INFSCI 2160: Homework #2

This homework will cover:

- Naive Bayes
- Random forests
- Multinomial logistic regression
- t-SNE
- K-means clustering
- Uniform manifold approximation and projetion

In [1]:
```
import pandas as pd
pd.set_option("display.max_rows", None)
pd.set_option("display.max_columns", None)
```

Part I: Naive Bayes

We will start with the adult_data.txt dataset provided by UCI: https://archive.ics.uci.edu/ml/datasets/Adult (https://archive.ics.uci.edu/ml/datasets/Adult)

A clean .txt file is available on Piazza for download.

Code to download a .txt file is provided below.

In [2]:
```
adult_df = pd.read_csv('adult_data.txt', sep = ",", header = None)
```

In [3]:
```
adult_df.columns = ["Age", "WorkClass", "fnlwgt", "Education", "EducationNum",
                    "MaritalStatus", "Occupation", "Relationship", "Race", "Ge
nder",
                    "CapitalGain", "CapitalLoss", "HoursPerWeek", "NativeCount
ry", "Income"]
```

In [4]:
```
adult_df.head()
```

Out[4]:

	Age	WorkClass	fnlwgt	Education	EducationNum	MaritalStatus	Occupation	Relati
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husba
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husba
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife

In [5]:
```
adult_df.shape
```

Out[5]: (32561, 15)

Question #1: We will predict the working class of the adult dataset using Naive Bayes and multinomial logistic regression. First, predict the working class with Naive Bayes. Createa a confusion matrix and evaluate your model in terms of accuracy. Choose either Gaussian Naive Bayes or Bernoulli Naive Bayes.

- 9 points
  - 2 point for successful Naive Bayes as either Gaussian or Bernoulli
    - 1 point for successful model run
    - 1 point for successful dataprep according to assumptions discussed in class
  - 2 point for identifying accuracy on trian and test sets
  - 1 point for confusion matrix

HINT: Read the documentatino for multinomial NB here: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html (https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html). Different prior options may yield different results!

Part II: Multinomial Logistic Regression

Question #2: Using the same dataset, predict WorkClass using multinomial logistic regression. Evaluate your model in terms of accuracy. Then, create overlapping ROC curves for each predicted class. What do these ROC curves mean to you? Do they give you any insight?

Scikit-learn logistic regression documentation for reference: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

9 points:

- 1 point for successful logistic regression
- 1 point for evaluating accuracy of model on train vs. test data
- 1 point for successful overlapping ROC plots for each class
- 2 points for any explanation

Regularization (such as L1 and L2 regularization) can be (and should be) used for multinomial logistic regression as well.

Part III: Random Forests

Question #3: Using the same dataset, we will now predict if individuals make more or less than $50k ('Income' field is now our target). Make this prediction using a random forest model. How does your model perform? How would you describe the fit? How did you tune your hyperparameters? Name a hyperparameter that you tuned and describe how it affected your model.

HINT: You will be expected to run more than one model and change hyperparameters. Keep all the code for each model you create. Gridsearch is acceptable.

8 points:

```
-1 point for successful random forest model creation
-1 point for evaluating your model by an acceptable metric for binary classificatio
n between train and test
-1 point for explanation of hyperparameter tuning
-1 point for citing a hyperparameter and describing how it affected your model perf
ormance
```

Random forest documentation: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

Part IV: Unsupervised Learning

Question #4: We will analyze MLB baseball pitcher performance over the past 5 years. The data is from fangraphs.com, and is available on Piazza under "pitcher_war.csv".

Train test split is NOT required in unsupervised learning since it is not possible to overfit

We have very high dimensional data with many continuous variables. Using K-means clustering, create representational groups for pitchers. Use any variables that you see fit to do this. Next, find the best value of K according to the sum of squared errors (SSE). Which value of K did you choose? Why? Finally, make a scatter plot relating Strikes on the x-axis to WAR on the y-axis. Color the points by the clustered group.

5 points

```
- 1 point for successful K-means clustering
- 1 point for evaluating a range of K's to find the best fit according to SSE
- 1 point for describing why that value of K is best
- 1 point for successful scatter plot
- 1 point for coloring observations on scatter plot by cluster group
```

Part V: t-SNE and UMAP

Question #5: Using the same data for K-means clustering, perform t-SNE and a supervised UMAP. When plotting, plot the observations by the first two components of t-SNE or UMAP, respectively. Color the points of the scatterplot by WAR. Describe the trends in the scatter plots you may see in t-SNE and UMAP.

Train test split is NOT required in unsupervised learning since it is not possible to overfit

Please tune parameters to see how they affect your plot results

8 points:

```
- 1 point for successful execution of t-SNE
- 1 point for successful execution of supervised UMAP
- 1 point for successfully creating scatterplot of 1st 2 t-SNE components
- 1 point for successfully creating a scatter plot of 1st 2 UMAP components
- 1 point for coloring observations by WAR in t-SNE scatterplot
- 1 point for coloring observations by WAR in UMAP scatterplot
- 1 point of trends you see in t-SNE scatterplot
- 1 point of trends you see in UMAP scatterplot
```