CS4710: Artificial Intelligence
Homework 5: Classification

Introduction:

For this assignment, you will be implementing one of the classification algorithms we saw in class. You will be given a training set of census data to use and you will be predicting the salaries of citizens based on various features in the data. There will be a competition to see who can classify the test data the best.

Background – The Data Set:

The data set we are providing can be found on Collab. There are two files of interest:

census.names*: This file contains a list of the features as well as the different values the features can take. For example:*

```
/* First line contains the output classes. You are predicting whether citizens make >50k or <=50k per year */

>50K  <=50K

/* For each feature (one per line) you are told if that feature is numeric or, if discrete, what values that feature can take */

age         numeric

workclass  Private  Self-emp-not-inc  Self-emp-inc  Federal-gov  Local-gov  State-gov  Without-pay  Never-worked

…

…
```

census.train*: The training data. Contains one training example per row. Contains features separated by one space each. The last value in each row is the output variable.*

```
60 Self-emp-not-inc HS-grad 9 Married-civ-spouse Exec-managerial Husband White Male 0 0 50 United-States >50K

28 Private 9th 4 Never-married Other-service Own-child White Female 0 0 35 El-Salvador <=50K

46 Private HS-grad 9 Married-spouse-absent Craft-repair Not-in-family White Male 0 0 40 Poland <=50K

46 Local-gov Some-college 10 Divorced Exec-managerial Not-in-family White Male 0 0 50 United-States <=50K

…

…
```

census.test*: Lastly, you will NOT be given the file containing the test data (that you should not be training on, only predicting). However, we will be using this file to test your classifier.*

<u>The Code:</u>

We will provide one Java class (Classifier.java). You should extend this class and implement the following methods:

Classifier.java

Classifier(String namesFilepath)	*Constructor. The one parameter is the path to the names file (see above). Initializes your classifier*
void train(String trainingDataFilepath)	*Trains your classifier. The last column of trainingDataFilepath will be the output variable. This array does NOT include feature headers.*
makePredictions(String testDataFilepath)	*Makes predictions on test data. This test data does NOT include the output variable, and so has one less column than training data above.*

<u>Running Your Code:</u>

We will run your code using our own test harness that instantiates your class, invokes train(), and then invokes makePredictions(). makePredictions() should output your predictions (the class only), one prediction per line to standard output. **Your code should not print out anything else during execution.**

<u>Requirements:</u>

You must do, at a minimum, the following:

- Extend Classifier.java and implement the three required methods.
- Compare your classifier's results against a simple classifier that simply guesses the classification. How much better is your approach?
- Do at least one more complicated experiment. This might include:
    - Implement another classifier and compare the results of the two
    - Tweak some aspect of your classifier (e.g., the hypothesis function?) and compare the two versions of your classifier
    - Tweak your training data in some way (e.g., more or less features?) and compare how the classification success changes
- Produce a pdf documenting your design, implementation, and results.

Writeup:

Produce a document that describes, at a minimum, the following aspects of the assignment:

- Describe, in detail, the functionality of your classifier and defend the motivations for your implementation.
- Present quantifiable results of your classification, including how much better it does than a random classifier.
- Present the data, results, and analysis of your more complicated experiment.

You Will Turn In:

…a zip up containing your source code, a pdf of your write-up, and the updated training data file (assuming you altered it for your own purposes). It should be clear which classification class you want us to run for the "competition".