Assignment 1

Submitted By: Fahad Aijaz & Rishab Garg

March 29, 2018

1 Question 1

When Does It Completely Fail: This can be shown easily in the example of 2 class values, that is if each one is given a prior that is in actuality the complement of its true value. By means of an equation, it can be demonstrated as: $posterior \propto (prior)(likelihood)$. If $\mathbb{P}(A)$ is the preferable probability of the prior, where $\mathbb{P}(A) \approx 1$, however it is awarded the probability $\mathbb{P}(A^C)$. Clearly it can be seen in this case the accuracy will be far off. For example, suppose the prior of $\mathbb{P}(unacc) \approx 1$ (in the *cars.csv* dataset), however if we assign $\mathbb{P}(unacc) = 0.01$ (which is possible, since we assign randomly), the total number of data available in this dataset will not let it converge to a better accuracy.

Why Does It Work: Demonstrated by an example with ease, the first attribute in cars is *vhigh*. And the rule we clearly infer is $vhigh \implies unacc$. Hence, this is where $IR(1Rule)$ inference will be eventually made. Even if there are a few wrong inferences, the probability will be skewed quickly enough to avoid any further inaccuracies, as the data is read. Furthermore, note that *unacc* is the most popular class. Even though we were to make the inference $high \implies unacc$ (which isn't true) we will still be able to score accurate results, we will get 324 results correct out of 433 after we have have read all *high*'s after *vhigh*'s. However, what rule follows is a fact: $high \implies unacc \vee acc$. Based on this some other combinations we will be easily able to determine,the correct class value, even though if the priors are not correct.

2 Question 2

2.1

Observation 1: Correlation of Size of Data And Accuracy

Dataset	Accuracy	Size
Mushroom.csv	95.667	8124
Hypothyroid.csv	95.667	3163
Car.csv	87.1528	1728
breast-cancer.csv	70.28	286

2.2 3163→8124≠Doubling Accuracy

Even though we have doubled the size of *Mushroom.csv* as compared to *Hypothyroid.csv*, that doesn't mean the accuracy will keep on increasing. However, going from *car.csv* to $breast-cancer.csv$ doess have a significant effect. Hence, the change is not necessarily directly proportional beyond a certain number, however for small data sizes, the change is evident.

2.3 Observation 2: Correlation of Class Value Range And Accuracy

Dataset	Accuracy	Class Values
Mushroom.csv	95.667	2
Hypothyroid.csv	95.667	2
Car.csv	87.1528	4
breast-cancer.csv	70.28	2

2.4 Class Value Range is High $\implies$ Less Accurate

This is an intuitive fact, that if we have only 2 possible class values we have a higher chance of being right as opposed to having 4 possible values.

3 Question 3

3.1 Cross-Validation (k-fold):

By trying cross-validation with various values of k, we have observed that the accuracy goes down in each of the cases. But this is to be expected, since cross-validation's rasion d'etre is to avoid overfitting etc, hence it gives stricter results. We applied different values of k, 90/10, 80/20, 60/40, but each of these have only minor differences in them, hence the differences are not worth quoting. Each one of these splits were tried 10 times and then the mean accuracy was calculated. Stratification, however may give more interesting results with cross-validation, especially in the case of the *car.csv* dataset since it contains greater number of class values.

Dataset	Mean Accuracy Of Cross-Validation
mushroom.csv	95.4366
hypothyroid.csv	95.33
car.csv	85.0289
breast-cancer.csv	65.862

The most significant drop occurs in $breast-cancer.csv$, which is due to its size and is to be expected. There were some key errors in $breast-cancer.csv$ which mean that some attributes were seen in the test phase for the first time, which may be a cause of the drop in the accuracy.

4 Question 7

4.1 Changes in Prediction: First Iteration x Second Iteration x Iteration of Convergence

Dataset	Iteration 0	Iteration 1	Iteration 2	Iteration of Convergence
mushroom.csv	75.57	78.54	78.532	>10
hypothyroid.csv	76.889	79.481	80.113	10
car.csv	70.0231	70.0231	70.0231	1
breast-cancer.csv	53.4965	54.1958	56.2937	10

Firstly, it should be noted that the swapping ideas presented in the lectures were implemented. Hence, only the best arrangement is chosen and henceforth used as a prior. This has proven to be highly effective, since a non-uniform distribution was used in constructing the priors, there is usually a skew towards one of the class values. This skew causes the subsequent predictions to become better quickly.

4.2 Conjecture Of What Is Behind The Scenes

In *cars.csv* the convergence is instaneous as compared to others. We have discussed earlier (in Question 1) that *cars.csv* has features which gives concrete inference rules regarding choosing class values for the training data. This may not be the case in for example $breast - cancer.csv$, hence we see it takes more iterations to converge. The difference is in the entropy of the data. The data reveals information that gives away easy inference in the case of *cars.csv* but not in the case of others. Furthermore, the aforementioned implies that the variation in attributes and the number of attributes will play an important role. If the data set has greater entropy, convergence will be faster and vice versa.