**WEEK 4&5**

1. (Class 5. Lecture 1) When you evaluate your classification output from the confusion matrix, how would you assess/compare the FP (False Positive) with the FN (False Negative)? Does one more affecting/negative factor than the other? If so, discuss how you came out with the assessment. Defend for yourself if there won’t be any differences.
2. What would you do if your dataset is too large to fit in the memory for developing a Decision Tree model?
3. There are four different test options in Weka: 1) Use training set, 2) Supplied test set, 3) Cross-validation and 4) Percentage split. Discuss when you would apply each method along with its pros and cons. Justify if you think there won’t be any differences.
4. When you develop a classification model like Decision Tree, you see many performance measures in the Weka Classifier Output. Define what TP, FP, and F-Measure are. How would you compare two classification models if you have developed from the same dataset? Are there any relationships between TP/FP and F-Measure?
5. Exercise #12 of Chapter 3 in our textbook.

FEEDBACK:

1: Used a good example. But lecture discussed Cost Matrix (-0.05)

2&3: Good job

5: You have used the textbook 1st ed. (-0.3) + 0.1

1.  
False Positive: In Hypothesis testing False Positive is a Type I error i.e rejecting the null hypothesis when it is true.  
False Negative: In Hypothesis testing Type II error is considered False Negative i.e failing to reject the null hypothesis when it is false.  
This can be explained with a classic example of a famous children book story of Shepherd Boy and the Wolf. In this story while tending to the flock of sheep the Shepherd boy falsly alarms the villagers of wolf attacking the sheep because he was bored and thought it would be fun. This is considered a False Positive. He continued this for several nights until the Wolf really appears at the scene posing threat to the lives of the sheep, in which case when he cried out for help none of the villagers turned up since they were fooled multiple times and they have had enough of his mockery.  
Hence a True Negative is when Wolf appears and the boy calls for help which would turn the event in his favor.  
True Negative is when there is no wolf and the boy goes on with his daily duty of tending to the sheep.  
False Positive is when there is no wolf but the boy calls for help which turned the villagers against him and resulted in everyone loss.  
False Negative is when there is a wolf threatening and the boy doesn't call for help.  
Thus it depends on the problem if False Positive is more affecting or False Negative. Some real life examples that can be considered are detecting Cancer in patients. If an individual is detected with Cancer which in reality is a False positive then he/she will be tested for a more Confirmatory test to consider if the Preliminary Test was indeed true or not. If the Secondary Test comes out to be negative that means the patient is free of Cancer in them and is good to go.  
If the Preliminary Test comes out to be negative and this in reality is a False negative then the doctors would choose to let go the patient since .a) It is a negative test result .b)The hospital isn't sitting idle and needs to tend to emergency/urgent care patients rather than doing confirmatory test .c)Confirmatory test take time to evaluate ,need specialized resources and are usually costly.  
Thus in which case a False Negative person's life is in danger and False Positive turns out to be better than False Negative.  
However in case of falsely detecting somebody as a terrorist could affect .a)the individuals life, mental health .b)delay in finding the real terrorist will affect the lives of the people.  
It is all a matter of perspective on case by case basis.

2.  
If the Datasets is too large for the memory we can:  
a)Increase memory: Using the Weka command line interface we can increase the upper limit(heap size) of the memory by modifying the Java memory allocation options.  
b)Decrease sample size: Using random sampling method we can use a smaller sample and depending upon the memory size we can use a incremental data loading technique to window in the data.  
c)Use Relational Database: They provide a means of storing large database and smaller batch of data can be loaded progressively.  
d)Divide the dataset: Divide the large datasets into n number of partitions and then create decision trees parallely. After all the decision trees are made independently we combine them using meta-learning by either running new samples through them or create a subset of trees based on a classification rule.

3.  
a) Use training set:  
First we train a 'classifier' by interpreting the association between the input variables and the class . In this case the classifier is applied on the same dataset where it was trained from. This can be explained by an example of giving the exact same set of questions to students in an exam that they have practiced upon in practice tests. This would not test their knowledge.  
b)Supplied test set:  
The trained classifier is applied on this set to provide an unbiased output of the model. This set is never used for training purposes and used to provide a final result model.  
c)Cross-validation:  
This method randomizes the dataset to create a certain number of equal sized blocks and then trains the classifier on all of those blocks except 1. These training sets can be further divided in 'k' number of blocks to perform cross validation of the parameters to finalize the best among them. The remainder block from the original partition is then used as a test set to test the classifier. This training-testing process is repeated over N times and the classifier parameter is repeated M times, randomizing the data everytime. This method provides an unbiased result however takes a longer processing time.  
d)Percentage split:  
This method splits the data into X and Y percentages for training and testing . For example 80% for training the classifier and the rest 20% for testing it. However this can provide a biased result.  
Use training set provides a biased result and is good when all the data is available to you. This can be used to understand the problem rather than predicting the data.  
Supplied Training Set is used when the training set is large to train the classifier upon.  
Cross-validation technique is used when the training set small or not present and this provides the most accurate outcome.  
Percentage split uses less time to provide an output.

4.  
TP: Is the correct prediction of true Positive value in the dataset

FP: Is the incorrect prediction of the positive value.  
Positive and Negative values affect the result output of the program and False or True provides a judgement if the prediction was correct or incorrect.

The performance metric we choose is dependent on the false positive/negative.  
Accuracy = (TP+TN)/(TP+FP+FN+TN): measures the closeness of the data to a certain known value. It is the ratio between the correct predictions and sum of all the predictions.  
Precision = TP/(TP+FP): Is the measure of closeness of data between each other. It is the ratio between true positive and all the positives (incorrect + correct prediction).  
Accuracy is hitting the bulls eye. Precision is hitting the bulls eye over and over again.  
Recall = TP/(TP+FN): is the ratio between the correct prediction and the reality.

F-measure: This parameter measures the accuracy of the test by computing the weighted harmonic mean of precision and recall values of the test.  
F1-Score = 2x(Recall x Precision)/ (Recall + Precision) .  
If the classification models are from the same dataset then we can compare the accuracy and precision values of these models. So if the classifier model A has an accuracy of 0.6 and classifier model B has an accuracy of 0.7, then we can clearly tell that model B is more accurate than model A.  
It depends on our goal to solve the problem if we want the model to be more precise or accurate.

5.  
These graphs tell us how the data is distributed. If the data points were in one straight line then it would imply the data is normally distributed. But as seen from the graph Petal length and petal width data are not normally distributed.  
Which means that the the first, third and fourth quartile have enough data to be graphed close enough to a straight line but the second quartile doesn't have enough data points and that is why the graph flattens .It would require data transformation techniques to satisfy the normality assumption.