**Assignment #1**

Part I - Decision Tree

1. The decision of going skiing depends on three features: snow, weather, and season. All variables are binary: **Skiing** (yes or no), **Snow** (fresh or frosted), **Weather** (windy or sunny), **Season** (low or high). The following table shows a data set that contains 10 data points.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Snow** | **Weather** | **Season** | **Skiing** |
| 1 | fresh | sunny | low | yes |
| 2 | fresh | sunny | high | yes |
| 3 | frosted | windy | high | no |
| 4 | fresh | windy | high | no |
| 5 | fresh | windy | low | yes |
| 6 | frosted | windy | low | no |
| 7 | frosted | sunny | high | yes |
| 8 | fresh | sunny | high | yes |
| 9 | fresh | windy | high | yes |
| 10 | frosted | sunny | low | no |

Now what you need to do is to carry out the first split for growing a decision tree model. Please use information gain based on Gini impurity as the criteria for splitting. **Note:** Your calculation will be stopped at the point you figure out which variable, **Snow**, **Weather**, or **Season**, you decide to use as the root node. You must write out all the intermediate steps.

Part II – Evaluation Measures

1. Given the following confusion matrix (obtained by using the default decision threshold of 0.5 for the probability estimates of a classifier on all the test examples), answer the subsequent questions.

|  |  |
| --- | --- |
| Predicted  + - | |
| 4 | 2 |
| 1 | 3 |

+

-

Actual  
  
1) What is the value of accuracy? What are the values of TPR (true positive rate), FPR (false positive rate)? Draw this point on the ROC graph.

2) If the cost of false positive prediction = $1 and cost of false negative prediction = $5, what is the total cost? How to adjust decision threshold (i.e., increase or decrease the decision threshold of classifying into positive class) to lower the total cost in this case?