**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.

Answer:

1. Skiing
   1. Yes = 6; No =4
   2. Gini Impurity = 1-(6/10)^2-(4/10)^2=0.48
2. Snow as root node
   1. Fresh 6/10 : Yes = 5; No = 1
      1. Gini Impurity = 1-(5/6)^2-(1/6)^2=0.2778
   2. Frosted 4/10 : Yes = 1; No = 3
      1. Gini Impurity = 1-(1/4)^2-(3/4)^2=0.3750
   3. Weighted average impurity of Snow = 0.2778\*(6/10)+0.3750\*(4/10)=0.31668
   4. Information gain = 0.48-0.31668=0.16332
3. Weather as root node
   1. Sunny 5/10 : Yes = 4; No = 1
      1. Gini Impurity = 1-(4/5)^2-(1/5)^2=0.32
   2. Windy 5/10 : Yes = 2; No = 3
      1. Gini Impurity = 1-(2/5)^2-(3/5)^2=0.48
   3. Weighted average impurity of Weather = 0.32\*(5/10)+0.48\*(5/10)=0.4
   4. Information gain = 0.48-0.4=0.08
4. Season as root node
   1. Low 4/10 : Yes = 2; No = 2
      1. Gini Impurity = 1-(2/4)^2-(2/4)^2=0.5
   2. High 6/10 : Yes = 4; No = 2
      1. Gini Impurity = 1-(4/6)^2-(2/6)^2=0.4444
   3. Weighted average impurity of Weather = 0.5\*(4/10)+0.4444\*(6/10)=0.46664
   4. Information gain = 0.48-0.46664=0.01336
5. Conclusion: using SNOW as our leaf node will result in a larger information gain.

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.

1. **The value of accuracy = (4+3)/(4+1+2+3) = 0.7**
2. **TRP (true positive rate) = TP/(TP+FN)=4/(4+2)=0.6667**
3. **FPR (false positive rate) = FP/(FP+TN)=1/(1+3)=0.25**

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?

|  |  |  |
| --- | --- | --- |
|  | Predicted (+) | Predicted (-) |
| Actual (+) | $0 | $5 |
| Actual (-) | $1 | $0 |

False positive prediction cost = 1\* $1 = $1

False negative prediction cost = 2 \*$5 = $10

Total cost = $1 + $10 =$11

Conclusion: The cost of a false-negative prediction is much higher than the cost of a false-positive prediction. This means that the cost of a positive prediction is lower than the cost of a negative prediction. Therefore, lowering the threshold (allow more prediction to be positive) can reduce the total cost.