Practice Problems for Exam 1

Son Nguyen

9/25/2019

# Problem 1.

Given outcomes of a predictive model:

|  |  |  |
| --- | --- | --- |
| Passenger ID | Probability of Survived | Truth |
| 1 | 0.55 | Survived |
| 2 | 0.2 | Survived |
| 3 | 0.94 | Survived |
| 4 | 0.63 | Not Survived |
| 5 | 0.9 | Survived |
| 6 | 0.35 | Not Survived |
| 7 | 0.84 | Not Survived |
| 8 | 0.38 | Not Survived |
| 9 | 0.01 | Not Survived |
| 10 | 0.68 | Survived |
| 11 | 0.71 | Survived |
| 12 | 0.45 | Survived |

1. Form the confusion matrix when the cutoff value for positive outcome is c=0.5. Calculate the sensitivity, specificity, precision, and F1-Score when c = 0.5
2. Sketch the ROC curve.
3. Sketch and interpret the Cumulative Lift
4. Sketch the Cumulative % Response

# Solution

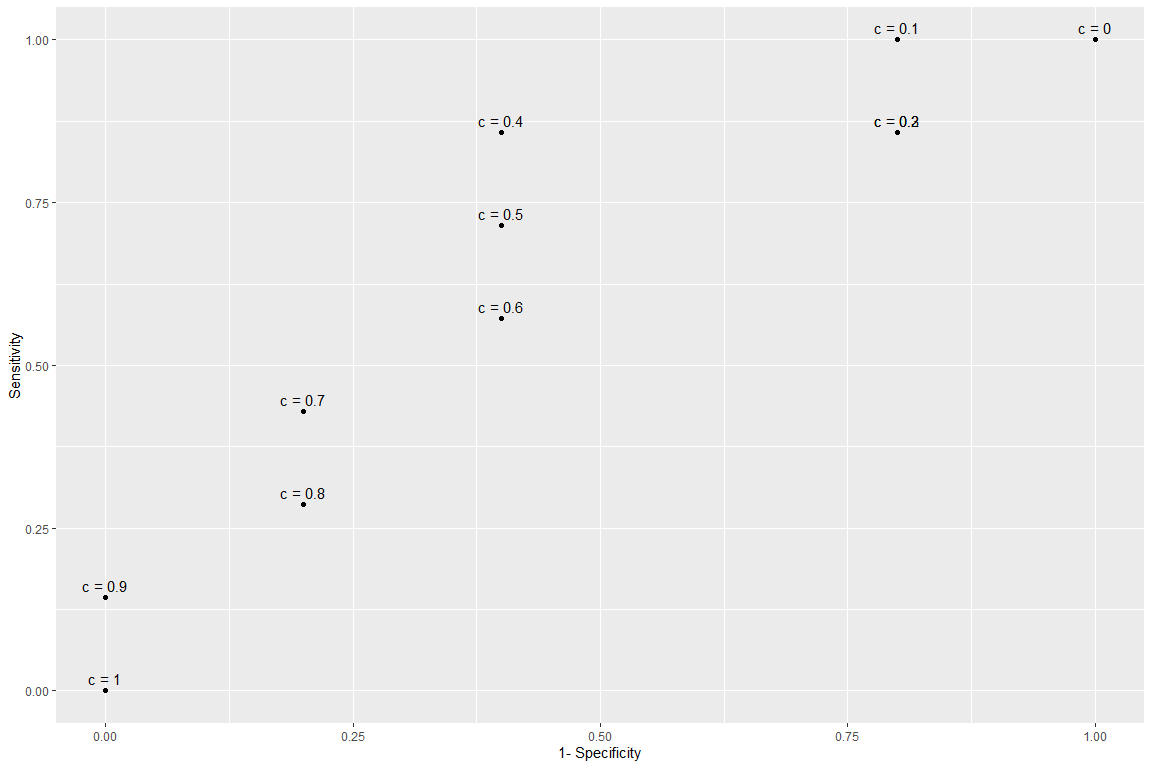
1. Confusion Matrices

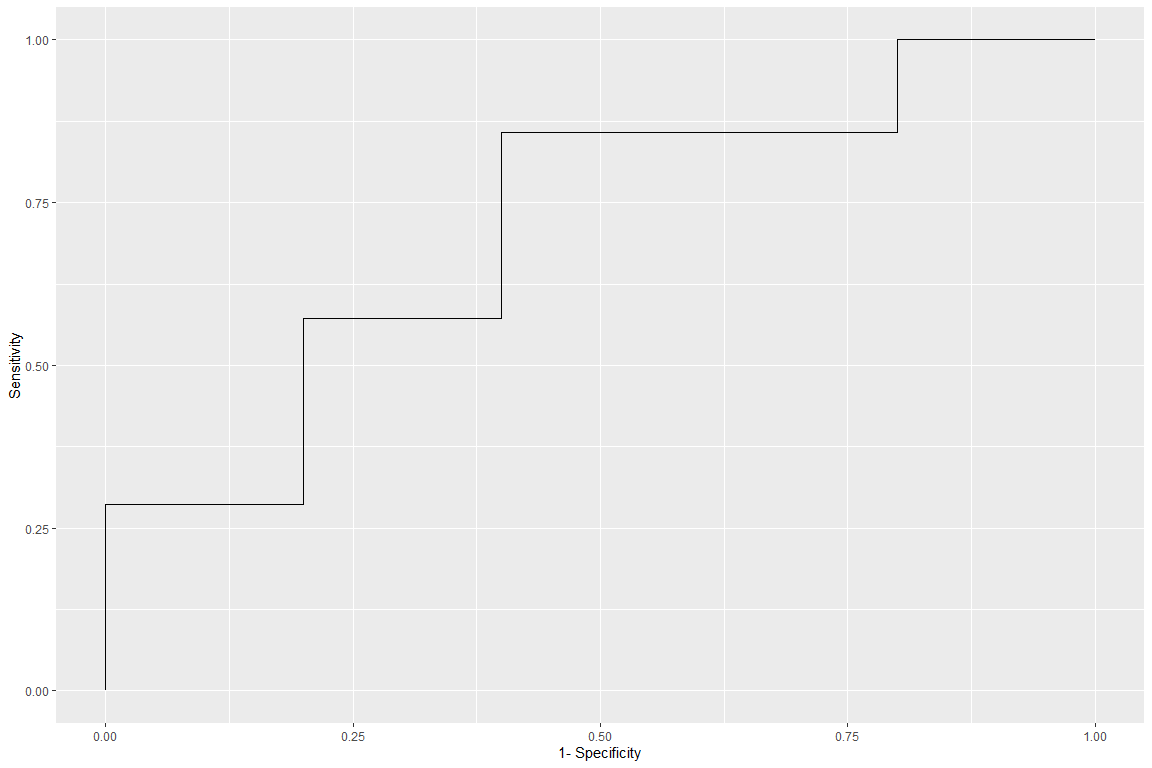
|  |  |  |
| --- | --- | --- |
|  | Predicted Positive | Predicted Negative |
| **Actual Positive** | TP = 5 | FN = 2 |
| **Actual Negative** | FP = 2 | TN = 3 |

1. ROC Curve

|  |  |  |
| --- | --- | --- |
| Order | Predicted Probabilities | True Values |
| 1 | 0.94 | 1 |
| 2 | 0.90 | 1 |
| 3 | 0.84 | 0 |
| 4 | 0.71 | 1 |
| 5 | 0.68 | 1 |
| 6 | 0.63 | 0 |
| 7 | 0.55 | 1 |
| 8 | 0.45 | 1 |
| 9 | 0.38 | 0 |
| 10 | 0.35 | 0 |
| 11 | 0.20 | 1 |
| 12 | 0.01 | 0 |

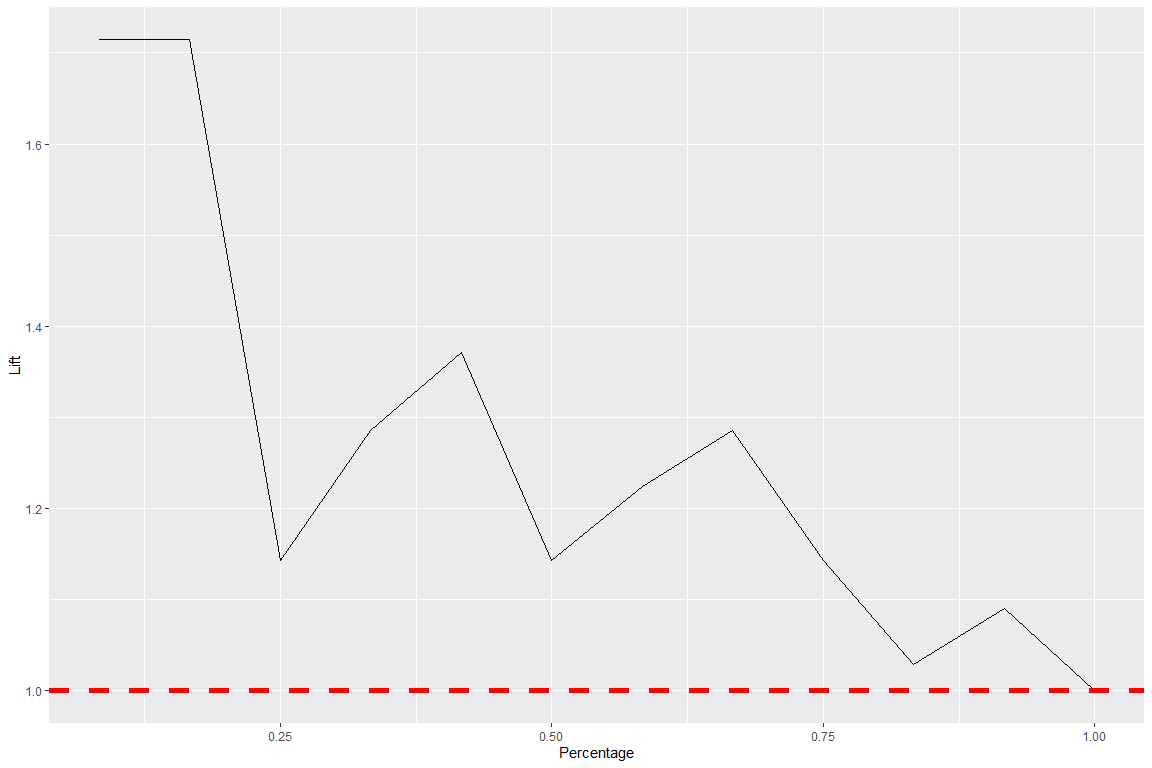
|  |  |  |
| --- | --- | --- |
| Cut-off Values | Sensitivity | Specificity |
| c = 0 | 1.0000000 | 0.0 |
| c = 0.1 | 1.0000000 | 0.2 |
| c = 0.2 | 0.8571429 | 0.2 |
| c = 0.3 | 0.8571429 | 0.2 |
| c = 0.4 | 0.8571429 | 0.6 |
| c = 0.5 | 0.7142857 | 0.6 |
| c = 0.6 | 0.5714286 | 0.6 |
| c = 0.7 | 0.4285714 | 0.8 |
| c = 0.8 | 0.2857143 | 0.8 |
| c = 0.9 | 0.1428571 | 1.0 |
| c = 1 | 0.0000000 | 1.0 |





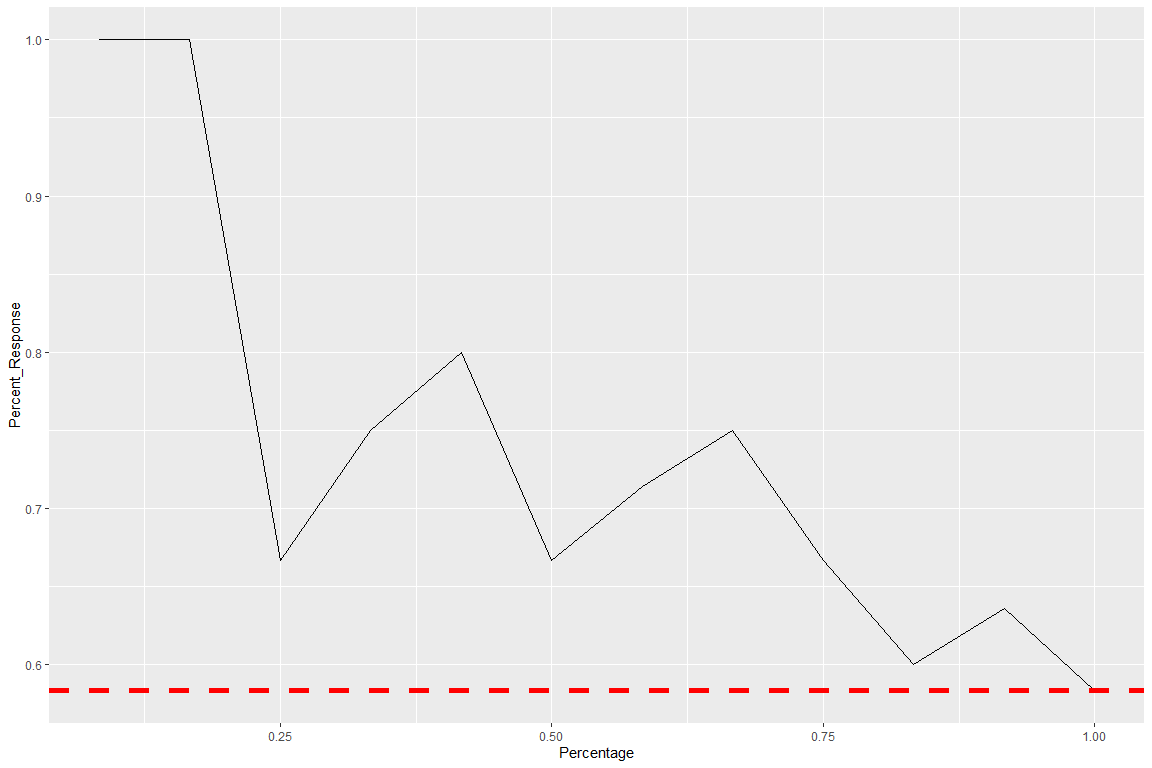
1. Cumulative Lift

|  |  |
| --- | --- |
| Percentage | Lift |
| 0.0833333 | 1.714286 |
| 0.1666667 | 1.714286 |
| 0.2500000 | 1.142857 |
| 0.3333333 | 1.285714 |
| 0.4166667 | 1.371429 |
| 0.5000000 | 1.142857 |
| 0.5833333 | 1.224490 |
| 0.6666667 | 1.285714 |
| 0.7500000 | 1.142857 |
| 0.8333333 | 1.028571 |
| 0.9166667 | 1.090909 |
| 1.0000000 | 1.000000 |



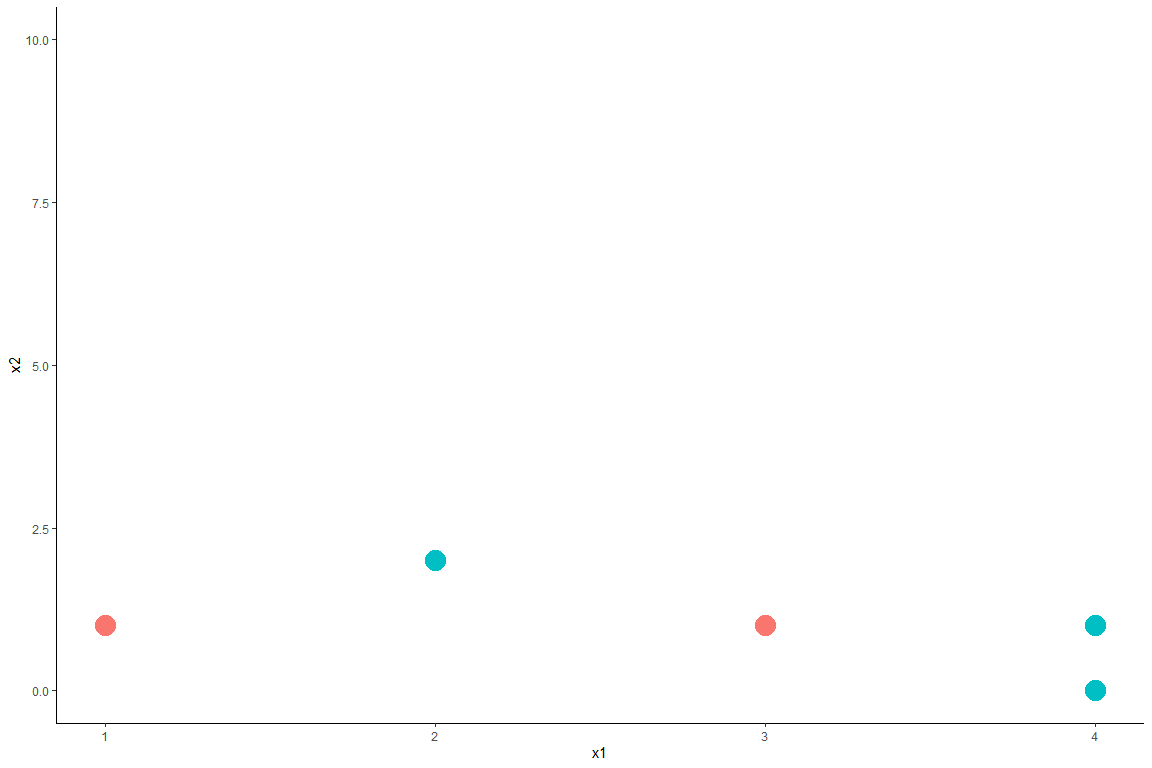
1. Cumulative % Response

|  |  |
| --- | --- |
| Percentage | Percent\_Response |
| 0.0833333 | 1.0000000 |
| 0.1666667 | 1.0000000 |
| 0.2500000 | 0.6666667 |
| 0.3333333 | 0.7500000 |
| 0.4166667 | 0.8000000 |
| 0.5000000 | 0.6666667 |
| 0.5833333 | 0.7142857 |
| 0.6666667 | 0.7500000 |
| 0.7500000 | 0.6666667 |
| 0.8333333 | 0.6000000 |
| 0.9166667 | 0.6363636 |
| 1.0000000 | 0.5833333 |



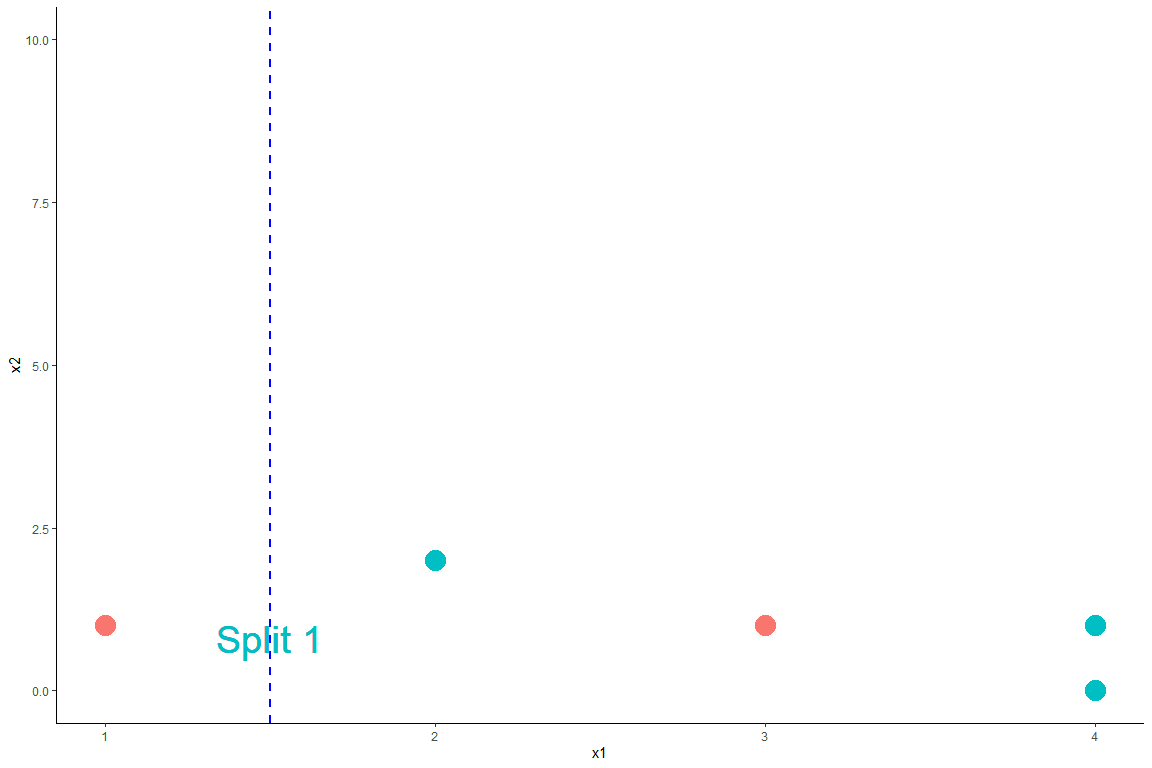
# Problem 2

Given the data points. Calculate the impurity gain of all possible splits by (1) Classification Errors, (2) Gini Index and (3) Entropy. Decide the best split based on each measure.

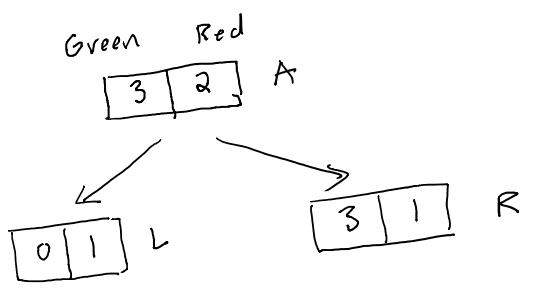


# Solution

### (1) IG By Classification Error:



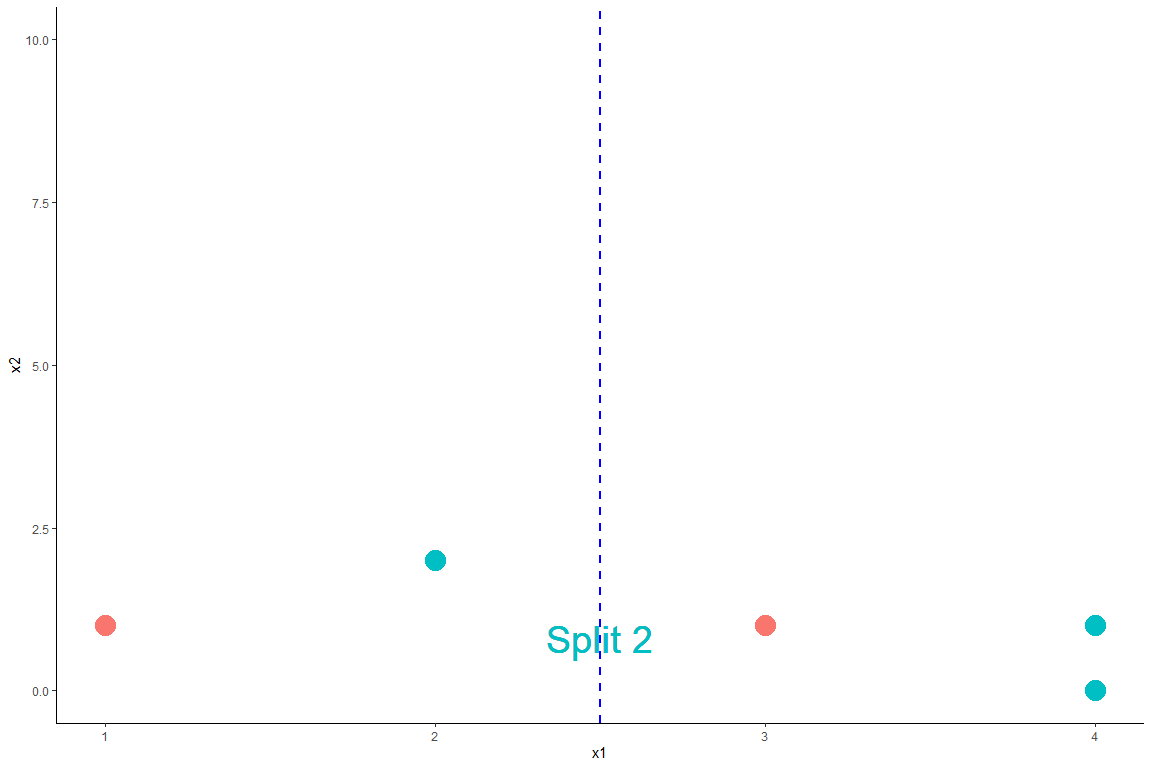
* Let **green** and **red** be class 0 and class 1, respectively.

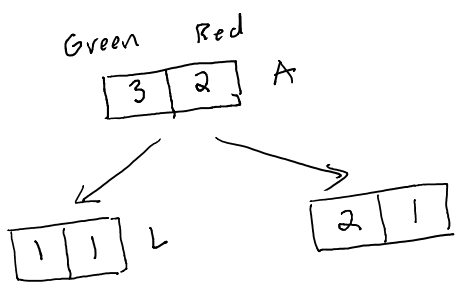


**IG By Classification Error of Split 1**

* For this split:
* Node *parent,* A: . Thus,
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 1 By Classification Error:

**IG By Classification Error of Split 2**

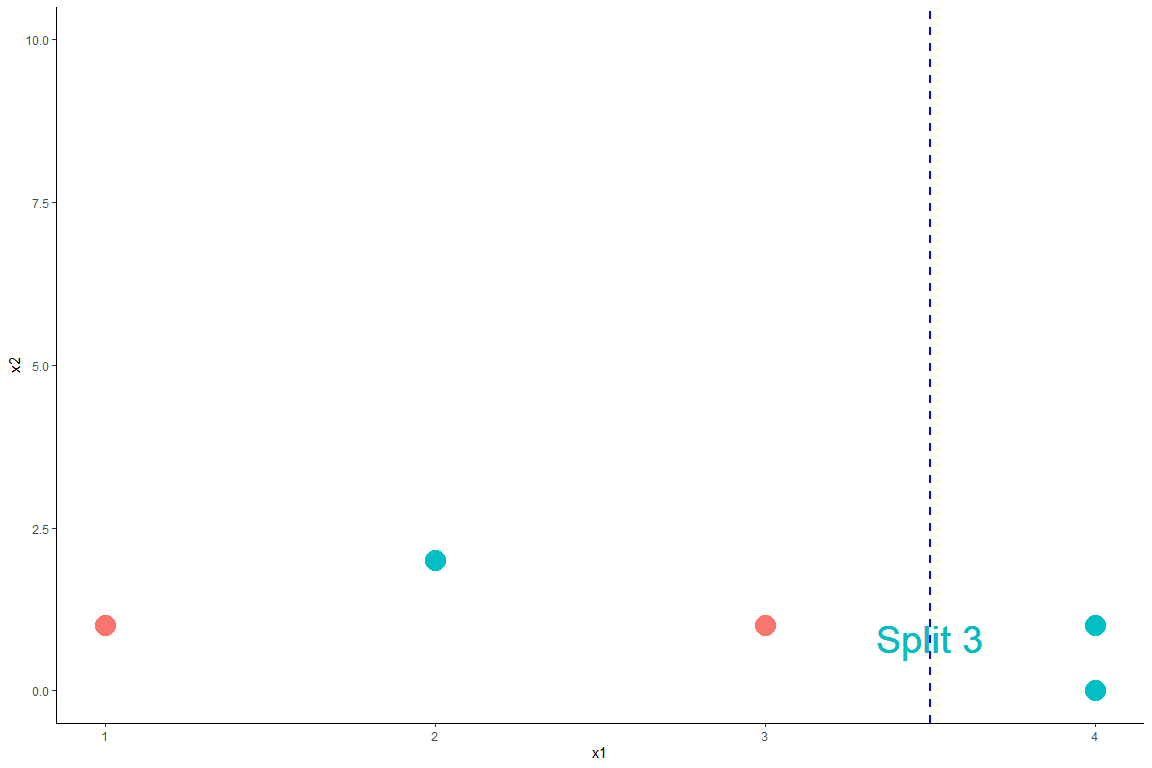


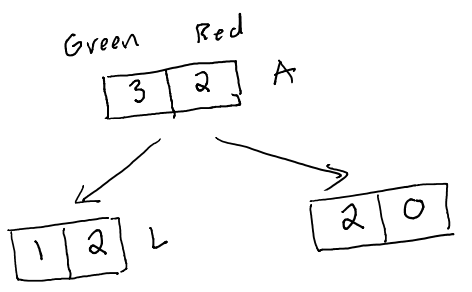


* Node *parent,* A: . Thus,
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 2:

**IG By Classification Error of Split 3**

For Split 3:





* Node *parent,* A: . Thus,
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 3:

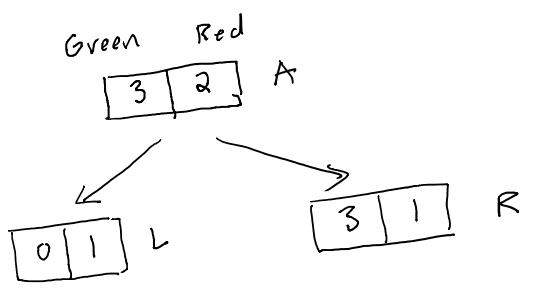
**Comparing IG By Classification Error**

|  |  |
| --- | --- |
|  | IG |
| Split 1 | 0.2 |
| Split 2 | 0 |
| Split 3 | 0.2 |

By classification error, Split 1 and Split 3 are tie as the best because they have the same impurity gain.

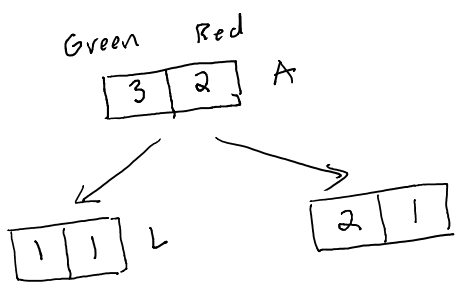
### (2) IG By Gini Index

For Split 1:



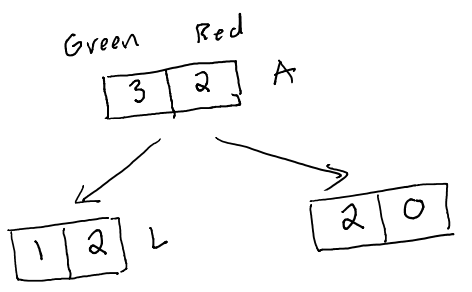
* Node *parent,* A: . Thus,
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 1:

For Split 2:



* Node *parent,* A: . Thus,
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 2:

For Split 3:



* Node *parent,* A:
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 3:

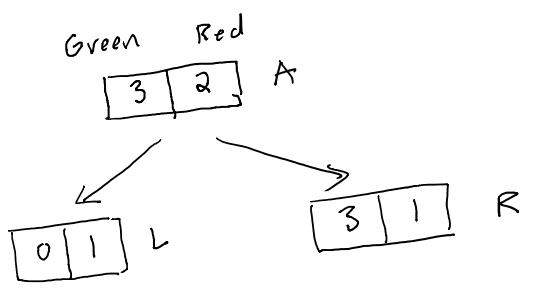
Comparing IG By Gini Index

|  |  |
| --- | --- |
|  | IG |
| Split 1 | 0.18 |
| Split 2 | 0.016 |
| Split 3 | 0.216 |

By Gini Index, Split 3 is the best because it has the greatest impurity gain.

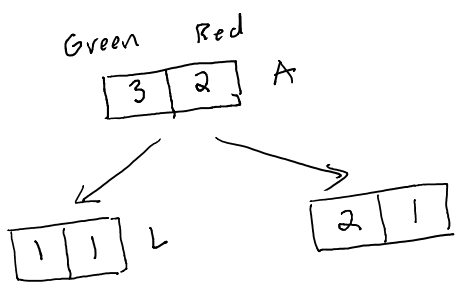
### (3) IG By Entropy

For Split 1:



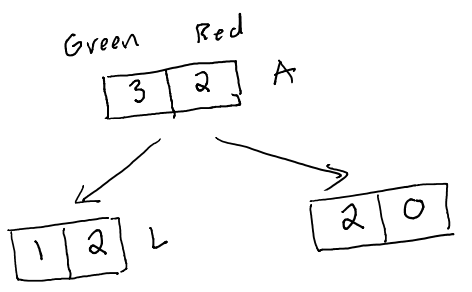
* Node *parent,* A: . Thus,
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 1:

For Split 2:



* Node *parent,* A: . Thus,
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 2:

For Split 3:



* Node *parent,* A:
* Node *child left,* L: . Thus,
* Node *child right,* R: . Thus,
* Impurity Gain of Split 3:

Comparing IG By Entropy

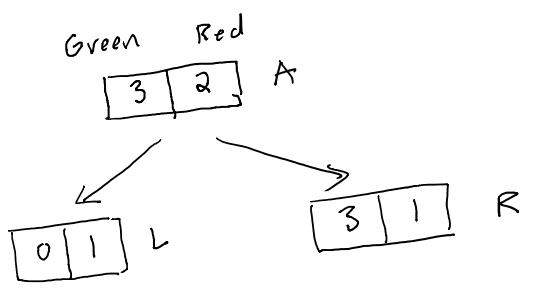
|  |  |
| --- | --- |
|  | IG |
| Split 1 | 0.322 |
| Split 2 | 0.02 |
| Split 3 | 0.42 |

By Gini Index, Split 3 is the best because it has the greatest impurity gain.

# Problem 3.

Compute the Chi-square values of all possible splits in Problem 2 then decide the best split based on this measure.

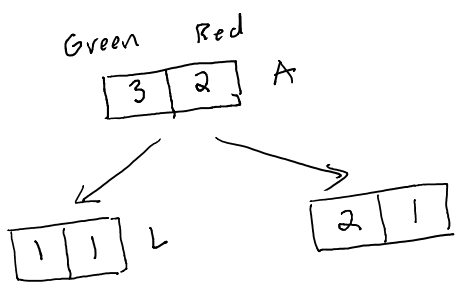
### For Split 1



|  |  |  |  |
| --- | --- | --- | --- |
|  | Greens | Reds |  |
| Left Branch | 0 (Cell 1) | 1 (Cell 2) | 1 |
| Right Branch | 3 (Cell 3) | 1 (Cell 4) | 4 |
|  | 3 | 2 |  |

* (Cell 1): ,
* (Cell 2): ,
* (Cell 3): ,
* (Cell 4): ,
* Plug in, we have:

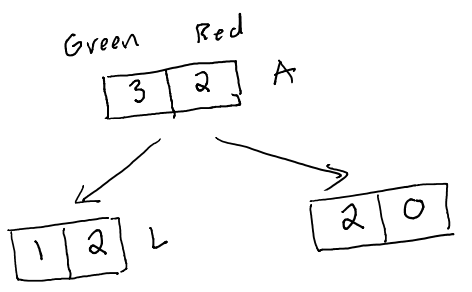
### For Split 2



|  |  |  |  |
| --- | --- | --- | --- |
|  | Greens | Reds |  |
| Left Branch | 1 (Cell 1) | 1 (Cell 2) | 2 |
| Right Branch | 2 (Cell 3) | 1 (Cell 4) | 3 |
|  | 3 | 2 |  |

* (Cell 1): ,
* (Cell 2): ,
* (Cell 3): ,
* (Cell 4): ,
* Plug in, we have:

### For Split 3



|  |  |  |  |
| --- | --- | --- | --- |
|  | Greens | Reds |  |
| Left Branch | 1 (Cell 1) | 2 (Cell 2) | 3 |
| Right Branch | 2 (Cell 3) | 0 (Cell 4) | 2 |
|  | 3 | 2 |  |

* (Cell 1): ,
* (Cell 2): ,
* (Cell 3): ,
* (Cell 4): ,
* Plug in, we have:

### Comparing the three splits

|  |  |
| --- | --- |
|  |  |
| Split 1 | 1.875 |
| Split 2 | 0.139 |
| Split 3 | 2.222 |

Split 3 is the best because it has the greatest !

# Problem 4-7: [Classification Tree Practice Problem](classification_tree2.html)

# Problem 8: [Random Forest Problem](rf2.html)