Miniproject 1

Group 3

Ashley Iannuzzelli

Joshua Ayers

Javier Mendez

Andy Senda

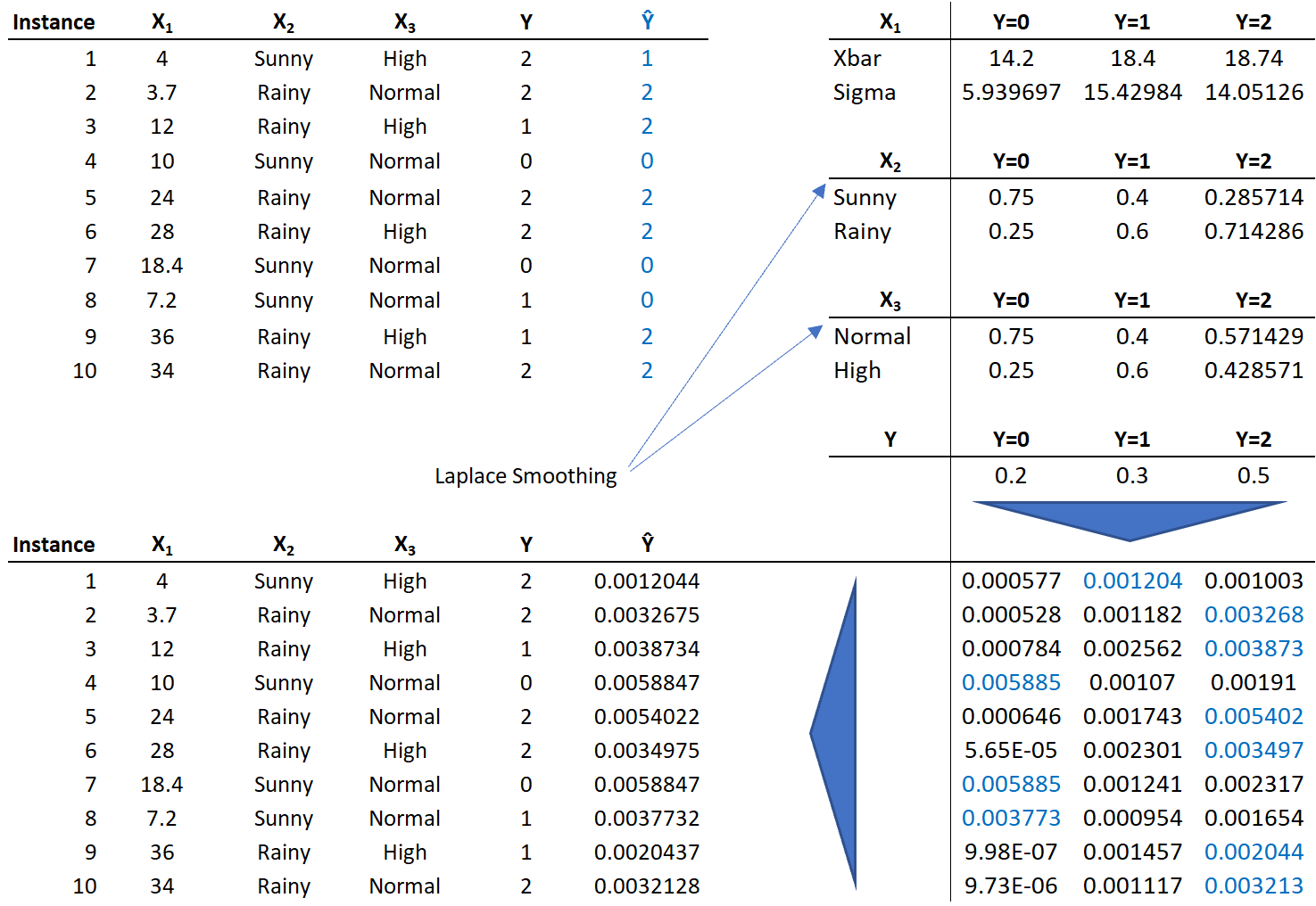
QUESTION 1 - Naïve Bayes Without Software

**a)** **Fill the predicted class on the right most column.**

Answer provided in blue (see right most column).

| **Instance** | **X1** | **X2** | **X3** | **Y** | **Ŷ** |
| --- | --- | --- | --- | --- | --- |
| 1 | 4 | Sunny | High | 2 | 1 |
| 2 | 3.7 | Rainy | Normal | 2 | 2 |
| 3 | 12 | Rainy | High | 1 | 2 |
| 4 | 10 | Sunny | Normal | 0 | 0 |
| 5 | 24 | Rainy | Normal | 2 | 2 |
| 6 | 28 | Rainy | High | 2 | 2 |
| 7 | 18.4 | Sunny | Normal | 0 | 0 |
| 8 | 7.2 | Sunny | Normal | 1 | 0 |
| 9 | 36 | Rainy | High | 1 | 2 |
| 10 | 34 | Rainy | Normal | 2 | 2 |

Supporting Work



Starting at the upper right quadrants, probabilities for each input are calculated based on the observed support for each target. Subsequently, column are multiplied to obtain the numerator portion on Bayes’ formula: 𝑃(𝑋 = 𝑥| 𝑌 = 𝑦)𝑃(𝑌 = 𝑦). Lastly class is selected based on the column with largest probability.

**b)** **Create the confusion matrix for all classes.**

Answer provided in blue.

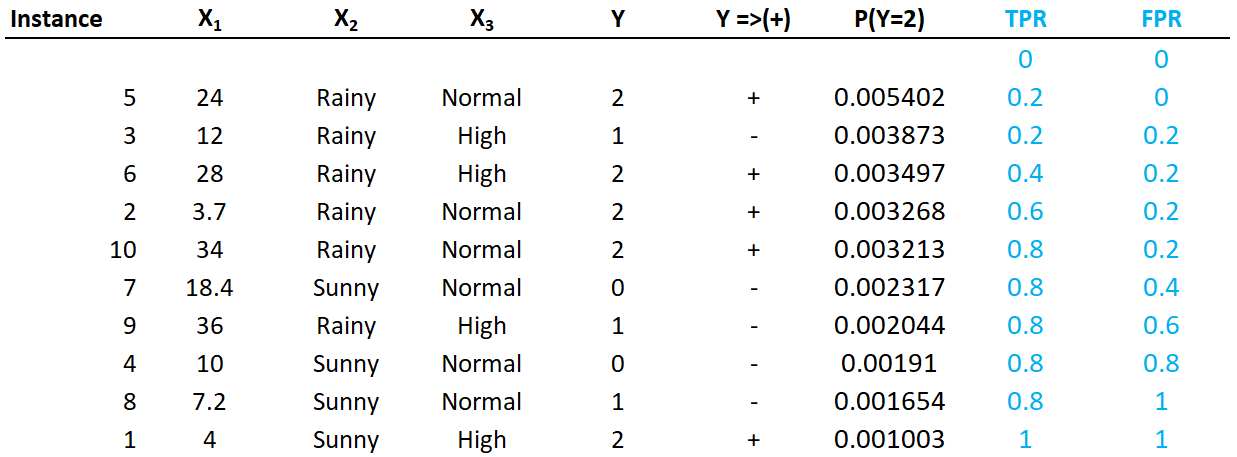
|  | | Predicted | | |
| --- | --- | --- | --- | --- |
| Y=0 | Y=1 | Y=2 |
| Actual | Y=0 | 2 | 0 | 0 |
| Y=1 | 1 | 0 | 2 |
| Y=2 | 0 | 1 | 4 |

**c)** **Assume class 2 is the positive class. Calculate TPR/FPR, plot ROC, and find area under the curve.**

Answer provided in blue and white.

| | **TPR** | **FPR** | | --- | --- | | 0 | 0 | | 0.2 | 0 | | 0.2 | 0.2 | | 0.4 | 0.3 | | 0.6 | 0.2 | | 0.8 | 0.2 | | 0.8 | 0.4 | | 0.8 | 0.6 | | 0.8 | 0.8 | | 0.8 | 1 | | 1.0 | 1 | |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

Supporting Work



Original table was sorted based on P(Y=2) column and actual (Y) was converted to positive vs negative class. P(Y=2) column leveraged the data from bullet “a)” above.

QUESTION 2 - Naïve Bayes Classifier Using Python

**a)** **Evaluate generalization error as follows**

The training data has a generalization error of: \_\_\_\_\_.

The 20% of test data has a generalization error of: \_\_\_\_\_.

The cross validation with 5-folds has a generalization error of:\_\_\_\_\_.

**b)** **Comment on any differences in these estimates of generalization error.**

The data...

**c) Provide the code and a confusion matrix, summary statistics, and ROC curves calculated from the cross-validation only.**

* Class A has a ton of data…
* Class B has an ROC of 1.0…