ECEN758 – Assignment 5

Rachel Hsingtze Yeh

hsingtze1007@tamu.edu

**Exercise 18.5 Q2:**

Using the Naïve Bayes Classifier theorem, one can find out the prior probability and mean for each cluster as:

And using calculated mean to centralize the data for each cluster:

Then for each cluster, the centralized data can calculate the variance for each dimension.

Then for the testing point (T, F, 1.0) , choosing the largest probability to determine the cluster by applying the equation each cluster’s prior probability, mean, and variance:

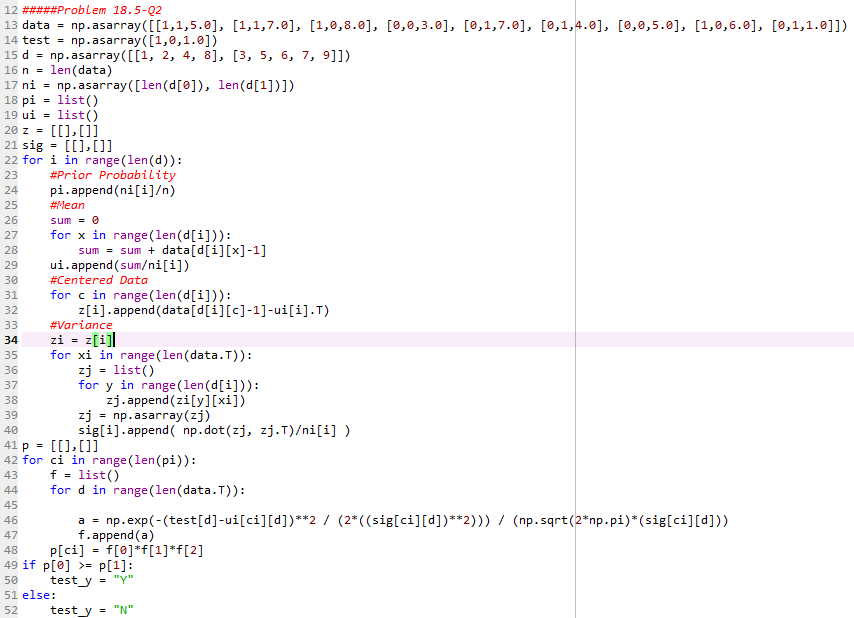
**Results:**

Setting the T = 1.0 and F = 0.0, use the python code the result can be concluded as:

|  |  |  |
| --- | --- | --- |
|  | Class 1 (Y) | Class 2 (N) |
| P(Ci) | 0.4444 | 0.5556 |
| Mean | [0.75, 0.5, 5.25] | [0.2, 0.6, 5.0] |
| Variance | [0.1875, 0.25, 2.1875] | [0.16, 0.24, 6.0] |
| **P (Test Point)** | **0.0007287** | **0.00123** |

From the above table, one can see that the probability of test point in N is larger than probability of test point in Y, then the test point is be placed to **Class N**.

**Code:**



**Exercise 18.5 Q3:**

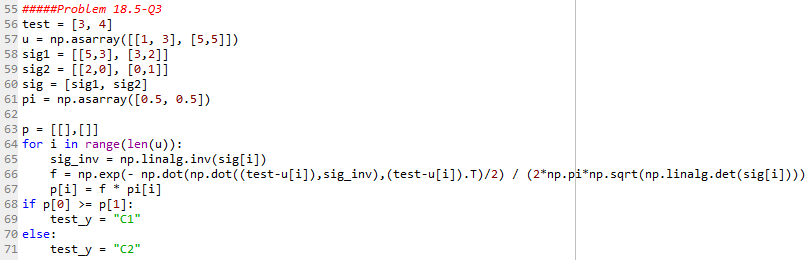
Using the full Bayes Classifier theorem, the probability of predict class is given as:

So the results of the probability with the given mean, covariance, and prior probability is conclude as:

|  |  |  |
| --- | --- | --- |
|  | Class 1 | Class 2 |
| P(Ci) | 0.5 | 0.5 |
| Mean | (1, 3) | (5, 5) |
| Covariance |  |  |
| **P (Test point)** | **0.048266** | **0.012555** |

It’s clearly see that the probability of class 1 is larger than class 2 for the test point, so test point is predicted to the **Class 1**.

**Code:**



**Exercise 19.4 Q2:**

For the Decision Tree Classifier, we have the equation for entropy, split entropy, information gain as:

**Fist Split Point**: find out the maximum gain for each split point is:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  | 0 | 1 | 0.6667 | 0.2516 |
|  |  |  | 0.97095 | 0 | 0.80913 | 0.10917 |
| Car = Sport |  |  | 0.91829 | 0 | 0.45915 | 0.459145 |
| Car = Vintage |  |  | 0 | 0.97095 | 0.80913 | 0.10917 |
| Car = SUV |  |  | 0 | 1 | 0.6667 | 0.2516 |

Compare with all the gain, the maximum gain is Car = Sport, which become the first split point.

* Looking for the side of can conclude that L = {} and H = {x2, x4, x6} where the purity is 100% and no more split can provide.
* Looking for the side of can find that L = {x1, x3} and H = {x5} where the purity is not 100% and need to split again.

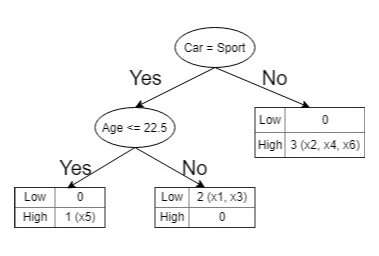
**Second split point**: we only have set and do the same procedure for finding out the maximum gain:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  | 0 | 0 | 0 | 0.9183 |

Since we only have one split point which is also the maximum gain, we can have:

* Looking for can conclude L = {}, H = {x5} where the purity is 100% and no more split.
* Looking for can conclude L = {x1, x3} and H = {} where the purity is 100% and no more split.

**Decision Tree:**



**Test Point: (Age = 17, Car = Vintage)**

Fellow by decision tree, we have so we get the class of **High**.