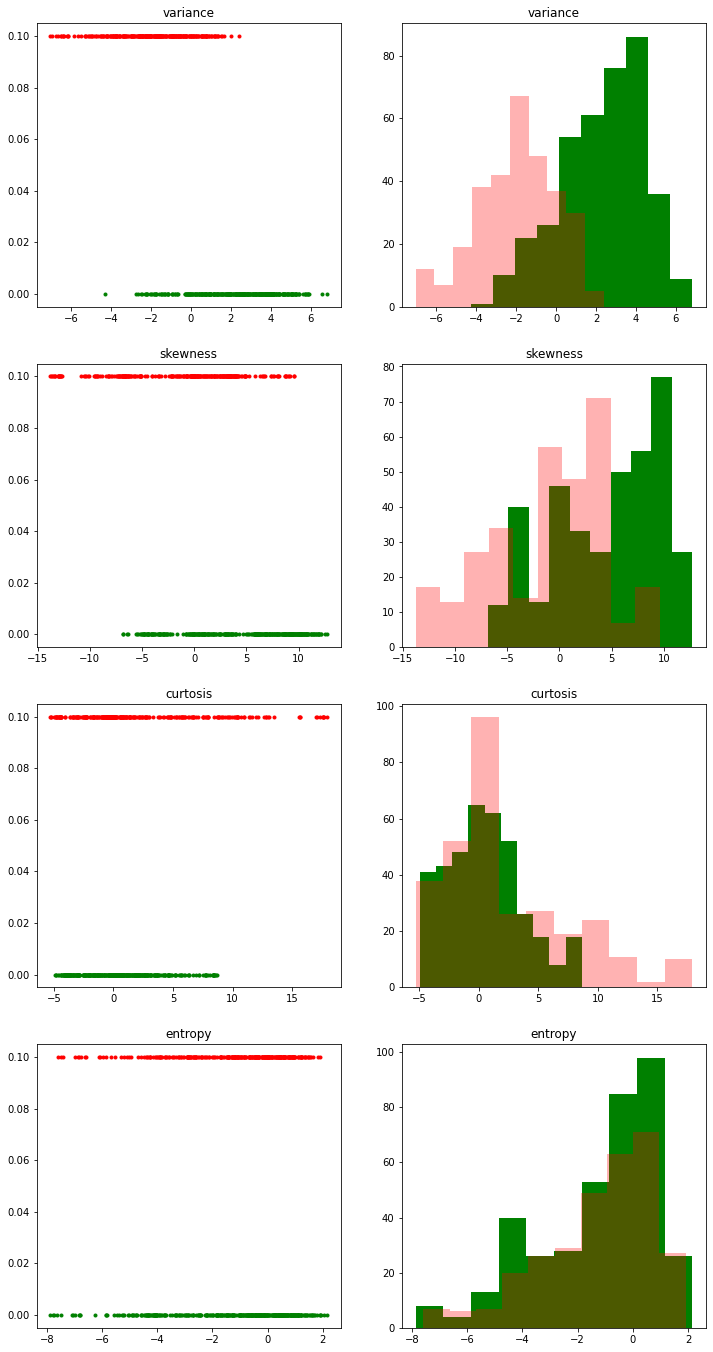
**Question1**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class | Mu(f1) | Sig(f1) | Mu(f2) | Sig(f2) | Mu(f3) | Sig(f3) | Mu(f4) | Sig(f4) |
| 0 | 2.27 | 2.02 | 4.25 | 5.13 | 0.79 | 3.23 | -1.14 | 2.12 |
| 1 | -1.86 | 1.88 | -0.99 | 5.40 | 2.14 | 5.26 | -1.24 | 2.07 |
| all | 0.43 | 2.84 | 1.92 | 5.86 | 1.39 | 4.31 | -1.19 | 2.10 |

**Question2**

Based on these charts, I personally think that it is unnecessary to use all features except ‘variance’. The other data points in other three features are overlapping as we can see from the histogram. Though the question asks me to choose three feature for comparison; however, after trying different combination of features, using only ‘variance’ could bring the highest accuracy.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP** | **FP** | **TN** | **FN** | **TPR** | **TNR** | **Accuracy** |
| **330** | **57** | **244** | **55** | **0.85** | **0.81** | **0.83** |

4.

5. Summary:

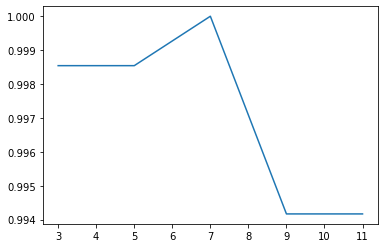
For ‘entropy’, ‘curtosis’, and ‘skewness’, the distribution of two classes of these three features are seriously overlapping, as we can see from the histogram. Therefore, it is not a good feature to classify the data into two classes. Only ‘variance’ provides a data distribution which we can find a relatively obvious difference between two class point.

This case proves the mindset again, simple idea always works better.

6. Yes, my classifier is better than 50%

**Quesiton3**

2. The optimal k\* is k = 7



3.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP** | **FP** | **TN** | **FN** | **TPR** | **TNR** | **Accuracy** |
| **330** | **57** | **244** | **55** | **0.85** | **0.81** | **0.83** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | TPR | TNR | Accuracy |
| 385 | 0 | 301 | 0 | 1 | 1 | 1 |

4. KNN is better than my simple classifier

5.

|  |  |
| --- | --- |
| My classifier | 1 |
| KNN(n=7) | 0 |

**Question4**

1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | All | F1 miss | F2 miss | F3 miss | F4 miss |
| Accuracy | 99% | 95% | 97% | 96% | 98% |

2.

The accuracy doesn’t increase in any of the 4 cases compared with accuracy when all 4 features are used

3.

When ‘**Variance’** missed, it contributed the most to loss of accuracy

4.

‘**Entropy’** contributed the least to loss of accuracy when being removed

**Question5**

2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TP | FP | TN | FN | Accuracy | TPR | TNR |
| 373 | 1 | 300 | 12 | 98% | 96% | 99% |

3.

Logistic regression is better than my simple classifier

4.

KNN(k = 7) is better than logistic regression

5.

Logistic regression predict 0 , it is the same with KNN

**Question6**

1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | All | F1 miss | F2 miss | F3 miss | F4 miss |
| Accuracy | 98% | 82% | 89% | 87% | 98% |

2.

There is no increase when remove any of features

3.

‘Variance’ will contribute the most loss

4.

‘entropy’ will contribute the least loss

5.

The significance of features is the same as I obtained using knn