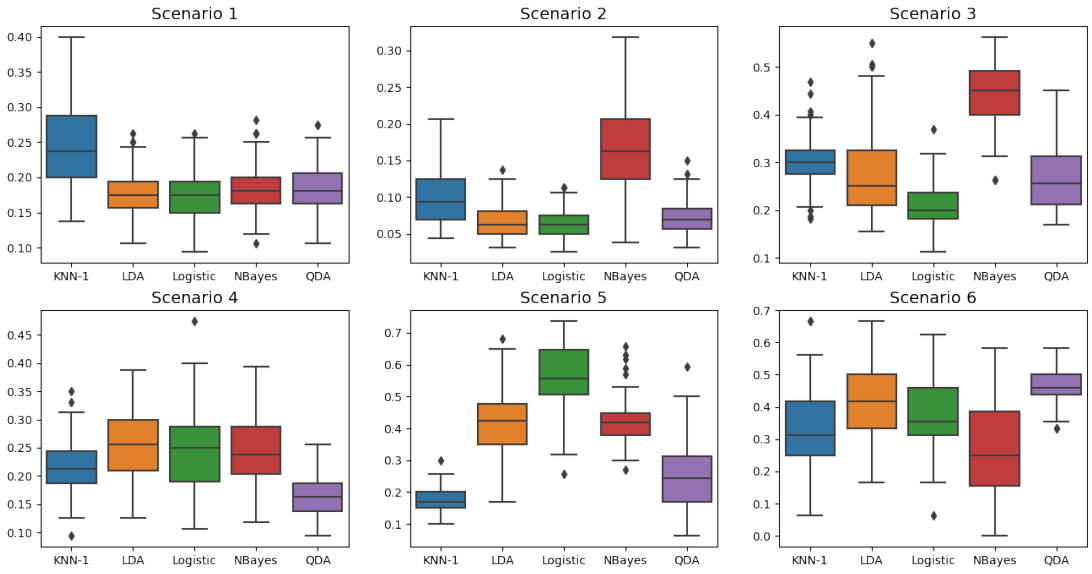
STA4042 24-25T1 Homework 2: Classifications

Inquery: 223040237@link.cuhk.edu.cn

September 26, 2024

Goal

The goal of this assignment is to recreate to some extent the boxplots 4.11 and 4.12 in the textbook (The plots also appeared in lecture 4-5). You will deliver 5 classification methods and apply them to 6 scenarios of data:



Data descriptions (You will need to write your own code to generate the scenario 2, 3 and 5):

- Scenario 1: Given as described in the textbook.
- Scenario 2: Data generated from a normal distribution; same covariance matrix in each class, positive correlation between the predictors (Be careful to choose covariance matrices that are positive definite).
- Scenario 3: Each class generated from a Student t-distribution with identity covariance, still 20 observations per class.
- Scenario 4: Given as described in the textbook.

- Scenario 5: Data generated from a normal distribution with uncorrelated predictors. Then the responses Y were sampled from logistic function applied to a complicated non-linear function of the predictors.
- Scenario 6: Given as described in the textbook.

Questions

Part 1 - With sample data (10+5+10+10+10+10 = 55 points)

1. Write your own function for KNN classification and verify your KNN-1 accuracy. You may take sklearn result for reference.
2. Run your own KKN function for K> 1 and verify your accuracy. You may take sklearn result for reference.
3. Write your own function for Logistic Regression and verify your accuracy. You may take sklearn result for reference.
4. Write your own function for LDA and verify your accuracy. You may take sklearn result for reference.
5. Write your own function for QDA and verify your accuracy. You may take sklearn result for reference.
6. Write your own function for Naive Bayes and verify your accuracy. You may take sklearn result for reference.

Part 2 - With 6 scenarios of data (15+12+18 = 45points)

7. Fill in the data generation code for Scenario 2,3 and 5.
8. Generate 6 boxplots. For each scenario of data, you will generate dataset and run for 100 times, therefore you will have 100 error rates for each method for each scenario.
9. Describe the major error rate characteristics for each scenario and give your justifications (For each scenario, do some methods work not well? Do some methods work better? Why? [Hint]: Think about if the model assumptions and the data distributions match each other.)

We are aware that you might not always get identical results with sklearn.