STAT 428 Presentation: Statistical Tests and Recommendations

Dev Patel

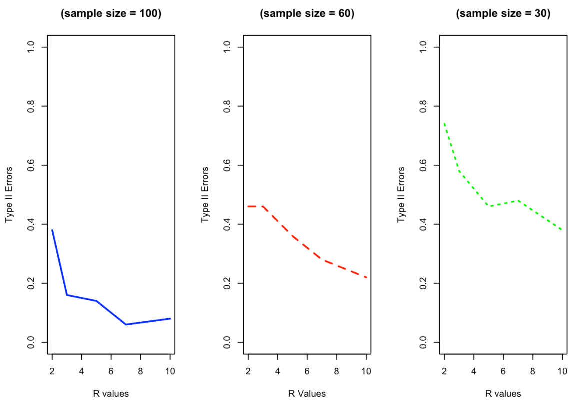
2023-11-27

Introduction

Statistical Tests and Recommendations - Nearest Neighbor Test

- Energy Distance Test
- Hotelling’s T -square test
- Graph Based Two Sample Test
- Parameters = Settings for the tests to run optimally
- Type II error = Likelihood of concluding that the two samples are different when in fact they are not (false negative)

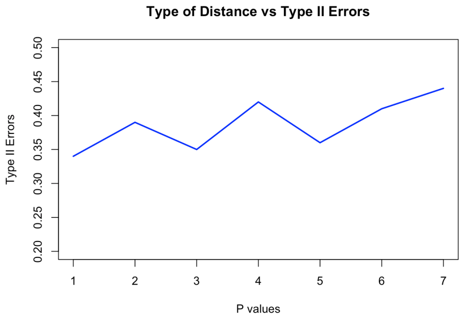
Nearest Neighbor Test Choice of Parameter

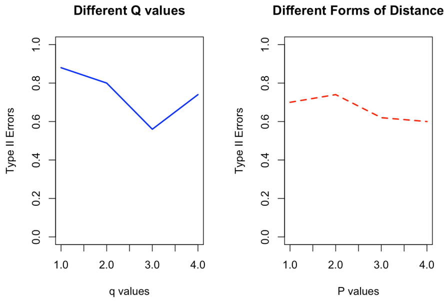


Nearest Neighbor Test Parameter (Nearest Neighbors)

- The choice of desired type II error and sample size affects what nearest neighbor value to select
- Advantages: Very clear which parameter choice is optimal (relative to other tests)

Graph based and Energy distance test Choice of Parameter





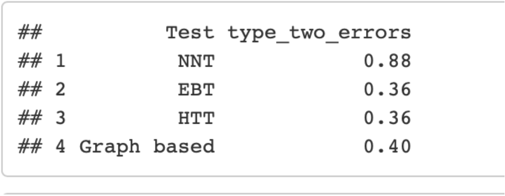
Desired Type II error influences choice of parameters for both tests (left = energy distance, two on right = graph)

Disadvantages: Not clear and obvious which parameter choice is optimal

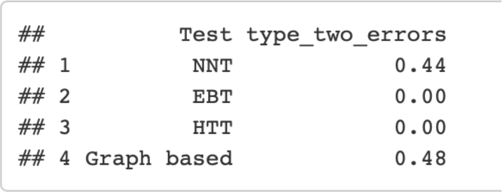
HTT Advantage: No resources need to be spent on selecting parameter for test

4/7

Test Comparisons Results for Dimensionality



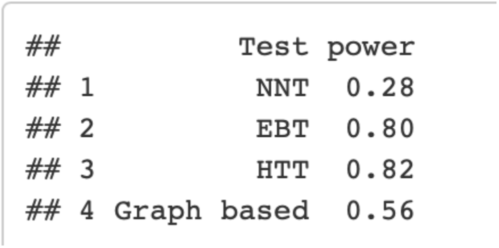
```
##          Test type_two_errors
## 1         NNT            0.88
## 2         EBT            0.36
## 3         HTT            0.36
## 4 Graph based            0.40
```



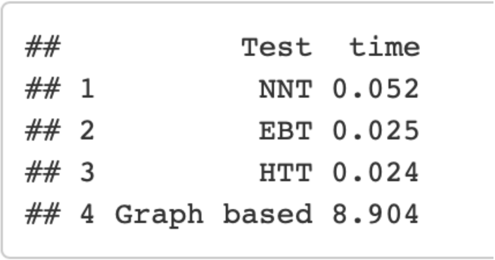
```
##          Test type_two_errors
## 1         NNT            0.44
## 2         EBT            0.00
## 3         HTT            0.00
## 4 Graph based            0.48
```

- Right (dimensions = 2) and left (dimensions = 7)
- Graph based Test is more suitable for low dimensional data
- Other tests can be more suitable for high dimensional data

Test Comparisons Results for Power and efficiency



```
##           Test power
## 1          NNT  0.28
## 2          EBT  0.80
## 3          HTT  0.82
## 4 Graph based  0.56
```



```
##           Test  time
## 1          NNT 0.052
## 2          EBT 0.025
## 3          HTT 0.024
## 4 Graph based 8.904
```

- HTT is most computationally efficient, graph based is most inefficient
- HTT is most powerful test (power = ability of test correctly identifying two samples to be different)

Final Recommendation

Suggested test to utilize would be HTT for the reasons

1. Most powerful test (great ability to correctly identify two samples to be different)

2. Most efficient test (takes less time to run = fewer resources needed)

3. Can work for both low dimensional and high dimensional data well

4. Little/no parameter selections compared to other tests (less time and fewer resources needed to implement)