STAT428 Final Project Presentation

Andy Luo

12/13/2020

Statistical Words

- Power: performance of the tests
- Variance: the spread of data from the average value

Nearest Neighbor Test

- Customized parameter: the number of neighbors when conducting the test (integer)
- When the dimension is fixed, the performance increases when the number of neighbors increases
- To ensure the computational time of this test, we need to avoid it being too large
- If the number of neighbors is too small, we can only obtain less information
- This test is sensitive to the dimension of data depending on the value of the number of neighbors

Energy Distance Test

- Dimensions of data does not affect the performance of the test.
- Customized parameter: the form of distance
- Usually, we consider the Lp distance or the Minkowski distance.
- The values of distance are equivalent in terms of the influence to the performance
- This test is not sensitive to the dimension of data.

Hotelling's T-square Test

- This test is sensitive to the dimension of data.
- Relatively faster runtime
- Performance depends on the significance difference between two treatments(distribution) and dimensions

Graph-based Two-sample Test

- Customized parameter: the threshold Q and distances
- The performance of the test depends on an appropriate Q under different dimensions.
- Same as the energy distance test, the value of distances are equivalent with no influence to the performance
- This test is not sensitive to the dimension of data.

Suggestions of Choosing Tests

- If the dimension is small, all four tests have high powers and the Hotelling's T-square Test will be sensitive to the treat differences.
- If the dimension is large, all four tests except the Hotelling's T-square Test have high powers.
- Hotelling's T-square Test has the fastest running time among the four tests given the same condition. If the difference between treatments(distributions) is obvious and the dimension is small, Hotelling's T-square Test would be a good choice.
- When the dimension is large, energy distance test, graph-based two sample test, and nearest neighbor test would have good performances but their customized parameters need to be appropriate to avoid counterproductive.
- If you care about the running time solely, Hotelling's T-square Test would be the best and energy distance test would be the worst.