Self-assessment #4

Sicheng Wang

Table of contents

1 Take-away **1**

2 Challenges **2**

3 Resources **2**

4 Explore **2**

1 Take-away

(1) Exploratory analysis and predictive analysis test data in multiple ways comparing with inferential data analysis.

(2) PCA is a linear transformation to reduce dimensionality where t-SNE is an non-linear transformation to visualize high-dimensional data.

(3) Classification task is label something, while regression task is measure something.

(4) In machine learning, variables are called features.

(5) The proportion for training and testing is 75:25 or 80:20.

(6) Training is equal to recipe (features) + specification (algorithm)

(7) `strata` is used for sampling.

(8) `kableExtra` create a formatted table, `kable()` create a simple table.

(9) Null hypothesis assume there is no difference, while alternative hypothesis assume there is effect exists.

2 Challenges

The challenges is mainly because IDA, PDA and EDA relate to statistical or mathematical concepts, which needs to be combined with model selection, algorithm selection and results interpret.

3 Resources

Because the math involved, it needs me to search for a lot of information like PCA, it discusses how to decentralize the data. Like bootstrapping, resampling the data with replacement. Like cross-validation, resampling without replacement. Other content like `set.seed`, used for reproduce the results.

4 Explore

The thing I need to learn more is some content related to mathematics which could help me understand the algorithm and independently choose model. Below is just an example about K-means in clustering analysis. It mentions the number of K determines the number of cluster. Normalized the data before measure data relatedness.