Data Analytics for Personalized Genomics and Precision Medicine Data & Python

Lecturer: Yu LI (李煜) from Department of Computer Science and Engineering (CSE)

Lecture 9 – Clustering and Classification Performance Evaluation
Monday, October 3rd, 2022

Outline of the Lecture:

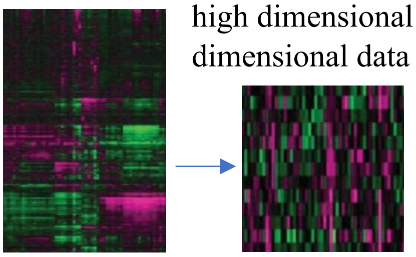
- Why feature selection & dimension reduction
- Feature selection

1. Why feature selection & dimension reduction

- Problem: Not every data we store are useful. The original data is huge, noisy, irrelevant, duplicated.

  Ex) Human has roughly 30.6M cells and it takes up about 2672GB≈2.7TB if we store it in Excel.

- Purpose: Feature selections and Dimension reduction are tools that help to change high dimensional microarray gene expression data into low dimensional data and subset of genes.



- Advantage:

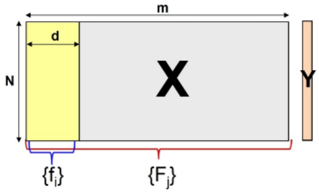
  (1) By reducing the size of the data, storage efficiency can be increased.

  (2) By removing irrelevant data and noise, prediction performance can increase and can understand the results clearly.

2. Feature selection

- There are two steps to reduce the dimensionality: Feature selection and Feature extraction. (In lecture 10, only feature selection was discussed.)
- Feature selection: It is a procedure of selecting the most relevant features to build a better, efficient data. The best subset of features should be selected among all features. There are some methods in feature selecting: Feature ranking, Feature subset selections Filter and Wrapper.

Procedure: Among the features select the most relevant features. Reduce the number of features from m to d.



(When the label Y is used, it is supervised and without it, it is unsupervised.)

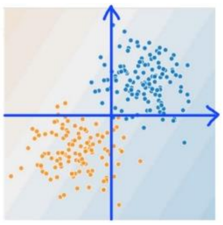
- Feature ranking: To select the best features, we need to discover the most relevant feature. In order to check which one is the best feature we need to look into the correlation between the features and class.
  - Mutual information I(i) / Fisher score F: The higher the I(i), F, the feature would be more related to the class.
  - Issues of individual features ranking:

    - Relevance doesn't imply usefulness and usefulness does not imply relevance. Only relying on relevance might lead to redundant subset.

    Ex) Subset of Height & Major might have higher relevance while it might not be useful compared to subset of Height & Weight.

    - Some features can be useful when it combines with another feature.

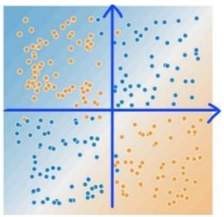
Ex)



1 feature to distinguish yellow and blue.

$Y > 0$, Blue // $Y < 0$, Yellow

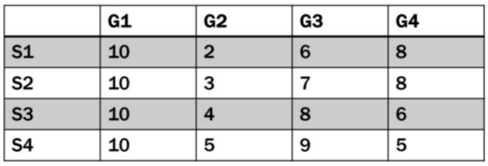
$X > 0$, Blue // $X < 0$, Yellow



At least 2 features needed to correctly distinguish

$X \times Y > 0$, Blue // $X \times Y < 0$, Yellow

- Filter: It is a method of filtering out the features which is very similar to data processing but in the feature level.
  - Features with higher variance is more useful
  - Features with smaller correlation is more useful



	G1	G2	G3	G4
S1	10	2	6	8
S2	10	3	7	8
S3	10	4	8	6
S4	10	5	9	5

Ex) In this case, the feature G1 has 0 variance meaning that it is not a useful feature and can be eliminated. In addition, G2 and G3 shows high correlation so one of them can be deleted either. Through filter method, data can be preserved and reduced to more efficient level.

- Wrapper: This method uses classification performance for selection.

  (1) Find the first best feature using the cross-fold validation.

  (2) Add the second feature using cross-fold validation together with the first feature.

  (3) Repeat the steps until newly added feature does not improve the performance.