BMEG3105 Data Analytic for Personalized Genomics and Precision Medicine

Scribing for Lecture 11: Dimension reduction and Visualization – October 16

Lecturer: Yu Li

Scriber: HONG Yuandai Date: October 23, 2024

Warning: these notes may contain factual errors

Outline of lecture 11

1.The necessary of features selection and dimension reduction

2.Features selection

- **2.1 Features Ranking**
- **2.2 Feature subset selection: Filter and Wrapper**

3.Dimension Reduction-Principal Component Analysis

Content

1.The necessity of features selection and dimension reduction

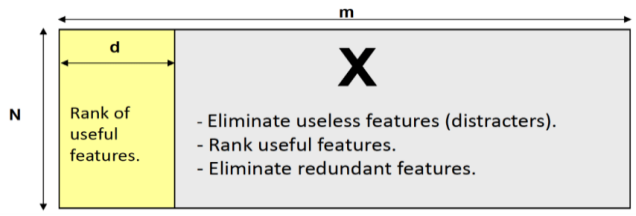
In reality, the big data we encounter can be really huge and include many redundant data. These data can be noise, unrelated or duplicated and occupy a large amount of memory space and lead to an increase in processing time. To eliminate such redundant features and increase the efficiency, feature selection and dimension reduction are necessary.

Benefits of feature selection and dimension reduction:

- Data Compression
- Improve prediction performance
- Understand the prediction results
- Facilitate data visualization

2.Features selection

2.1 Features Ranking



Features selection means selecting or extracting the most useful features to build more efficient and accurate learning machines. Since only useful data are kept for processing, the dimensions(features) reduced.

To rank the usefulness, we may check the relevance between the features to the targeted label by their correlation, mutual information or Fisher score.

Noted that the relevance doesn't imply usefulness and usefulness doesn't imply relevance. For example, gender of students is our targeted label, we have three features Height, Weight and Major of the students, the correlation of

Height & Gender = 0.8

Weight & gender = 0.7

Major & Gender = 0.6,

although Height and Weight are the two most relevant features to gender, but the usefulness of Height & Weight < Height &Major in the classification. We are selecting the k best features instead of best k features, we are choosing the *complementary* features. Also a useless feature by itself can be useful when it paired with others.

2.2 Feature subset selection: Filter and Wrapper

Filter

A feature selection method based on the variance thresholds and the information gain. In filter, classification performance is not involved in the selection loop. In filter, the features with higher variance contain more information, and the ideal information gain of the features should be different and so the features are complementary and avoid duplicated features.

Example :

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

In the above table, feature G1's variance is 0, so it is discarded. G2 and G3 are highly correlated, so we will only choose one of them. The variance of G4 is high and its information gain is different from others, so it is a ideal feature.

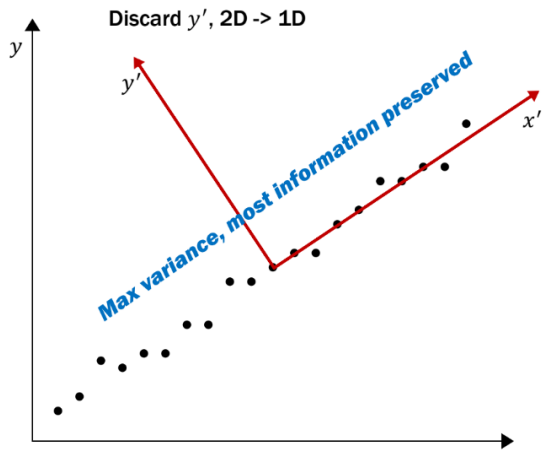
Wrapper

A feature selection method guided by the classification performance. Wrapper is computational expensive since it need to evaluate the classfication performance to guide the selection. There are two kinds of wrapper: Recursive Feature Elimination and **Sequential Feature Selection**.

The sequential feature selection starts with finding the first best feature using cross-fold validation, and then add the second feature and so on, until the new feature doesn't improve the performance. The result of the sequenctial feature selection may neglect some useful combinatuion since it confused on the features with good classification performance while those features with poor classification performance can be useful too when paired with others.

3.Dimension Reduction-Principal Component Analysis

During the dimension reduction, we care about the variance and distance. If a multi-dimensional scatter points show high correlation, we can try to reduce the dimension through principal component analysis(PCA). These highly correlated data has a certain dimension that contains the most variance(in other words information), we project the data onto the dimension and preserved the most variance. The residual variance orthogonal to this dimension are captured by other dimensions and be discarded, and so the dimension reduced.



PCA procedures:

- Construct an n by d data matrix X and normalize the data to get X'
- Calculate the covariance matrix of X'

$$\Sigma = \frac{1}{n-1} X'^T X', \Sigma: \text{a } d \text{ by } d \text{ matrix}$$

- Find the eigenvectors and eigenvalues of the covariance matrix
- Combine the eigenvectors to get the principal components
- Project the data to the principal component by performing $\hat{X} = X'P$