BMEG 3105 Fall 2024

Data analytics for personalized genomics and precision medicine

Lecturer: Yu LI (李煜) from CSE

Scriber: Lai Ching Yan

Student ID: 1155172990

Wednesday 16 October, 2024

Lecture 11: Dim Reduction

Agenda

- Why do we need feature selection and dim reduction
- Feature selection
- Dim reduction

Why do we need feature selection and dim reduction

1. Bio-data can be huge

   Gene expression profile: 25000 genes

   If we have a single-cell RNA sequence with 10000 cells

   -> We need around 1.2TB to store the huge data

   -> which is the data storage of a personal computer

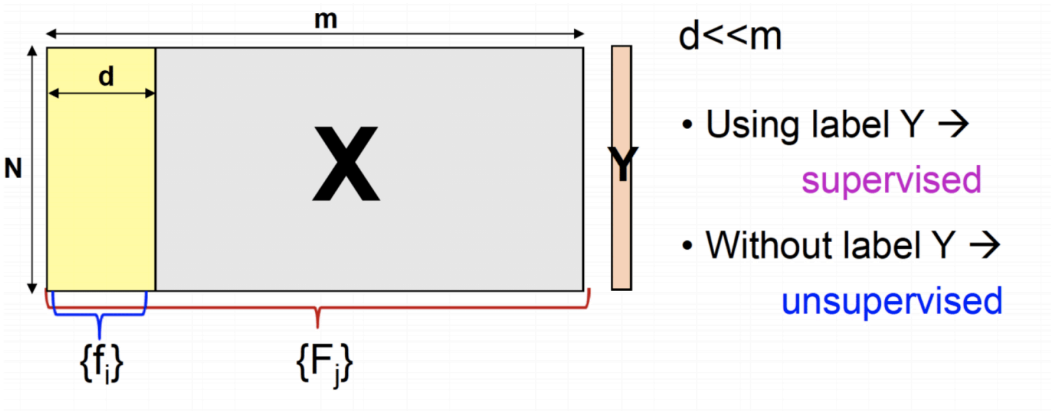
2. Why do we need to select and reduce?
   - Inside huge data, there may consist of a lot of irrelevant genes that we do not want to include in analysis
   - We do not want to include all of the similar genes
   - Some genes are complementary

     Combine them into one value can be more useful

3. The benefit of feature selection and dim reduction
   - Data compression
   - Improve prediction performance
   - Understand the prediction results
   - Facilitate data visualization

Feature Selection

To select/extract the most useful or relevant features to build more efficient, precise and fast learning machines.



In the figure, m is the total number of features, d is the number of useful features, and X is the unused one.

Feature selection vs Feature extraction

Feature selection

- Choose the best subset of all genes
- E.g. Filter and wrapper

Feature extraction

- Extract new features by linear or non-linear combination of the original feature, which can be equivalent to one gene plus another
- e.g. PCA, SVD

To determine whether the feature is useful, we can calculate the correlation between features and class, mutual information I(i), and fisher score F. However, relevance does not imply usefulness, and usefulness does not imply relevance also. In the lecture, the following example is given:

Weight VS gender = 0.7

Height VS gender = 0.8

Major VS gender = 0.6

Although weight and height seem to be useful features, since they are correlated, it may not be the best choice. Instead, Height and major are preferable as the complementary information from the major is much more than the complementary information of weight against height.

Filter

Classification performance will not be involved in selecting useful features. Features with higher variance carry more useful information.

	G1	G2	G3	G4	Cancer
S1	10	2	6	8	Yes
S2	10	3	7	8	Yes
S3	10	4	8	6	No
S4	10	5	9	5	No

From the above example, G1 is removed because all values in G1 are the same. G2 or G3 will be preserved because their correlation is 1. Therefore, in the end, we would keep G2 and G4 or G3 and G4.

Wrapper

It uses the classification performance to guide selection.
The process recursively eliminates and selects features by sequential feature selection.
Take the above table as an example, initially, we find the best feature among G1, G2, G3, and G4 by using cross-fold validation. After calculation, let's say G4 is the best, then we pair up G4 with the remaining feature (i.e. G1, G2, G3) to become G1+G4, G2+G4 and G3+G4 to carry out cross-validation again. Repeat the process until the new feature does not improve the performance.

Dim Reduction

PCA

If a two-dimensional scatter of points shows a high degree of correlation, then we can use PCA to reduce the number of dimensions (e.g. 2D to 1D)

- PCA procedure

1. Construct a ***n*** by ***d*** data matrix and normalize the data to get X'
2. We calculate the covariance of X' by the below formula
   - $\Sigma = \frac{1}{n-1} X'^{T} X'$
3. Find the eigenvectors and eigenvalue of the covariance matrix
4. Combine eigenvectors to find the principal component
5. Project the data to the M eigenvectors' direction by the below formula
   - $\hat{X} = X'P$