Lecture 11: Feature Selection and Dimension Reduction

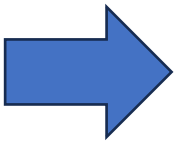
Reasons of conducting feature selection and dimension reduction

Bio – Data

Huge: If the gene expression profile is of 25,000 genes (features) and the single-cell RNA-seq is with 10000 cells (data points), then we need to deal with a 10000 by 25000 matrix (Huge capacity).

Irrelevant and noisy: There are some irrelevant genes that we do not have to include them in our analysis.

Duplicated: There are some highly correlated genes that we do not have to include them in our analysis. Also, some genes are complementary, which needs to combine them into one value.



Dimension Reduction

Benefits

Data Compression: Efficient storage and retrieval

Improve Prediction Peformance and Clustering Speed: Remove unrelated genes.

Facilitate Data Visulization: Dimension reduction to 2D, show the distance between cells visually.

*However, dimension reduction cannot increase the number of data points.

Ways to reduce dimensionality

- Feature Selection
  1. Choose the best subset genes form all the genes: using label Y (supervised)
  2. Feature ranking: discover the most relevant features (target label)

How to measure which ones are useful?

→Correlation between feature & class

→Mutual information

→Fisher score

Issues of individual features ranking

→Relevance does not imply usefulness, vice versa.

→Leads to the selection of a redundant subset.

→A variable that is useless by itself can be useful with others.

  3. Feature subset selection: Filter and Wrapper

Filter

- Classification performance is not involved in the selection loop.
- Variance threshold: features with a higher variance contain more useful information.
- Information gain: features should be different.

Wrapper

- Using the classification performance to guide selection
- Computational expensive and time-consuming
- Recursive feature elimination
- Sequential feature selection

1. No feature
2. Find the 1st best feature using cross-fold validation.
3. Add 2nd feature.
4. …
5. Until the new feature does not improve the performance.

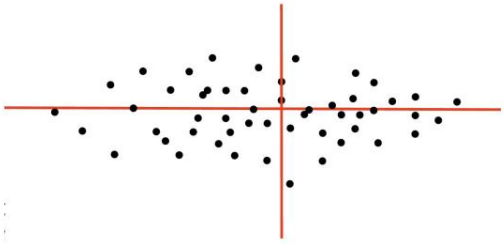
- Feature Extraction
    1. Extract new feature by linear or non-linear combination of the original features. (New feature = Gene1 + Gene 2)
    2. New features may not have physical interpretation (usually for non-linear), which is built for dimension reduction.
    3. Methods: PCA, SVD, Isomap, LLE, CCA, et.al.

Principle components analysis (PCA)

Objective: - Capture the intrinsic variability in the data

- Conduct more efficient description after vector space transform

- Reduce the dimensionality of a data set to ease interpretation



1st dimension captures max variance.

2nd dimension captures the max amount of residual variance, at right angles (orthogonal) to the first.

It is supposed to ignore the remaining axis since the 1st dimension has captured much information.

Suppose there is a n by d data matrix (X)

Steps: 1. Normalize each feature to make the average of each feature 0.

The normalized matrix is called X'

2. Calculate the covariance matrix of X'

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

3. Find the eigenvectors and eigenvalues of Σ

4. M eigenvectors with the M largest eigenvalues (Principal components)

5. Project the data to the M eigenvectors' direction

$$\hat{X} = X'P$$

Application:

PCA should be used mainly for variables which are strongly correlated. If the relationship is weak between variables, PCA does not work well to reduce data, which will lead to essential information loss.