BMEG3105	Fall 2023

Data Analytics for Personalized Genomics and Precision Medicine
Lecture 11: Feature selection & dimension reduction

Lecturer: Yu LI (李煜) from CSE
Liyu95.com, liyu@cse.cuhk.edu.hk
Scriber: Lo Ka Yee
SID: 1155143047

11 October 2023

I. Recap from previous lecture
II. Reasons for feature selection & dimension reduction
III. Feature selection
IV. Dimension reduction

I. Recap from previous lecture

- N-fold vs leave-one-out validation
  - N-fold: build n classifiers to validate the model, grouping of data can be random
  - Leave-one-out: number of classifiers equal to number of data points
- How logistic regression is used for multi-class classification
  - Build logistic regression function for each class
  - Prediction is done by assigning the class with highest value

II. Reasons for feature selection

- Huge volume of Bio-data
- Bio-data consist of noisy, unrelated, and duplicated data
  - Irrelevant genes
  - Highly correlated genes
  - Complementary genes
- Benefits of feature selection and dimension reduction
  - Data compression for efficient storage and retrieval
  - Improve prediction performance
  - Understand the prediction results
  - Facilitate data visualization

III. Feature selection

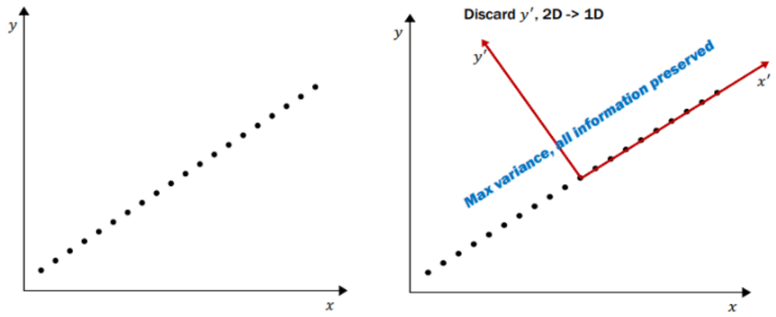
- Feature: genes; Data point: cells
- Select/ extract the most relevant features to build a better model
- Methods to reduce dimensionality
    - Feature selection
        - Choose the best subset genes from all the genes
    - Methods of feature ranking (find the most relevant features)
        - Correlation
            - Calculate the correlation between individual feature and class
        - Mutual information
        - Fisher score
    - Issues of individual features ranking
        - Relevance and usefulness are not correlated
        - Selection of redundant subset
        - Some features may be useful only with other features

    - Feature subset selection: Filter and Wrapper
        - Filter
            - Classification performance not involved
            - Higher variance -> more useful information
            - Information gain should be different for features

        - Wrapper
            - Sequential feature selection
                - Selection based on classification performance of features
                - Computational expensive
                - Recursive feature elimination
                - Sequential feature selection
                - Process:
                    - Build a model for each feature and find out the best feature
                    - Add the second feature cross validation to check the performance
                    - Add feature until the new feature does not improve performance

IV. Dimension reduction

- Feature extraction
  - Extract new features by linear or non-linear combination of the original features
  - Principal components analysis (PCA)
    - Vector space transformation



    - In this case: After vector transformation, x' can capture the maximum variance, while y' can capture none. y' is removed, so that one dimension can be removed, but information is preserved

  - How to do PCA
    - Normalize each feature in a data matrix X to get X' so that the average of each feature is 0.
    - Calculate the covariance matrix of X'
      - $\Sigma = \frac{1}{n-1} X'^T X'$, Σ: a d by d matrix
    - Find the eigenvectors and eigenvalues of Σ
    - The principal components are the M eigenvectors with the M largest eigenvalues
    - Project the data to the M eigenvectors' direction

  - PCA Example illustration:
    - Matrix X:

X1	1	1	1
X2	2	2	2
X3	3	3	3

- Normalization of X to X'

X1	-1	-1	-1
X2	0	0	0
X3	1	1	1

- Calculate the covariance matrix of X'

$$\Sigma = \frac{1}{n-1} X'^T X' = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

- Find the eigenvalues and vectors of Σ

$$\Sigma * V = \lambda * V$$

$$|\Sigma - \lambda I| = 0$$

$$\begin{vmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & 1-\lambda \end{vmatrix} = 0$$

$$(1-\lambda)^3 + 1 + 1 - (1-\lambda) \\ - (1-\lambda) - (1-\lambda) = 0$$

- We will find that Λ_1=3, $\Lambda_{2,3}$=0, subsituting the eigenvalues into the equation, we can find the respective eigenvectors.

$$\lambda_1 = 3 \qquad V_1 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \end{bmatrix} \qquad \lambda_{2,3} = 0 \qquad V_{2,3} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The V_1 here is normalized.

- Project the data to M eigenvectors' direction

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

$$P = \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \end{bmatrix} \qquad \hat{X} = X'P = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} * \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \end{bmatrix} = \begin{bmatrix} -\sqrt{3} & 0 \\ 0 & 0 \\ \sqrt{3} & 0 \end{bmatrix}$$

Therefore, we can obtained a reduced data matrix:

X1	$-\sqrt{3}$	0
X2	0	0
X3	$\sqrt{3}$	0