Name: Mok Nga Sze

SID: 1155193690

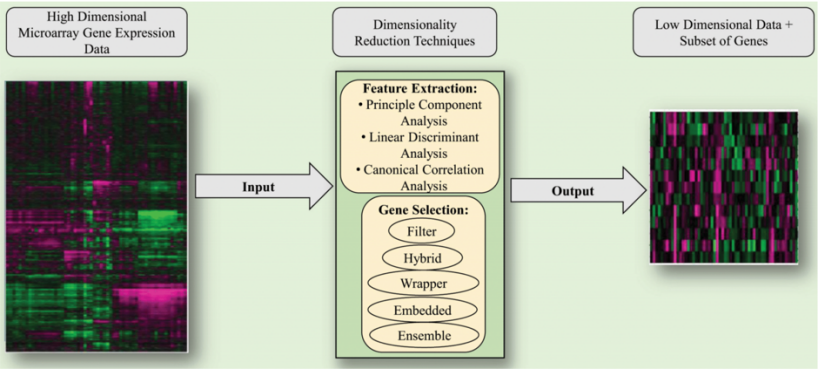
# Feature selection & dimension reduction

## Lecture outline:

1) Why feature selection & dimension reduction
2) Feature selection
3) Dimension reduction (PCA, Example of PCA)

## Why feature selection & dimension reduction

1) Bio-data can be huge
2) Bio-data can be noisy, unrelated, and duplicated (We do not have to include all genes, combine them into one value is more useful)
   - Irrelevant genes
   - Highly correlated genes
   - Complementary genes



## Benefit of feature selection and dimension reduction

1) Data compression
   - Efficient storage and retrieval
2) Improve prediction performance
   - Remove unrelated inputs
3) Understand the prediction results
   - What genes are related to the cancer prediction
4) Facilitate data visualization
   - Understand the distance between cells visually

## Feature Selection/Extraction

- Select/extract the most relevant one to build better, faster, and easier to understand learning machines

## How to reduce dimensionality

1) Feature selection
   - Choose the best subset genes from all the genes
   - Feature ranking
   - Feature subset selection: Filter and Wrapper
2) Feature extraction
   - Extract new features by linear or non-linear combination of the original features
   - New features may not have physical interpretation/meaning (usually for non-linear)
   - PCA, SVD, Isomap, LLE, CCA…

## Feature ranking

- Build better, faster, and easier to understand learning machines
- Discover the most relevant features w.r.t. target label

## How to measure which ones are useful

- Correlation between feature and class
- Mutual information I (The higher I, the attribute is more related to the class)
- Fisher score F (The higher F, the attribute is more related to the class)

## Issues of individual features ranking

- Relevance vs usefulness
- Leads to the selection of a redundant subset
- A variable that is useless by itself can be useful with others

## Subset feature selection

1) Filter
   - Classification performance is not involved in the selection loop
   - Variance thresholds: Features with a higher variance contain more useful information
   - Information gain: Features should be different

   e.g.

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

2) Wrapper
   - Using the classification performance to guide selection
   - Computational expensive

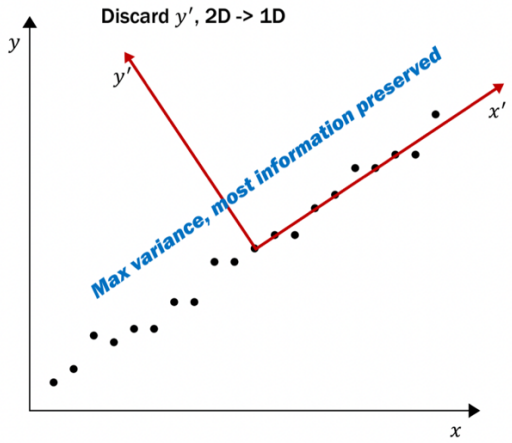
- Recursive feature elimination
- Sequential feature selection
- Process:
  1) Build a model for each feature and find out the best feature
  2) Add the second feature cross validation to check the performance
  3) Add feature until the new feature does not improve performance

e.g.

| | 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 |

## Principal components analysis (PCA)

- A two dimensional scatter of points that show a high degree of correlation
- We care about variance (information) and distance




- After vector space transform, we have more "efficient" description
- 1st dimension captures max variance
- 2nd dimension captures the max amount of residual variance, at right angles (orthogonal) to the first
- The 1st dimension may capture so much of the information content in the original data set that we can ignore the remaining axis

## How to do PCA?

1) Normalize each feature in a data matrix X to get X' so that the average of each feature is 0
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) The principal components are the M eigenvectors with the M largest eigenvalues
5) Project the data to the M eigenvectors' direction

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

Example of PCA

- Matrix X:

$X$

| | | | |
|---|---|---|---|
| X1 | 1 | 1 | 1 |
| X2 | 2 | 2 | 2 |
| X3 | 3 | 3 | 3 |

1) Normalization of X to X'

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

$X'$

2) Calculate the covariance matrix of X'

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

3) Find the eigenvalues and vectors of Σ

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

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

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

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

$$\lambda = 3 \; or \; \lambda = 0$$

$$\Sigma * V = \lambda * V \qquad \lambda_1 = 3 \; or \; \lambda_2 = 0$$

$$(\Sigma - \lambda I) * V = 0$$

$$\begin{bmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & 1-\lambda \end{bmatrix} * \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0$$

$$\lambda_1 = 3 \qquad V_1 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \end{bmatrix}$$

4) 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 V1 here is normalized)

5) Project the data to M eigenvectors' direction

**2D** $$P = \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \end{bmatrix} \qquad \hat{X} = X'P = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} * \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \\ \frac{\sqrt{3}}{3} & 0 \end{bmatrix} = \begin{bmatrix} -\sqrt{3} & 0 \\ 0 & 0 \\ \sqrt{3} & 0 \end{bmatrix}$$

6) Therefore, we can obtain a reduced data matrix:

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

$\hat{X}$