BMEG3105

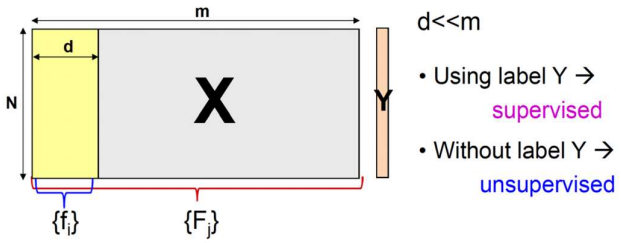
Scribing for Lecture 11 – Dim Reduction & Visualization

Lecturer: Yu Li

1. **The need of feature selection and dimension reduction**
   Bio-data can be huge, which may contain more than 1TB of genetic information. However, some data are not too useful for processing, such as duplicated or non-expressive genes.
   To reduce the need of memory space and processing time, Selection of data and features are needed.

2. **Benefits of feature selection and dimension reduction**
   - More effective of storage and retrieval of data
   - Remove unrelated/redundant input
   - Understand which genes are related to the prediction and which are not
   - Allows data visualization

3. **Feature selection**
   To select/extract the most useful features to build more efficient and accurate learning machines.



In the figure, d represents the useful data while X represents the not useful ones. Only those useful data are kept for processing.
To determine whether the features are useful or not, we may calculate the correlation between the features and class, mutual information I(i) or Fisher score. However, relevant data does not mean useful data, and vice versa. For example:
Height VS gender = 0.8
Weight VS gender = 0.7
Major VS gender = 0.6
Although height and weight seem to be the two most useful data, due to the reason height and weight having high correlation with each other, choosing height and major over height and weight is more preferrable. A useless feature can be useful when paired with others.

- Filter
  Classification performance is not used to determine the usefulness of a feature.
  Features with greater variances are considered being more useful.
  The features should have difference.

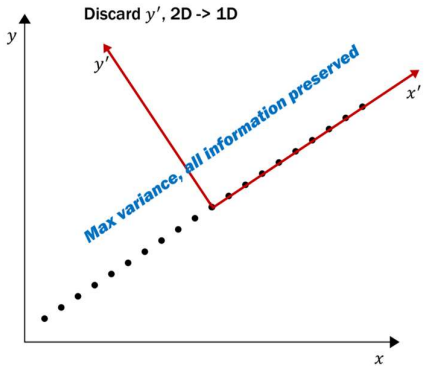
	G1	G2	G3	G4
S1	10	2	6	8
S2	10	3	7	8
S3	10	4	8	6
S4	10	5	9	5

  On the above table, if filter is used, G1 will be removed since there is no difference between the values. Then either G2 or G3 will be preserved since they have a correlation of 1. At the end, it's either G2 and G4 or G3 and G4.

- Wrapper
  It utilizes the classification performance to guide performance.
  It eliminates recursive features, and selects features sequentially.
  Taking the above table as an example again, initially, the usefulness of G1, G2, G3 and G4 are determined individually by cross-fold validation. After the best is determined (e.g. G4), the best one is paired up with the remaining features (e.g. G4+G3, G4+G2…) and carries out cross-fold validation again. This process repeats until a new feature does not improve the performance, and that new feature will thus be removed.

4. Dimension reduction (PCA)

If a multi-dimensional scatter of points shows high correlation, we can try to reduce the number of dimensions (e.g. 2D to 1D). It is noticeable that a certain dimension will catch most of the variance, that the negligible residual variance orthogonal to the first is captured by the remaining dimensions, yet since it's insignificant, PCA can be carried out. Therefore, to determine whether PCA can be performed, we can note its distance and variance.



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

    $$\Sigma = \frac{1}{n-1} X'^T X'$$

  - Find the eigenvectors and eigenvalues of the covariance matrix
  - Combine the eigenvectors to give the principal components
  - Project the data to the eigenvectors' directions, by performing matrix calculation

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