BMEG3105 Lecture 6 Scribing

Scriber: Lai Yuk Fai (1155190593)

Recap from last lecture:

1. From sequence to gene expression
2. Genome assembly
3. Mapping
4. Data Cleaning:
    - Denoise data
    - Remove outliers
    - Handling missing data
    - Remove duplicates
    - Categoric data encoding
    - Data normalization
5. Data Exploration:
    - Summary Statistics: Location, Spread, Frequency
    - Visualization: Distribution & Trend, Histograms, Box plots

Today's Agenda:

1. Clustering
    - Why Clustering?
        i. Understanding:
            - As a stand-alone tool to get insight into data distribution
            - As a pre-processing step for other algorithms
        ii. Summarization:
            - Reduce the size of large data sets
            - Preserve privacy
    - What is clustering analysis?
        i. Finding groups of objects such that the objects in a group will be similar (or related) to one another and different form (or unrelated to) the objects in other groups
    - What are needed to do clustering?
        i. Data to be clustered
        ii. Similarity measurement
        iii. Clustering algorithm
2. Similarity and Dissimilarity
    - Similarity: numerical measure of how alike two data objects are ([0,1], higher = more similar)
    - Dissimilarity: numerical measure of how different two data objects are (0 = identical, lower = more similar)

- Cosine similarity: (where • indicate vector dot product and $|d|$ is the length of the vector d)

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{(|d_1| * |d_2|)}$$

- Correlation: (measures the linear relationship between objects)

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- Euclidean distance: (where m is the number of dimensions (attributes) and p_k and $q_k$are, respectively, the k-th attributes (components) or data objects p and q)

$$Ed(p,q) = \sqrt{\sum_{k=1}^{m} (p_k - q_k)^2}$$

- Minkowski distance: (where r is a parameter, m is the number of dimensions (attributes) and p_k and $q_k$are, respectively, the k-th attributes (components) or data objects p and q)

$$dist(p,q) = (\sum_{k=1}^{m} |p_k - q_k|^r)^{\frac{1}{r}}$$

  i. r = 1 → City block (Manhattan, taxicab, L_1 norm) distance
     - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
  ii. r = 2 → Euclidean distance
  iii. r = ∞ → "Supremum" (L_{max} norm, L_∞ norm) distance
     - This is the maximum difference between any component of the vectors

3. Hierarchical Clustering
   - Produces a set of nested clusters organized as a hierarchical tree
   - Can be visualized as a dendrogram (a tree like diagram that records the sequences of merges)
   - They may correspond to meaningful taxonomies (gene clusters, phylogeny reconstruction, animal kingdom)
   - Steps:
     i. Compute the Similarity or Distance matrix
     ii. Let each data point be a cluster
     iii. Merge the two closest clusters x n times
     iv. Update the similarity or distance matrix x n times

v. Until only a single cluster remains

- How to update the distance matrix after merging?
  i. E.g. use linear correlation:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

  ii. Treat each gene as a cluster
  iii. Merge the gene
  iv. Update with minimum distance (Largest correlation)

Potential Project 2

1. Data preprocessing for the gene expression matrix
   - Data collecting and merging
   - Exploration
   - Visualization
   - Data cleaning
   - Get distance matrix
   - Perform clustering
2. Dimension reduction

Advanced Topic

1. Mahalanobis distance:

$$mahalanobis(p,q) = (p-q)^T \sum^{-1} (p-q)$$

   - Where the summation sign is the covariance matrix