BMEG 3105 Fall 2024
Data analytics for personalized genomics and precision medicine
Lecture 8

Recap from last lecture

Clustering

1. Data to be clustered
2. Similarity measurement
3. Clustering algorithms

Similarity measurement

- Cosine similarity
- Correlation
- Euclidean distance
- Manhattan distance
- Mahalanobis distance

Clustering

Hierarchical clustering

1. Compute the similarity or distance matrix
2. Let each data point be its own cluster
3. While number of clusters > 1:
    a. Merge 2 closest cluster
    b. Update similarity or distance matrix

Distance calculation between clusters

- Min
- Max
- Group Average
- Distance between centroids

Mahalanobis distance

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

Calculating distance considering the data distribution

Classification

Find a method to assign the class of previously unseen records based on their other attributes and the training set as accurately as possible.

1. Training data with class
2. Classification method
3. Data to be classified

K-nearest neighbour

1. Training process:
    a. Store the available training instances
2. Predicting process:
    a. Find the K training instances that are closest to the query instance
        i. K normally between 5-10
        ii. Chosen by cross-validation
    b. Return the most frequent class label among those K instances

Data should be normalized

Clustering vs Classification

	Clustering	Classification
Goal	Find similarity (clusters) in the data	Assign class to the new data
Data	Data without class	Training data with class and testing data without class
Classes	Unknown number of classes	Known number of classes
Output	The cluster index for each point	The class assignment of the testing data
Algorithm	One phase	Two phases (training and application)

- Unsupervised learning
    - Machine learning algorithms to analyse and cluster unlabelled data
    - Example: clustering and dimension reduction
- Supervised learning

- Machine learning algorithms to classify and predict outcomes, trained on labelled data
- Example: classification and regression