CS 466
Introduction to Bioinformatics

Instructor: Jian Peng
Teaching Assistant: Wesley Qian
Finding hidden structure in data
Expression analysis

- Blood
- Liver
- Brain
Single-cell expression analysis
Clustering: examples

**Image segmentation**
Goal: Break up the image into meaningful or perceptually similar regions
Network clustering
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns
Clustering

- **Basic idea:** group together similar instances
- **Example:** 2D point patterns

- **What could “similar” mean?**
  - One option: small Euclidean distance (squared)
    \[
    \text{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2
    \]
  - Clustering results are crucially dependent on the measure of similarity (or distance) between “points” to be clustered
Given: $N$ unlabeled examples $\{x_1, \ldots, x_N\}$; the number of partitions $K$
Goal: Group the examples into $K$ partitions

- The only information clustering uses is the similarity between examples
- Clustering groups examples based on their mutual similarities
Clustering algorithms

1. **Flat or Partitional clustering** (K-means, Gaussian mixture models, etc.)
   - Partitions are independent of each other

2. **Hierarchical clustering** (e.g., agglomerative clustering, divisive clustering)
   - Partitions can be visualized using a tree structure (a dendrogram)
   - Does not need the number of clusters as input
   - Possible to view partitions at different levels of granularities (i.e., can refine/coarsen clusters) using different K
K-means

- **Input:** $N$ examples $\{x_1, \ldots, x_N\}$ ($x_n \in \mathbb{R}^D$); the number of partitions $K$
- **Initialize:** $K$ cluster centers $\mu_1, \ldots, \mu_K$. Several initialization options:
  - Randomly initialized anywhere in $\mathbb{R}^D$
  - Choose any $K$ examples as the cluster centers
- **Iterate:**
  - Assign each of example $x_n$ to its closest cluster center
    
    $$C_k = \{n : k = \arg \min_k ||x_n - \mu_k||^2\}$$

    ($C_k$ is the set of examples closest to $\mu_k$)
  - Recompute the new cluster centers $\mu_k$ (mean/centroid of the set $C_k$)
    
    $$\mu_k = \frac{1}{|C_k|} \sum_{n \in C_k} x_n$$

  - Repeat while not converged
K-means for segmentation

K=2

K=3

K=10

Original
When will K-means fail?

Non-convex/non-round-shaped clusters: Standard $K$-means fails!

Clusters with different densities
Hierarchical clustering

A hierarchical approach can be useful when considering versatile cluster shapes:

10-means

By first detecting many small clusters, and then merging them, we can uncover patterns that are challenging for partitional methods.
Agglomerative clustering:

- First merge very similar instances
- Incrementally build larger clusters out of smaller clusters

Algorithm:

- Maintain a set of clusters
- Initially, each instance in its own cluster
- Repeat:
  - Pick the two closest clusters
  - Merge them into a new cluster
  - Stop when there’s only one cluster left

Produces not one clustering, but a family of clusterings represented by a dendrogram
We need a notion of similarity between clusters.
Single linkage uses the minimum distance.
Example
Complete linkage uses the maximum distance.
Group average linkage uses the average distance between groups.
Mouse tumor data from [Hastie et al.]
Application to breast cancer expression data
Human population