What is Clustering Analysis?

- Goal: Group objects so that each cluster has **high internal similarity** and is **well-separated** from others
    - *Intra-cluster differences → small*
    - *Inter-cluster differences → large*
- Core components:

1. Data to cluster
2. Similarity measure
3. Clustering algorithm

Distance & Similarity Measures

Minkowski Distance (generalized L_r norm):

- **r=1**: Manhattan (a.k.a. city block/taxicab). Includes Hamming distance for binary data
- **r=2**: Euclidean
- **r→∞**: Supremum (L∞), i.e., maximum component-wise difference

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Other measures mentioned: cosine similarity, correlation, Mahalanobis distance

---

Hierarchical Clustering (Main Topic)

- Produces a **nested set of clusters** represented by a *dendrogram*
- Useful in taxonomy-like structures (e.g., biology, gene clustering, phylogeny)

Steps:

1. Compute similarity/distance matrix
2. Each data point starts as a single-element cluster
3. Iteratively merge the **closest clusters**
4. After each merge, update the distance matrix
5. Repeat until only one cluster remains

Methods for Updating: min, max**,** Group average**,** Centroid distance

---

Running Example (Gene Expression Clustering)

Gene	wt	mutant_1	mutant_2	mutant_3
At4g35770	1.5	3	3	1.5
At1g30720	4	7.5	7.5	5
At4g27450	1.5	1	1	1.5
At2g34930	10	25	23	15
At2g05540	1	1	2	1

- Similarity measured with **linear correlation**

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

- Stepwise process shown:
    1. Each gene has its own cluster
    2. Merge pairs with highest correlation
    3. Update matrix (using min correlation as criterion)
    4. Continue until a hierarchy is built

Mahalanobis Distance (Advanced Topic)

- Considers both data distribution & covariance
- Uses covariance matrix (Σ) → requires its inverse for computation
- Intuition: Instead of raw numerical difference, scale by variance/covariance
    - Example: quiz score differences of same magnitude may not be equivalent if variances differ
    - Essentially measures "number of standard deviations apart" between points
- Often combined with **normalization** (e.g., min-max scaling)

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

Min-max normalization

Person	Height	Weight
P1	0.625	0.875
P2	0	0
P3	0.25	0.375
P4	1	1

---

11. Conclusion

- Clustering = unsupervised method that relies on similarity measures
- Hierarchical clustering builds a dendrogram structure
- Choice of distance metric (Euclidean, correlation, Mahalanobis)