Cluster Similarity

**Similarity** is most often measured with the help of a *distance function*. The smaller the distance, the more similar the data objects (points).

A function $d: M \times M \rightarrow \mathbb{R}$ is a **distance** on $M$ if it satisfies for all $\hat{x}, \hat{y}, \hat{z} \in M$ (where $M$ is an arbitrary non-empty set and $\mathbb{R}$ is the set of real numbers):

- $(i)$ $d(\hat{x}, \hat{y}) = 0 \iff \hat{x} = \hat{y}$ (coincidence axiom)
- $(ii)$ $d(\hat{x}, \hat{y}) = d(\hat{y}, \hat{x})$ (symmetry)
- $(iii)$ $d(\hat{x}, \hat{z}) \leq d(\hat{x}, \hat{y}) + d(\hat{y}, \hat{z})$ (triangle inequality)
Distance Functions: Minkowski

The Minkowski family of distance functions is defined as:

\[ d_k(x, y) = \left( \sum_{i=1}^{n} (x_i - y_i)^k \right)^{\frac{1}{k}} \]

where \( n \) is the number of features and \( k \) is a parameter.

- \( k = 1 \): Manhattan distance (or city block distance)
- \( k = 2 \): Euclidean distance
- \( k \to \infty \): maximum distance (i.e. \( d_\infty(x, y) = \max_{i=1}^{n} |x_i - y_i| \))
Distance Functions: Mahalanobis

Mahalanobis distance is defined as:

\[ d(\hat{x}, \hat{y}) = \sqrt{(\hat{x} - \hat{y})^T \Sigma^{-1} (\hat{x} - \hat{y})} \]

where \( \Sigma \) is the covariance matrix of \( \hat{x} \) and \( \hat{y} \).

Euclidean Distance

Mahalanobis Distance
Exclusive vs. Overlapping vs. Fuzzy

• An **exclusive clustering** assigns each object to a single cluster.
• An **overlapping clustering** or non-exclusive clustering is used to reflect the fact that an object can **simultaneously** belong to more than one group (class).
• A **fuzzy cluster** assigns objects to a cluster with a membership weight that is between 0 (absolutely doesn’t belong) and 1 (absolutely belongs).
Clustering Using Mixture Models

• Often, it is convenient and effective to assume that data has been generate by a statistical process.
• To describe the data, we can find the statistical model (described by a distribution and a set of parameters for that distribution) that best fits the data.
• **Mixture models** use statistical distributions to model the data, each distribution corresponding to a cluster and the parameters of each distribution describing the corresponding cluster.
Mixture Models

- Mixture models view the data as a set of observations from a mixture of different probability distributions.

- The Idea:
  - Given several distributions, usually of the same type, but with different parameters, randomly select one of these distributions and generate an object from it.
  - Repeat the process $m$ times, where $m$ is the number of objects.
The EM Algorithm

EM algorithm

1: Select an initial set of model parameters.

2: repeat

3: **Expectation Step** For each object, calculate the probability that each object belongs to each distribution, i.e., calculate $prob(distribution \ j | x_i, \Theta)$.

4: **Maximization Step** Given the probabilities from the expectation step, find the new estimates of the parameters that maximize the expected likelihood.

5: until The parameters do not change (or are below a threshold).
Consider two initial distributions. Here, we use two Gaussian distributions, each with its own mean, $\mu$, and variance, $\sigma$.

First, we make an initial guess for $\mu_1$ and $\mu_2$ (perhaps initializing $\sigma_1$ and $\sigma_2$ to a constant).
The EM Algorithm: A Visual Description

For the expectation step, we compute the probability that a point came from a particular distribution. These values can be expressed by an application of Bayes rule.
The EM Algorithm: A Visual Description

After computing the cluster membership probabilities for all the points, we compute new estimates for \( \mu_1 \) and \( \mu_2 \). This is the maximization step. Here, the new estimate for the mean of a distribution is just a weighted average of the points, where the weights are the probabilities that the points belong to the distribution.
The EM Algorithm: A Visual Description

We repeat these two steps until the estimates of $\mu_1$ and $\mu_2$ either don’t change or change very little.
Mixture Models: A Mathy Description

Assume $K$ distributions and $m$ objects, $X = \{x_1, \ldots, x_m\}$. Let the $j^{th}$ distribution have parameters $\theta_j$, and let $\Theta$ be the set of all parameters, i.e. $\Theta = \{\theta_1, \ldots, \theta_K\}$. The probability that the $j^{th}$ distribution is chosen to generate an object is given by the weight $w_j$, $1 \leq j \leq K$, where these weights (probabilities) sum to one, i.e., $\sum_{j=1}^{K} w_j = 1$. Then the probability of an object $x$ is given by:

$$\text{prob}(x_i | \Theta) = \sum_{j=1}^{K} w_j p_j(x_i | \theta_j)$$

The probability of object $x$ if it comes from the $j^{th}$ distribution.
Mixture Models: A Mathy Description

If the objects are generated in an independent manner, then the probability of the entire set of objects is just the product of the probabilities of each individual $x_i$

$$prob(X|\Theta) = \prod_{i=1}^{m} prob(x_i|\Theta) = \prod_{i=1}^{m} \sum_{j=1}^{K} w_j p_j(x_i|\theta_j)$$

The probability of $i^{th}$ object.

The probability of $i^{th}$ object if it comes from the $j^{th}$ distribution.
Gaussian Mixture: A Mathy Description

Consider a mixture model in terms of Gaussian distributions. The probability density function for a one-dimensional Gaussian distribution at a point $x$ is:

$$prob(x_i | \Theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

The parameters of the Gaussian distribution are given by $\theta = (\mu, \sigma)$, where $\mu$ is the mean of the distribution and $\sigma$ is the standard deviation.
MLE: A Mathy Description

Consider a set of $m$ points that are generated from a one-dimensional Gaussian distribution:

$$prob(X|\Theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\log prob(X|\Theta) = - \sum_{i=1}^{m} \frac{(x_i - \mu)^2}{2\sigma^2} - 0.5m \log 2\pi - m \log \sigma$$

One approach to estimate $\mu$ and $\sigma$ is to choose the values of the parameters for which the data is most probable (most likely), known as the maximum likelihood principle (MLE).
MLE: A Mathy Description

The concept is called the maximum likelihood principle because, given a set of data, the probability of the data, regarded as a function for the parameters, is called a likelihood function.

\[ \text{likelihood}(\Theta|X) = (\Theta|X) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \]

For practical reasons, the log likelihood is more commonly used. Note that the parameter values that maximize the log likelihood also maximize the likelihood.

\[ \log \text{likelihood}(\Theta|X) = \ell(\Theta|X) = -\sum_{i=1}^{m} \frac{(x_i - \mu)^2}{2\sigma^2} - 0.5m \log 2\pi - m \log \sigma \]
Example of EM: Initialization

Begin the EM algorithm by making initial guesses for $\mu_1$ and $\mu_2$, say, $\mu_1 = -2$ and $\mu_2 = 3$. Thus, the initial parameters, $\theta = (\mu, \sigma)$, for the two distributions are, respectively, $\theta_1 = (-2,2)$ and $\theta_1 = (3,2)$. The set of parameters for the entire mixture model is $\Theta = \{\theta_1, \theta_2\}$. 
Example of EM: Expectation Step

For the expectation step of EM, we want to compute the probability that a point came from a particular distribution. These values can be expressed by a straightforward application of Bayes’ rule:

\[
prob(distribution \ j|x_i, \theta) = \frac{0.5 \ prob(x_i|\theta_j)}{0.5 \ prob(x_i|\theta_1) + 0.5 \ prob(x_i|\theta_2)}
\]

where 0.5 is the probability (weight) of each distribution and \( j \) is 1 or 2.
Example of EM: Maximization Step

After computing the cluster membership probabilities for all the points, we compute new estimates for $\mu_1$ and $\mu_2$ in the maximization step of the EM algorithm.

$$\mu_j = \sum_{i=1}^{m} x_i \frac{\text{prob}(\text{distribution } j| x_i, \Theta)}{\sum_{i=1}^{m} \text{prob}(\text{distribution } j| x_i, \Theta)}$$

where $j$ is a particular distribution and $m$ is the number of objects.
Mixture Model Clustering Using EM

- The EM algorithm can be slow.
- Does not work well when clusters contain only a few data points or if the data points are nearly co-linear.
- The problem in estimating the number of clusters or choosing the exact form of the model to use.
- More general than $k$-means because they can use distributions of various types.
- Easy to characterize the clusters produced.
Clustering Evaluation

1. Determining the **clustering tendency** of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.

2. Determining the correct number of clusters.

3. Evaluating how well the results of a cluster analysis fit the data **without** reference to external information.

4. Comparing the results of a cluster analysis to externally known results, such as externally provided class labels.

5. Comparing two sets of clusters to determine which is better.
Types of Cluster Evaluation

**Unsupervised:** Measures the goodness of a clustering structure without respect to external information.

**Supervised:** Measures the extent to which the clustering structure discovered by a clustering algorithm matches some external structure.

**Relative:** Compares different clusterings or clusters with a supervised or unsupervised evaluation measure that is used for the purpose of comparison.
Overall Cluster Validity

In general, we can consider expressing overall cluster validity for a set of $K$ clusters as a weighted sum of the validity of individual clusters.

$$\text{overall validity} = \sum_{i=1}^{K} w_i \text{validity}(C_i)$$

The validity function can be cohesion (compactness, tightness), separation (isolation), or some combination.
Prototype-Based Validity

- **Cohesion** can be defined as the sum of the proximities with respect to the prototype of the cluster.
- **Separation** can be measured by the proximity of the two cluster prototypes.
The Silhouette Coefficient

This method combines cohesion and separation:

1. For the $i^{th}$ object, calculate its average distance to all other objects in its cluster, $a_i$.

2. For the $i^{th}$ object and any cluster not containing the object, calculate the object’s average distance to all the objects in the given cluster. Find the minimum such value with respect to all clusters, $b_i$.

3. For the $i^{th}$ object, the silhouette coefficient is:
   \[ s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \]
The Silhouette Coefficient

• The silhouette coefficient combines both cohesion and separation.

• The value can vary between -1 and 1.
  – A negative value is undesirable, because this corresponds to a case in which $a_i$, the average distance to points in the cluster, is greater than $b_i$, the minimum average distance to points in another cluster.
  – We want the silhouette coefficient to be positive $a_i < b_i$, and for $a_i$ to be as close to 0 as possible.
Hierarchical Cluster Evaluation

• The **cophenetic distance** between two objects is the proximity at which an agglomerative hierarchical clustering technique puts the objects in the same cluster for the first time.
  – For example, if at some point in the agglomerative hierarchical clustering process, the smallest distance between the two clusters that are merged is 0.1, then all points in one cluster have a cophenetic distance of 0.1
Example of Cophenetic Distance
(Single Link Clustering)

Hierarchical Tree Diagram

Cophenetic Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>p6</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>p2</td>
<td>0.22</td>
<td>0</td>
<td>0.15</td>
<td>0.15</td>
<td>0.14</td>
<td>0.15</td>
</tr>
<tr>
<td>p3</td>
<td>0.22</td>
<td>0.15</td>
<td>0</td>
<td>0.15</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>p4</td>
<td>0.22</td>
<td>0.15</td>
<td>0.15</td>
<td>0</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>p5</td>
<td>0.22</td>
<td>0.14</td>
<td>0.15</td>
<td>0.15</td>
<td>0</td>
<td>0.15</td>
</tr>
<tr>
<td>p6</td>
<td>0.22</td>
<td>0.15</td>
<td>0.11</td>
<td>0.15</td>
<td>0.15</td>
<td>0</td>
</tr>
</tbody>
</table>
Summarization of Clustering

- The greater the similarity (or homogenity) within a group and the greater the difference between groups, the better or more distinct the clustering.
- Clusters can be partitional (i.e., $k$-means) or hierarchical (i.e., agglomerative hierarchical).
- Clusterings can be exclusive, overlapping, or fuzzy (i.e., mixture models).
- Cluster evaluation assessed the validity of clusters.
Summarization of Clustering & Association
Summarization of Clustering & Association

• Association analysis and cluster analysis are both types of unsupervised learning (the process of trying to find hidden structure in unlabeled data).

• Association analysis is the process of discovering interesting relations between variables in large datasets.

• Cluster analysis is the process of grouping data objects based only on the description of objects and their relationships.
And now...

*Lets see some cluster evaluation!*