Data mining
Measures – similarities, distances

University of Szeged
Looking for similar data points

- can be important when for example detecting
  - plagiarism
  - duplicate entries (e.g. from search results)
  - recommendation systems (customer A is similar to customer B; product X is similar to product Y)

- What do we mean under similar?
  ⇒ Objects that are only little distance away from each other.
  ⇒ How shall we define some distance?
Axioms of distance metrics

- Function $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined over the n-dimensional point pair $(a, b)$ is a distance metric iff it fulfills the following requirements:
  1. $d(a, b) \geq 0$ (non-negativity)
  2. $d(a, b) = 0 \iff a = b$ (positive definiteness)
  3. $d(a, b) = d(b, a)$ (symmetry)
  4. $d(a, b) \leq d(a, c) + d(c, b)$ (triangle inequality).
Tightly connected concepts

One can easily turn some distance to similarity and vice versa

E.g. given a distance measure $d(a, b)$, we can define similarity $s(a, b)$ as:

- $s(a, b) = -d(a, b)$
- $s(a, b) = \frac{1}{1 + d(a, b)}$
- $s(a, b) = \exp^{-d(a, b)}$
- $s(a, b) = \cos(d(a, b))$, if $d(a, b)$ is given as an angle
Characterization of distances

- **Euclidean vs. non-Euclidean distances**
  - Euclidean distances: distances are determined by the positions of the data points in the (Euclidean) space
  - non-Euclidean distances: distances of points are not directly determined by their positions

- **Metric vs. non-metric distances**
  - Metric distance: all of the axioms of distance metrics hold for them
  - Non-metric distance: at least one of the axioms of distance metrics does not hold for them
    - Example? $d(1\text{PM}, 2\text{PM})$
Minkowski distance

- generalization of Euclidean distance

\[ d(a, b) = \left( \sum_{i=1}^{N} (|a_i - b_i|^p)^{1/p} \right) \]

- \( p = 1 \Rightarrow \) Manhattan distance (\( \ell_1 \) norm) \( \rightarrow \) 7 in the example
- \( p = 2 \Rightarrow \) Euclidean distance (\( \ell_2 \) norm) \( \rightarrow \) 5 in the example
- \( p = \infty \Rightarrow \) Maximum (\( \ell_{\text{max}} \) norm) \( \rightarrow \) 4 in the example
Cosine similarity

- the cosine of the angle enclosed by vectors $a$ and $b$
- Pros? Cons?
- $s_{\cos}(a, b) = \cos \Theta = \frac{a^T b}{\|a\|\|b\|}$ (Proof: at the blackboard)
- Scalar product in case of binary data vectors?
Cosine distance

- Derived from cosine similarity as $d_{\text{cos}} = 1 - s_{\text{cos}}(a, b)$ or $d_{\text{cos}} = \text{arccos} \ s_{\text{cos}}(a, b)$
- $d(a, b) \geq 0$
- $s_{\text{cos}}(a, a) = 1 \Rightarrow d_{\text{cos}}(a, a) = 0$
- $s_{\text{cos}}(a, b) = s_{\text{cos}}(b, a) \Rightarrow d_{\text{cos}}(a, b) = d_{\text{cos}}(b, a)$
- Triangle inequality: rotating from $a$ to $c$ then from $c$ to $b$ has to be at least as much as rotating directly from $a$ to $b$
More 'exotic' distances – Handling inter-dependence among variables

- Mahalanobis distance
  \[ d_{mah}(a, b) = \sqrt{(a - b)^\top \Sigma^{-1} (a - b)} \]
  where \( \Sigma \) is the covariance matrix of the variables in the dataset

![Diagram showing Euclidean distance between points A, B, and C. The question is: d(A,C) < d(A,B)?
What is in Mahalanobis distance?

- Euclidean distance once the data is made uncorrelated
- How could one make $X$ uncorrelated? ($X \in \mathbb{R}^{n \times d}$)
  - We can assume that each feature has mean 0 $\rightarrow X^T X \propto \Sigma$
  - We need $L \in \mathbb{R}^{d \times d}$ such that $(L^T X^T)(XL) = I$
  - It follows that $\Sigma = (LL^T)^{-1} \equiv \Sigma^{-1} = LL^T$, which means $L$
    comes from the Cholesky decomposition of $\Sigma^{-1}$

Reminder

1.) $(AB)^{-1} = B^{-1}A^{-1}$, $(AB)^T = B^T A^T$ and $(A^T)^{-1} = (A^{-1})^T$
2.) Cholesky decomposition: any symmetric, positive definite matrices (such as $\Sigma$)
    have a special LU decomposition where $U = L^T$

\[
\begin{bmatrix}
4 & -4 \\
-4 & 5
\end{bmatrix} = \begin{bmatrix}
2 & 0 \\
-2 & \sqrt{?}
\end{bmatrix} \begin{bmatrix}
2 & -2 \\
0 & \sqrt{?}
\end{bmatrix}
\]

- How would the squared distance of two uncorrelated points look? 
  $(L^T(a - b))^T(L^T(a - b)) = (a - b)^T \Sigma^{-1} (a - b)$
Making data uncorrelated using Cholesky decomposition

\[ \Sigma = [10, 3; 3, 4], \mu = [0, 0], n=500 \]
Distances for distributions

- Bhattacharyya coefficient $BC = \sum_{x \in X} \sqrt{P(x)Q(x)}$
  - We would integrate for continuous variables
  - Quantifies the similarity between distributions
    $BC(P, Q) = 1 \iff P = Q$
Similarity, distance

Bhattacharyya and Hellinger distances

- BC is the basis for various distances
  - Bhattacharyya distance: \( d_B(P, Q) = -\ln BC(P, Q) \)
    - Does not obey triangle inequality
  - Hellinger distance: \( d_H(P, Q) = \sqrt{1 - BC(P, Q)} \)
    - Can be regarded as a special form of Euclidean distance
      \( \left( \frac{1}{\sqrt{2}} \| \sqrt{P(X)} - \sqrt{Q(X)} \|_2 \right) \)
    - E.g. for \( P \sim Bernoulli(0.2) \) and \( Q \sim Bernoulli(0.6) \) we have
      \( BC(P, Q) = \sqrt{0.12} + \sqrt{0.32} = 0.912 \) and
      \( d_H(P, Q) = \sqrt{1 - 0.912} = 0.296 \)
More exotic distances – Variable length feature vectors

- Feature vectors of variable length (e.g. in case of proteins and genes)
  - How similar/different are the two strings AAGCTAA and GGCTA?
- Edit distance: determines the number of deletion and insertion operations needed to transform string $a$ into form $b$
- Many alternations are known (e.g. weighted error types, Levenshtein distance)
- Can be solved with dynamic programming in time $O(mn)$ (where $m$ and $n$ are the lengths of the two words)
- Tight connection with the Longest Common Subsequence (LCS) problem
- $d_{ED}(a, b) = |a| + |b| - 2|LCS(a, b)| = 7 + 5 - 2 \times 4 = 4$
Edit distance – example

- $D[0, j] = j, \forall j \in \{0, 1, \ldots, n\}$
- $D[i, 0] = i, \forall i \in \{0, 1, \ldots, m\}$

$$D[i, j] = \min \begin{cases} d(i - 1, j) + 1, \text{ for deletion} \\ d(i, j - 1) + 1, \text{ for insertion} \\ d(i - 1, j - 1) + 2(1 - a(i) == b(j)), \text{ for replacement} \end{cases}$$

$\Rightarrow d_{ED}(a, b) = D[m, n]$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>5</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>^</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>^</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>
Does edit distance fulfill the metric axioms?

- All edits are weighted non-negatively $\Rightarrow d_{ED}(a, b) \geq 0$
- $d_{ED}(a, a) = |a| + |a| - 2 \times |LCS(a, a)| = 0$
- $d_{ED}(a, b) = d_{ED}(b, a)$ as insertion and deletion operations are weighted equally and inverses of each other.
- Triangle inequality: bringing $a$ into form $b$ in such a way that it is first transformed into $c$ needs at least as many deletions and insertions as transforming it directly into form $b$.
Jaccard similarity

- \( s_{\text{Jacc}}(A, B) = \frac{|A \cap B|}{|A \cup B|} \)

**Example**

- \( s_{\text{Jacc}}(A, B) = \frac{2}{10} = 0.2 \)

- Similarity of multisets
  - \( A = \{x, x, x, y\} \), \( B = \{x, x, y, y, z\} \) \( \Rightarrow s_{\text{Jacc}}(A, B) = \frac{|\{x, x, y\}|}{|\{x, x, x, y, y, z\}|} = \frac{3}{6} \)
Similarity, distance

Jaccard and Dice distances

- \( d_{\text{Jacc}}(A, B) = 1 - s_{\text{Jacc}}(A, B) \)
- one relative of Jaccard similarity: Dice coefficient
- \( s_{\text{Dice}}(A, B) = \frac{2|A \cap B|}{|A| + |B|} \)
- \( d_{\text{Dice}}(A, B) = 1 - \frac{2|A \cap B|}{|A| + |B|} \)