Data Mining / Intelligent Data Analysis: Metrics

Martin Russell

School of Computer Science, University of Birmingham

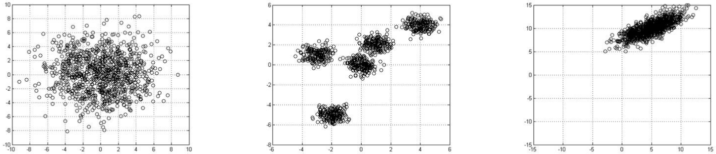
February 12, 2019

Overview

1. Motivation
   - Structure in Data
2. Metrics
   - Properties of metrics
   - The Euclidean metric
   - The L^p metrics
   - Unit spheres
3. Clustering
   - Centroids
   - Distortion
4. Summary

Discovering Structure in Data

Figure: Examples of structured data: spherically distributed single cluster (left), multiple-source data (centre), shifted correlated data (right)



- For example, each data point might be a vector of measurements from an array of sensors in a structure or a train. The clusters in the centre figure might correspond to different states of the structure/train
- **Clustering** discovers structure in multi-source data (center)

What is a metric?

- Let X be a set of vectors. $X \times X$ denotes the set of pairs $X \times X = \{(x, y) : x, y \in X\}$
- A **metric** is a function $d : X \times X \to \mathbb{R}^+$ such that

$$\begin{aligned} d(x,y) &\geq 0, \forall x,y \in X & (1)\\ d(x,y) &= 0, \text{if and only if } x = y & (2)\\ d(x,y) &= d(y,x), \forall x,y \in X & (3)\\ d(x,y) &\leq d(x,z) + d(z,y), \forall x,y,z \in X & (4) \end{aligned}$$

  Property (3) indicates that a metric must be bf symmetric. Property (4) is the **triangle inequality**.
- $\mathbb{R}^+$ is the set of **positive** real numbers
- A metric is sometimes called a **distance function**

The Euclidean metric

- Suppose $X = \mathbb{R}^N$ is N-dimensional space and $x, y \in X$, where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \tag{5}$$

The **Euclidean metric** is given by

$$d_2(x, y) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2} \tag{6}$$

This is the normal notion of distance in Euclidean space

The **squared** Euclidean metric

- Sometimes the square root is ommitted.
- The results is the *squared* Euclidean metric

$$d_2^{sq}(x, y) = \sum_{n=1}^{N} (x_n - y_n)^2 \tag{7}$$

- This is useful, for example, if all that is needed is to find the closest point to a reference point

The L^p metrics

- The Euclidean metric is one of a family of metrics called the L_p **metrics**, denoted by d_p
- In general, for any positive integer p

$$d_p(x,y) = (\sum_{n=1}^{N}(x_n - y_n)^p)^{\frac{1}{p}} \tag{8}$$

The cases $p = 1$ and $p = \infty$ are most common

$$\begin{aligned} d_1(x,y) &= \sum_{n=1}^{N} |x_n - y_n| \quad (9) \\ d_\infty(x,y) &= max_{n=1,\ldots,N}|x_n - y_n| \quad (10) \end{aligned}$$

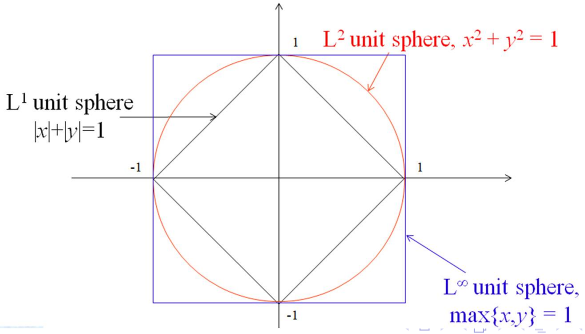
d_1 is referred to as the **City Block** metric

Unit spheres for d_1, d_2 and d_∞

- The **unit sphere** for a metric d is the set

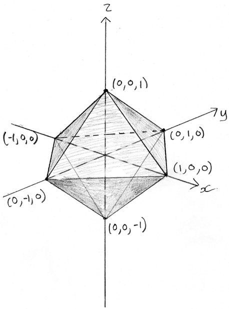
$$U_d = \{x : d(x, 0) = 1\} \tag{11}$$

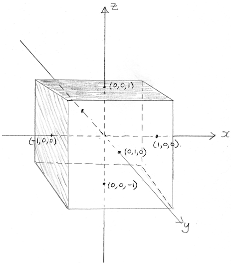
Figure: Unit spheres for d_1, d_2 and d_∞



Unite spheres (3D)

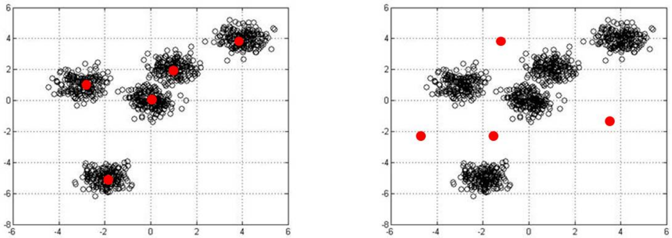
Figure: Unit spheres in 3-dimensions for the City-Block L^1 metric (left) and L^∞ metric (right)





Representing clusters as centroids

Figure: Good (L) and poor (R) representation of clusters with centroids



- *Centroids* are data points located to represent a set of clusters
- Requires correct number of centroids in correct locations
- In general, the number and location of the clusters is unknown

Distortion

- *Distortion* is a measure of how well a set of centroids $C = \{c_1, ..., c_K\}$ fits a set of data $X = \{x_1, ..., x_N\}$
- Let d be a metric
- Let $c_{i(n)}$ be the closest centroid to x_n $(n = 1, ..., N)$

$$d(x_n, c_{i(n)}) = min_{k=1,...,K} d(x_n, c_k) \tag{12}$$

- The *Distortion* for the centroids C relative to the data set X is

$$Dist(C, X) = \frac{1}{N} \sum_{n=1}^{N} d(x_n, c_{i(n)}) \tag{13}$$

Distortion

- $Dist(C, X)$ is the average distance between x_n and its closest centroid
- The best centroid set is the set $\bar{C}$ where

$$Dist(\bar{C}, X) = min_C Dist(C.X) \tag{14}$$

- How do we find $\bar{C}$ - the best set of centroids?
- In general we can't, but we can use a *clustering* algorithm to find a set of centroids that is *locally* optimal
- We'll see how to do this in the next lecture - *K-means* clustering

Summary

- Motivation - cluster analysis
- Properties of a metric
- The L^p family of metrics
- Centroids and distortion