Contents

1 Distant measures

Various types of distance measures and implementations

2 Introduction

A distance measure is simply a means of calculation between two points or objects. Distance measures are useful in identifying patterns in the input data. Also, it can be used to recognize similarities among the data.

2.1 Euclidean Distance

The **Euclidean distance** between two points in Euclidean space is the length of a line segment between the two points. It can be calculated from the cartesian coordinates of the points using the Pythagorean theorem. Euclidean distance works great when you have low-dimensional data and the magnitude of the vectors is important to be measured.

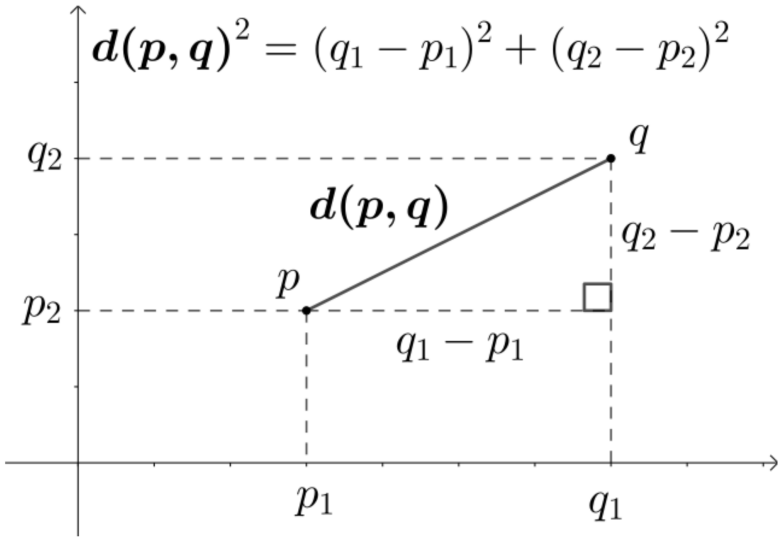

Figure 1: Pythagorean theorem

3 Distance Formulas

3.1 One dimension

The distance between any two points on the real line is the absolute value of the numerical difference of their coordinates. Thus if p and q are two points on the real line, then the distance between them is given by:

$$d(p, q) = |p - q|$$

$$d(p, q) = \sqrt{(p - q)^2}$$

3.2 Two dimensions

In the Euclidean plane, let point p have Cartesian coordinates *(p1, p2)* and let point q have coordinates *(q1, q2).*
Then the distance between p and q is given by,

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$$

3.3 Higher dimensions

In general, for points given by Cartersian coordinates in n-dimensional Euclidean space, the distance is,

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

3.4 Disadvantages

Euclidean distance is not scale in-varient which means that distance computed might be skewed depending on the units of the features. Typically, one needs to **normalize** the data before using this distance measure.
Moreover, as the dimensionality increases of your data, the less useful Euclidean distance becomes. This has to do with the curse of dimensionality which relates to the notion that higher-dimensional space does not act as we would, intuitively, expect from 2 or 3-dimensional space.

4 Cosine Similarity

It is a popular method for approximating how similar two vectors are. The intuition behind cosine similarity is relatively straightforward, we simply use the cosine of the angle between the two vectors to quantify how similar two vectors are.
From trigonometry we know that the

$$cos(0^\circ) = 1$$

$$cos(90^\circ) = 0 \quad and \quad 0 <= cos(\theta) <= 1.$$

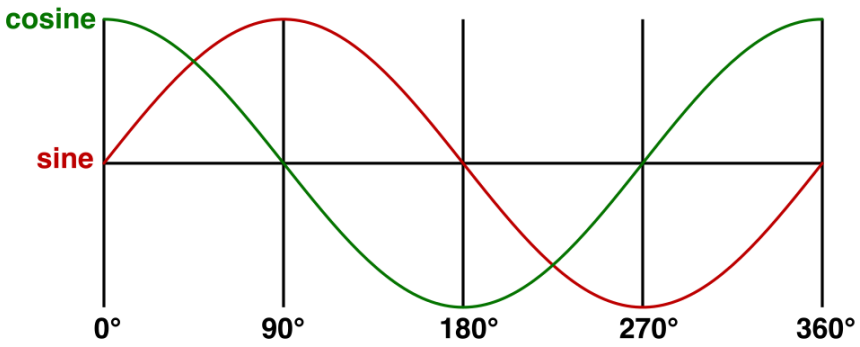

Figure 2: Sine and Cosine Wave

Figure 3: Trigonometry

The Dot Product of two Euclidean vectors **a and b** is defined by,

$$a \cdot b = \|a\| \, \|b\| \, cos\theta,$$

$$cos(\theta) = \frac{a \cdot b}{\|a\| \, \|b\|}$$

$$cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

Two vectors with exactly the same orientation have a cosine similarity of 1, whereas two vectors diametrically opposed to each other have a similarity of -1. Note that their magnitude is not of importance as this is a measure of orientation.

4.1 Disadvantage

One main disadvantage of cosine similarity is that the magnitude of vectors is not taken into account, merely their direction. In practice, this means that the differences in values are not fully taken into account. If you take a recommender system, for example, then the cosine similarity does not take into account the difference in rating scale between different users.

4.2 Usage

We use cosine similarity often when we have high-dimensional data and when the magnitude of the vectors is not of importance. For text analyses, this measure is quite frequently used when the data is represented by word counts. For example, when a word occurs more frequently in one document over another this does not necessarily mean that one document is more related to that word.

5 Reference

Why is Euclidean distance not a good metric in high dimensions?