Distance

Alex Sverdlov
alex@theparticle.com

1 Introduction

When we think of distance, we often think of straight line (or Euclidean) distance. There are many others, each applicable to various situations and domains. For example, when traveling, straight line distance is not as important as travel distance or travel time.

Distance is often used as proxy for *similarity*, which may be difficult to define.

2 Manhattan Distance

This is the distance that a taxi cab must travel between any two points in a grid city, and is measured as the sum of city blocks that need to be traveled. For any two points a and b,

$$d_1(a, b) = \sum_{i=1}^{n} |a_i - b_i|$$

3 Euclidean Distance

This is the familiar distance everyone learns about in school. For any two points a and b,

$$d_2(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

4 Chebyshev Distance

Also known as the chessboard distance, and represents the number of moves a king has to make to move between positions on the chess board. For any two points a and b,

$$d_{Chebyshev}(a, b) = \max_i (|a_i - b_i|)$$

In other words, maximum absolute difference in any coordinate.

5 p-norm Distance

Also known as Minkowski distance, is a generalization of manhattan & euclidian distance to higher powers. For any two points a and b,

$$d_p(a,b) = (\sum_{i=1}^{n} |a_i - b_i|^p)^{\frac{1}{p}}$$

Notice that $p = 1$ is Manhattan distance, $p = 2$ is Euclidean distance. And the limit as $p \to \text{inf}$ results in Chebyshev Distance.

6 Cosine Similarity

To compare vectors to each other, we often project one onto the other via an inner product (also known as *dot* product, represented here as $\langle\rangle$). For any two vectors a and b,

$$d_{cos}(a,b) = \frac{\langle a,b\rangle}{\|a\|\|b\|}$$

where $\|a\|$ is the magnitude of vector a. If a and b are unit length, we can omit division by the magnitude.

Note that the dot product is interpretable as the cosine of the angle between the two vectors. If both a and b are pointing in the same direction it will be 1. If they're orthogonal (at right angles to each other), the distance will be 0.

7 Correlation coefficient

Also known as Pearson correlation coefficient is a measure of linear relationship between random variables X and Y. For any two variables X and Y:

$$p_{XY} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

The values range from -1 to 1, where -1 means negatively correlated, and 1 means positively correlated. The value of 0 (or anything close to 0) implies no linear relationship: It does NOT mean the two variables are not related or are independent.

For sample data, the above is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

8 Kullback-Leibler divergence

To measure "distance" between probability distributions, we measure something called divergence, or relative entropy. For two discrete probability distributions P and Q:

$$D_{KL} = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

9 Problems

One of the problems with measuring similarity by way of distance is that it is too damn convenient. It's very easy to take X and Y and calculate a single number that we claim represents something significant about both X and Y. In fact, most machine learning methods rely on this!

The biggest problems involve finding the correct dimensions to measure. Only knowledge of the domain can determine that.

Once relevant dimensions are chosen, finding the correct scaling (or quantization) factors to use for each dimension.

After we have the feature vector, we can start selecting and experimenting with various ways of comparing them, hopefully something that makes sense semantically.

10 Conclusion

Distance measure is often the critical bit of any classification or clustering method. It is often overlooked—treated as unimportant. Pretty much every time you see anyone use Euclidean distance to compare feature vectors that represent raw business measurements, someone is being lazy (or ignorant of the business).

The measures mentioned in this paper are just a small sample of what is out there.