Similarity

Cosine Similarity

- If $X = Y$ (the two vectors have the same values)

$$\frac{X \cdot Y}{||X||\,||Y||} = \frac{X \cdot X}{||X||\,||X||} = \frac{x_1x_1 + \cdots + x_Nx_N}{\sqrt{x_1x_1 + \cdots + x_Nx_N} \cdot \sqrt{x_1x_1 + \cdots + x_Nx_N}} = 1$$

- If X is orthogonal to Y
  - $X = [-1, 3]$, $Y = [3, 1]$

$$\frac{X \cdot Y}{||X||\,||Y||} = \frac{(-1 \cdot 3) + (3 \cdot 1)}{||X||\,||Y||} = 0$$

- The length of vectors does not affect similarity

$$\frac{X \cdot Y}{||X||\,||Y||} = \frac{(5X) \cdot Y}{||\left(5X\right)||\,||Y||}$$

Euclidean Distance

- Euclidean distance is the length of the line connecting two points
- $d\left(X, Y\right) = \sqrt{\sum_i^N \left(x_i - y_i\right)^2}$
- Example:
  - $X = [1, 2]$, $Y = [3, 5]$
  - $d\left(X, Y\right) = \sqrt{\left(1-3\right)^2 + \left(2-5\right)^2}$
  - $d\left(X, Y\right) = \sqrt{13}$
- If the vectors are the same, $d\left(X, X\right) = 0$

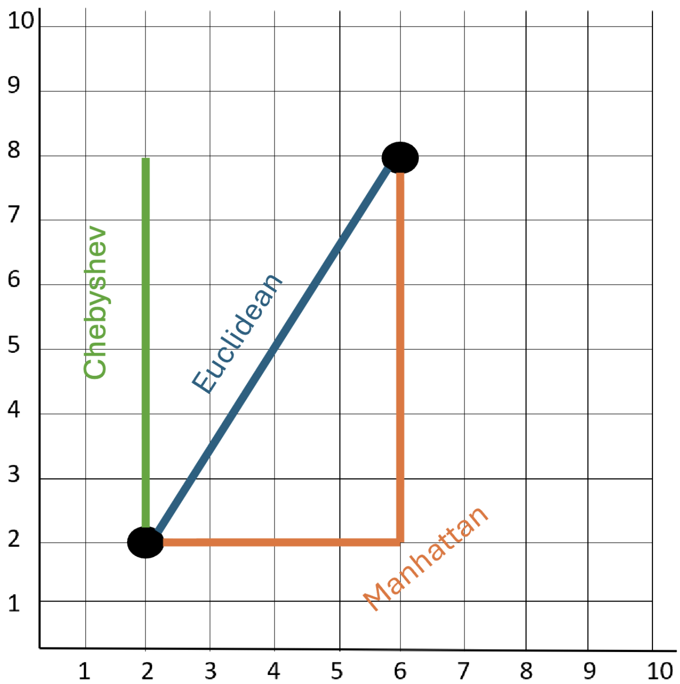
Manhattan Distance

- $d\left(X, Y\right) = \sum_i^N |x_i - y_i|$
- Also called $L1$ or taxicab distance
- Measured in straight lines along each axis
- The number of blocks that you need to drive to get from one point to another
- Example:
  - $X = [1, 2]$, $Y = [3, 5]$
  - $d\left(X, Y\right) = |1 - 3| + |2 - 5| = 5$
- If the vectors are the same, $d\left(X, X\right) = 0$

Minkowski Distance

- Euclidean and Manhattan distances are specific forms of the Minkowski distance
- $d(X, Y) = \left(\sum_{i}^{N} |x_i - y_i|^p\right)^{\frac{1}{p}}$
- If $p = 1$, we the Manhattan
- If $p = 2$, we get Euclidean
- If $p = \infty$, we get Chebyshev
  - $d(X, Y) = \max_i |x_i - y_i|$

Visualizing the Distances



Properties

- Euclidean distance and Manhattan distance both have the following properties
  - Symmetric: $d(X, Y) = d(Y, X)$
  - Positive
  - $d(X, Y) = 0$ if and only if the two vectors contain the same values
  - The triangle inequality holds for three points p, q, r. That is, $d(p, q) + d(q, r) \leq d(p, r)$

Binary Similarity – Jaccard Similarity Index

- If we have binary vectors (vectors only containing 1s and 0s), we can use specific similarities
- $J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$
- M_{11}: The number of times that both vectors have a 1

- M_{01}: The number of times that vector A has a 0 when vector B has a 1
- M_{10}: The number of times that vector A has a 1 when vector B has a 0
- M_{00}: The number of times that both vectors have a 0
- Example:
  - $A = [0, 0, 0, 1, 1, 1]$
  - $B = [0, 1, 0, 1, 0, 0]$
  - $M_{11} = 1$
  - $M_{01} = 1$
  - $M_{10} = 2$
  - $J = \frac{1}{1+2+1} = 0.25$