Distance measures

Reproduced from Jeffrey D. Ullman at Stanford

Axioms of a Distance Measure

- *d* is a *distance measure* if it is a function from pairs of points to real numbers such that:

* distribution $A \to B$

$\neq$

" $B \to A$

  1. $d(x,y) \geq 0$.
  2. $d(x,y) = 0$ iff $x = y$.
  3. $d(x,y) = d(y,x)$.
     - In fact, there are some asymmetric distance measures, so this constraint is not always required
  4. $d(x,y) \leq d(x,z) + d(z,y)$ (*triangle inequality*).

Euclidean Distance

- d(x,y) = square root of the sum of the squares of the differences between *x* and *y* in each dimension.
  - The most common notion of "distance."
- A.k.a., L_2 *norm*

Manhattan Distance

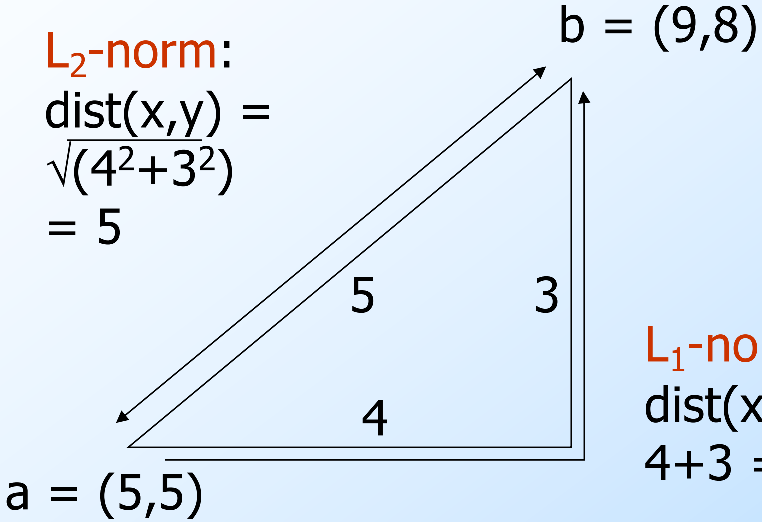
源自城市地图

- d(x,y) = sum of the differences in each dimension.
  - Distance if you had to travel along coordinates only.
- A.k.a., *L_1 norm*

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2|$$

Examples of Euclidean Distance and Manhattan Distance

L_2-norm:
$\text{dist}(x,y) = \sqrt{(4^2+3^2)} = 5$



L_1-norm:
$\text{dist}(x,y) = 4+3 = 7$

Other norms

- L_∞ *norm* : d(x,y) = the maximum of the differences between *x* and *y* in any dimension.
- Note: the maximum is the limit as *n* goes to ∞ of the L_n norm: what you get by taking the *n* th power of the differences, summing and taking the *n* th root.*

L_i-norm : $(x,y) = \sqrt[i]{|x_1-y_1|^i + |x_2-y_2|^i}$

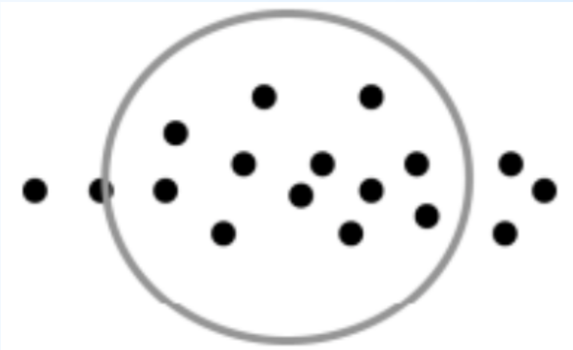
當 $r=\infty$ $(x,y)=\sqrt[\infty]{|x_1-y_1|^\infty+|x_2-y_2|^\infty}$

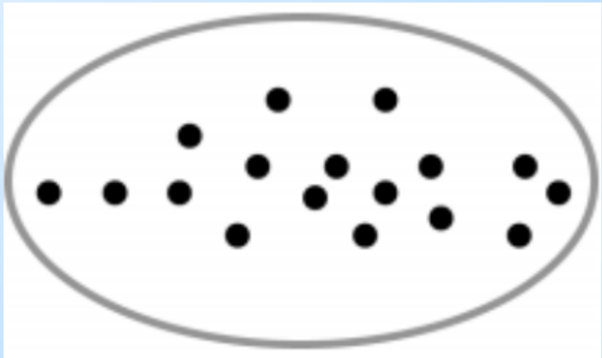
$=|x_1-y_1|$

看2組誰大

Mahalanobis distance

- A lot of times, data points do not form a circle shape
- We probably need to consider the variability of each dimension

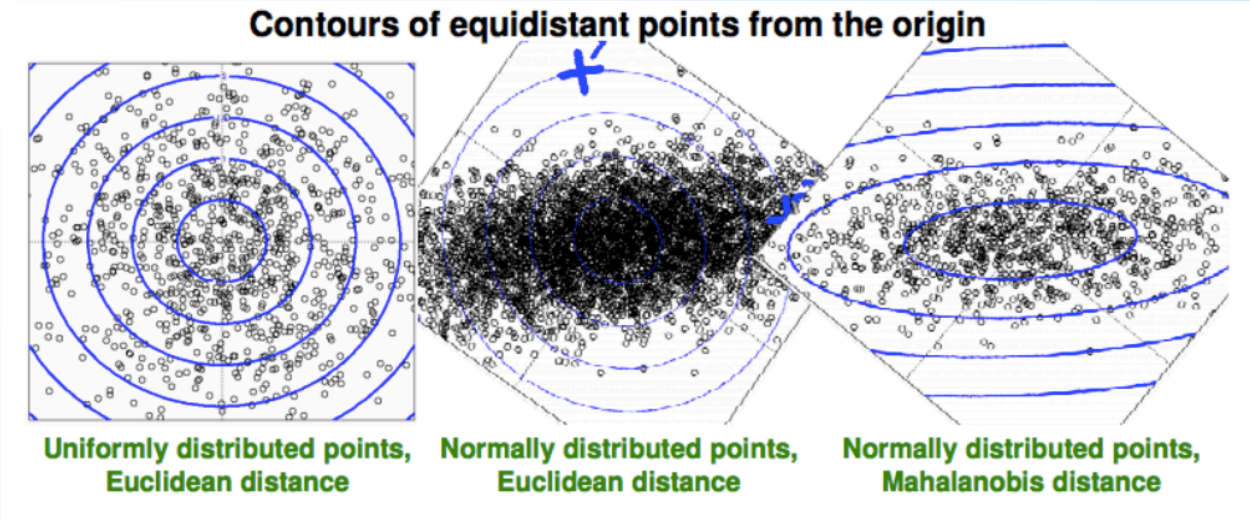




Mahalanobis distance

- $d(x, y) = \sqrt{\sum_{i=1}^{d} \frac{(x_i - y_i)^2}{s_i^2}}$
  - Take into account the variation of each dimension

Euclidean vs Mahalanobis distance



Source: J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets

Jaccard Distance for Sets (Bit-Vectors)

{a,c,d,e} {a,d,e}

- Example: p_1 = 10111; p_2 = 10011.
- Size of intersection = 3; size of union = 4, Jaccard similarity (not distance) = 3/4.
- d(x,y) = 1 – (Jaccard similarity) = 1/4.

$\frac{3}{4}$

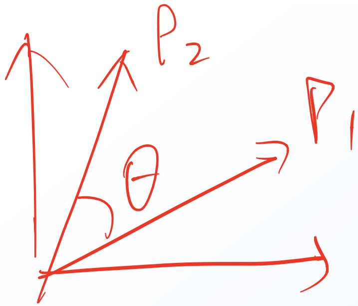
Why J.D. Is a Distance Measure

- $d(x,x) = 0$ because $x \cap x = x \cup x$.
- $d(x,y) = d(y,x)$ because union and intersection are symmetric.
- $d(x,y) \geq 0$ because $|x \cap y| \leq |x \cup y|$.
- $d(x,y) \leq d(x,z) + d(z,y)$ trickier – ignore the proof here

cosine similary

$A = [\,-1w,\ -2w\,]$

$B = [\,1w,\ 2w\,]$



$\frac{P_1 \cdot P_2}{|P_1||P_2|}$ $C = [101, 201]$

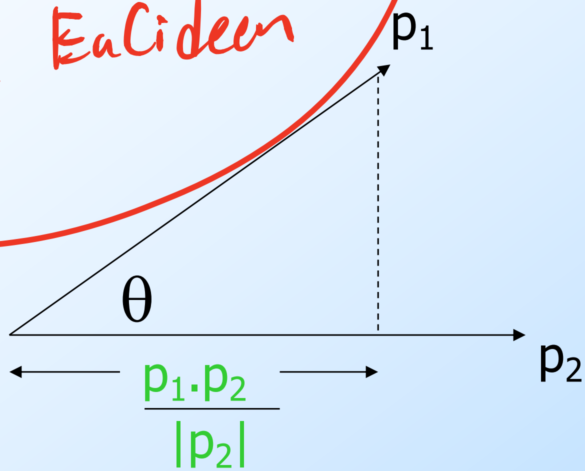
Cosine Distance

- Think of a point as a vector from the origin (0,0,…,0) to its location.
- Two points' vectors make an angle, whose cosine is the normalized dot-product of the vectors: $p_1.p_2/|p_2||p_1|$.
  - Example: $p_1 = 00111$; $p_2 = 10011$.
  - $p_1.p_2 = 2$; $|p_1| = |p_2| = \sqrt{3}$.
  - $\cos(\theta) = 2/3$; θ is about 48 degrees.

Cosine-Measure Diagram

* What situation ?

Cosine v.s. Euclidean



$$d(p_1, p_2) = \theta = \arccos(p_1 . p_2 / |p_2||p_1|)$$

Why C.D. Is a Distance Measure

- d(x,x) = 0 because arccos(1) = 0.
- d(x,y) = d(y,x) by symmetry.
- d(x,y) $\geq$ 0 because angles are chosen to be in the range 0 to 180 degrees.
- Triangle inequality: physical reasoning. If I rotate an angle from *x* to *z* and then from *z* to *y*, I can't rotate less than from *x* to *y*.

Edit Distance

- The *edit distance* of two strings is the number of inserts and deletes of characters needed to turn one into the other. Equivalently:
- $d(x,y) = |x| + |y| - 2|LCS(x,y)|$.
  - LCS = *longest common subsequence* = any longest string obtained both by deleting from *x* and deleting from *y*.

Example: LCS

- x = *abcde* ; y = *bcduve*.
- Turn x into y by deleting *a*, then inserting *u* and *v* after *d*.
  - Edit distance = 3.
- Or, LCS(x,y) = *bcde*.
- Note: |x| + |y| - 2|LCS(x,y)| = 5 + 6 –2*4 = 3 = edit distance.

Why Edit Distance Is a Distance Measure

- $d(x,x) = 0$ because 0 edits suffice.
- $d(x,y) = d(y,x)$ because insert/delete are inverses of each other.
- $d(x,y) \geq 0$: no notion of negative edits.
- Triangle inequality: changing *x* to *z* and then to *y* is one way to change *x* to *y*.

Variant Edit Distances

- Allow insert, delete, and *mutate*.
  - Change one character into another.
- Minimum number of inserts, deletes, and mutates also forms a distance measure.
- Ditto for any set of operations on strings.
  - Example: substring reversal OK for DNA sequences

Hamming Distance

- *Hamming distance* is the number of positions in which bit-vectors differ.
- Example: p_1 = 10101; p_2 = 10011.
- $d(p_1, p_2) = 2$ because the bit-vectors differ in the 3rd and 4th positions.

Why Hamming Distance Is a Distance Measure

- d(x,x) = 0 since no positions differ.
- d(x,y) = d(y,x) by symmetry of "different from."
- d(x,y) $\geq$ 0 since strings cannot differ in a negative number of positions.
- Triangle inequality: changing *x* to *z* and then to *y* is one way to change *x* to *y*.

Other distance measures

- Distance between two distributions
  - KL-divergence (a well-known asymmetric distance measure)
- Number of steps to move a king (in a chess game) from (x1, y1) to (x2, y2),
  - A king can move to any of it’s neighboring square
  - A.k.a., infinity norm, or Chebyshev distance
  - Distance = max(|x1-x2|, |y1-y2|)

Quiz

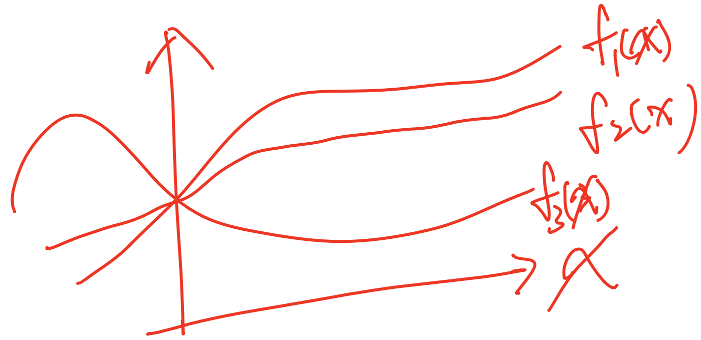
play the game of go

- Given an example in which Euclidean distance may be inapplicable or inappropriate
- How to define the distance between two sets (e.g., A=[1,2,3], B=[2,3,4], C=[5,6,7], S(A,B)=? S(A,C)=?) 用 jpart
- Doc1 has 100 word "w1" and 300 word "w2"; doc2 has 10 word "w1" and 30 word "w2", doc3 has 101 word "w1" and 200 word "w2"
  - Which doc is similar to doc1?

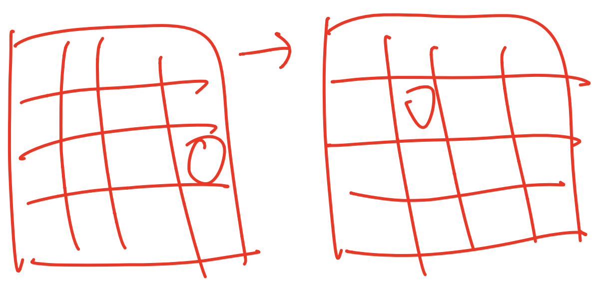
no mean of similar

no answer

* KL-divergence ex: find distribution



*



* Quidditch

depends on scen

cosine distance = 0
but in
eucld dis (d1, 3) > (d1, 2)