Introduction to Data Science

Distance

Arijit Mondal
Dept. of Computer Science & Engineering
Indian Institute of Technology Patna
arijit@iitp.ac.in

Measuring distance

- How to best measure the distance between points p and q in d-dimension?

Measuring distance

- How to best measure the distance between points p and q in d-dimension?
- The most obvious choice is Euclidean distance $d(p, q) = \sqrt{\sum_{i=1}^{d} (p_i - q_i)^2}$

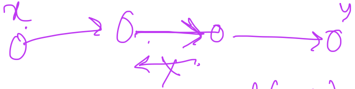
Measuring distance

- How to best measure the distance between points p and q in d-dimension?
- The most obvious choice is Euclidean distance $d(p,q) = \sqrt{\sum_{i=1}^{d}(p_i - q_i)^2}$
- Distance metric - distance measure needs to satisfy the following criteria
  - Positivity, $d(x,y) > 0$ ✓ if and only if
  - Identity, $d(x,y) = 0 \iff x = y$
  - Symmetric, $d(x,y) = d(y,x) \forall x, y$
  - Triangle inequality → $|a| + |b| \geq |c|$

Other type of metrics

- Not all measures are distance metric
- Example
  - Correlation coefficient → $(-1, 1)$ $\underline{x, y}$
  - Cosine similarity →
  - Travel time in a directed network
  - Cheapest airfare

$$\cos(p, q) = \frac{p \cdot q}{|p|\,|q|} \qquad -1, 1$$



$$d(x, y) \neq d(y, x)$$

Distance metric

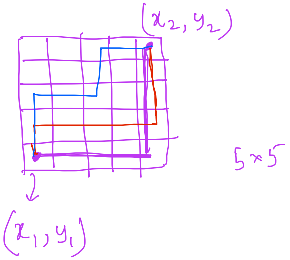
- Generic distance metric is defined as $d_k(p, q) = \sqrt[k]{\sum_{i=1}^{d} |p_i - q_i|^k}$

  d_2

  - Parameter k provides a way to trade off between the longest and the total dimensional differences
  - k can vary between 1 and ∞

Distance metric

- Generic distance metric is defined as $d_k(p, q) = \sqrt[k]{\sum_{i=1}^{d} |p_i - q_i|^k}$ ← L_k L^k
  - Parameter k provides a way to trade off between the longest and the total dimensional differences
  - k can vary between 1 and ∞
- L_1 — Manhattan distance



$L_1 = |(x_2 - x_1)| + |(y_2 - y_1)|$

Distance metric

- Generic distance metric is defined as $d_k(p, q) = \sqrt[k]{\sum_{i=1}^{d} |p_i - q_i|^k}$ ✓ $q \equiv$ origin
  - Parameter k provides a way to trade off between the longest and the total dimensional differences
  - k can vary between 1 and ∞
- L_1 — Manhattan distance
- L_2 — Euclidean distance

$\boxed{p}$ ✗ p_i^k ✗

$L_\infty = \max_i \{p_i\}$ ✓

Distance metric

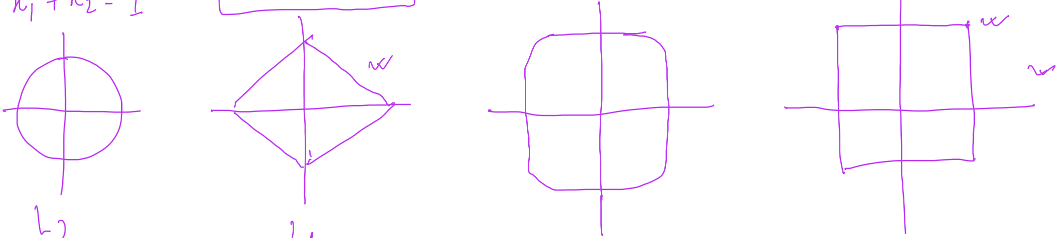
- Generic distance metric is defined as $d_k(p, q) = \sqrt[k]{\sum_{i=1}^{d} |p_i - q_i|^k}$
  - Parameter k provides a way to trade off between the longest and the total dimensional differences
  - k can vary between 1 and ∞
- L_1 — Manhattan distance
- L_2 — Euclidean distance
- L_∞ — Maximum component

Shape of equal distance

- L_1, L_2, L_5, L_∞

$x_1^2 + x_2^2 = 1$

$\boxed{|x_1| + |x_2| = 1}$



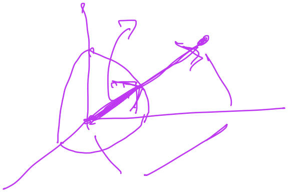
L_2 L_1 L_∞

$\max(|x_1|, |x_2|)$

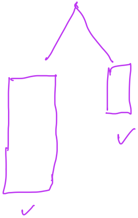
$$\sqrt[k]{\sum_{i=1}^{d} c_i |p_i - q_i|^k} \rightarrow z$$

Point vs Vector

- Vectors are usually a point in unit sphere, it provides only direction
- Norms
- Cosine similarity — $cos(p, q) = \frac{p \cdot q}{|p| \cdot |q|}$
- Cosine distance — $(1 - |cos(p, q)|)$ (triangle inequality does not hold)
- Angular distance — $d(p, q) = 1 - \frac{cos^{-1}(cos(p, q))}{\pi}$



Political, Sports, . . . –



CS244

Distance between probability distribution

- This is based on information theoretic notion of entropy
  - It measures uncertainty for the value of a sample drawn from the distribution
- Entropy — $H(P) = \sum_i p_i \log(1/p_i)$

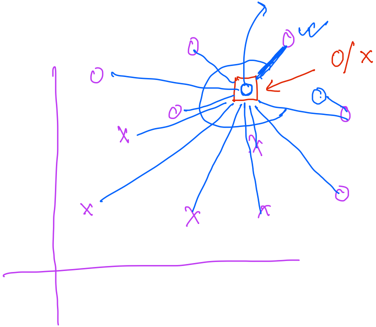
Distance between probability distribution

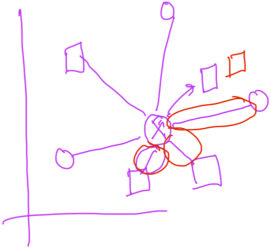
- This is based on information theoretic notion of entropy
  - It measures uncertainty for the value of a sample drawn from the distribution
- Entropy — $H(P) = \sum_i p_i \log(1/p_i)$
- Standard distance measure for probability distributions is KL-divergence (Kullbach-Leibler)
  $KL(P||Q) = \sum_i p_i \log_2(p_i/q_i)$ $\quad q_i = p_i$
- KL-divergence is not symmetric

Distance between probability distribution

- This is based on information theoretic notion of entropy
  - It measures uncertainty for the value of a sample drawn from the distribution
- Entropy — $H(P) = \sum_i p_i \log(1/p_i)$
- Standard distance measure for probability distributions is KL-divergence (Kullbach-Leibler) $KL(P||Q) = \sum_i p_i \log_2(p_i/q_i)$
- KL-divergence is not symmetric
- Jensen Shannon divergence metric — $JS(P,Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M)$ where $m_i = (p_i + q_i)/2$
- $\sqrt{JS(P,Q)}$ is a distance metric

Nearest neighbor

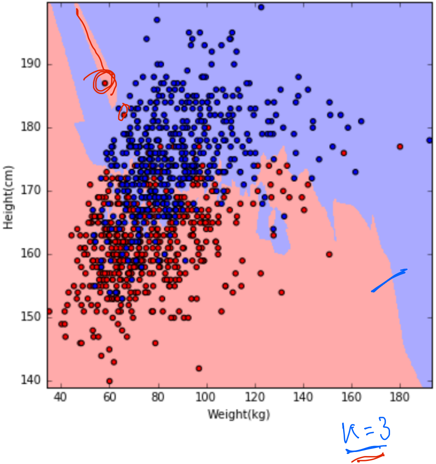




Nearest neighbor

- Simple, interpretable, non-linear
- Example — categorization of books, movies, cricketers, music, etc.

k-nearest neighbor



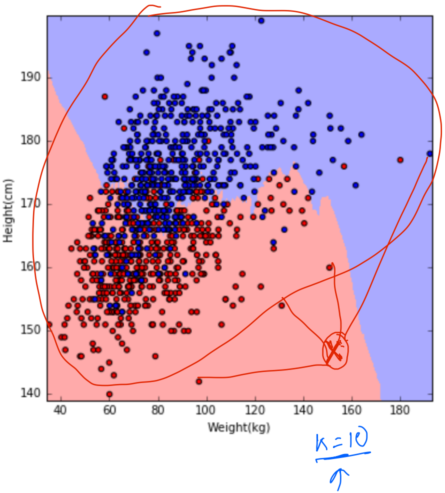


image source: Data Science Design Manual

Finding nearest neighbor