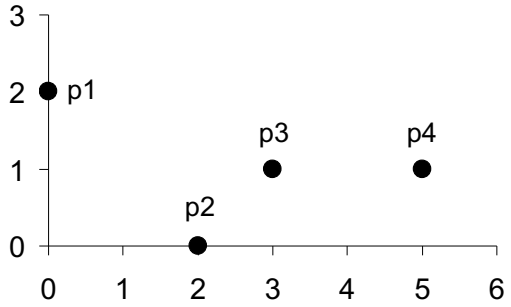
Feature spaces and proximity measures

- Famous example: Euclidean vector space *E*=(*Dom*, *dist*)
  - (*Dom*, *dist*) is a metric space
  - $Dom = \mathbb{R}^d$
  - $dist(x, y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$
- Notation:
  - Euclidean vector space =: "Feature space"
  - Vectors (Objects in the Euclidean feature space) =: "Feature vectors"
  - The *d* dimensions of the vector space =: "Features"
- Standardization is necessary, if scales differ!
  - e.g., age (e.g., range [0-100] vs salary (e.g., range: 10000-100000))

We will come back to this in a few slides

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Point coordinates



	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance matrix

Feature spaces and proximity measures

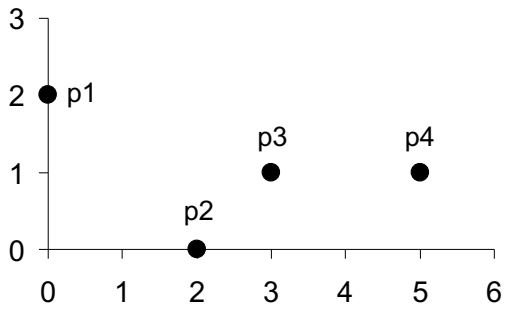
- Manhattan distance or City-block distance (L_1 norm)
  - $dist_1 = |p_1-q_1|+|p_2-q_2|+...+|p_d-q_d|$
  - The sum of the absolute differences of the p,q coordinates
- Euclidean distance (L_2 norm)
  - $dist_2 = ((p_1-q_1)^2+(p_2-q_2)^2+...+(p_d-q_d)^2)^{1/2}$
  - The length of the line segment connecting p and q
- Supremum distance (L_{max} norm or L_∞ norm)
  - $dist_\infty = \max\{|p_1-q_1|, |p_2-q_2|,...,|p_d-q_d|\}$
  - The max difference between any attributes of the objects.
- Minkowski Distance (Generalization of L_p-distance)
  - $dist_p = (|p_1-q_1|^p + |p_2-q_2|^p + ...+ |p_d-q_d|^p)^{1/p}$

Feature spaces and proximity measures

- Example

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Point coordinates



L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L1 distance matrix

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L2 distance matrix

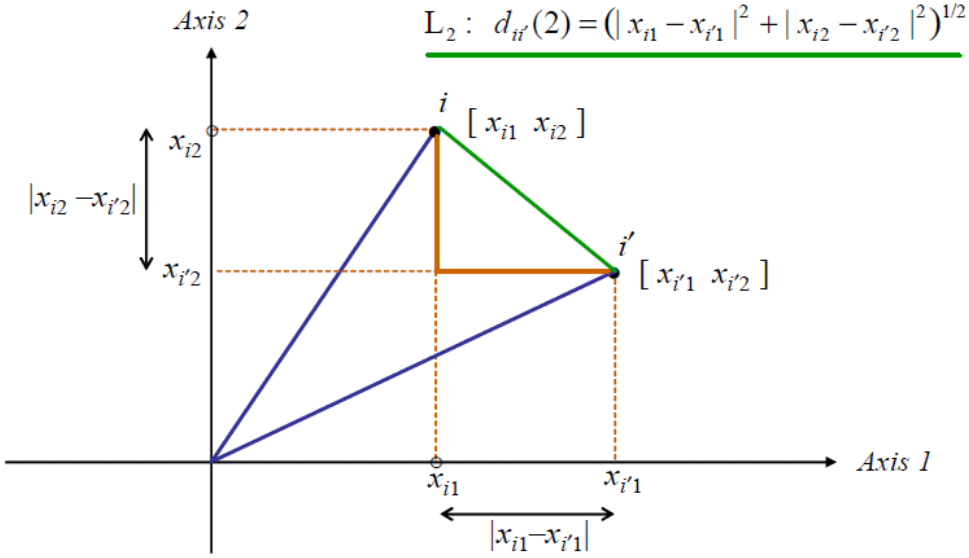
L	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

L_∞ distance matrix

Feature spaces and proximity measures

$$L_1: \; d_{ii'}(1) = |x_{i1} - x_{i'1}| + |x_{i2} - x_{i'2}|$$

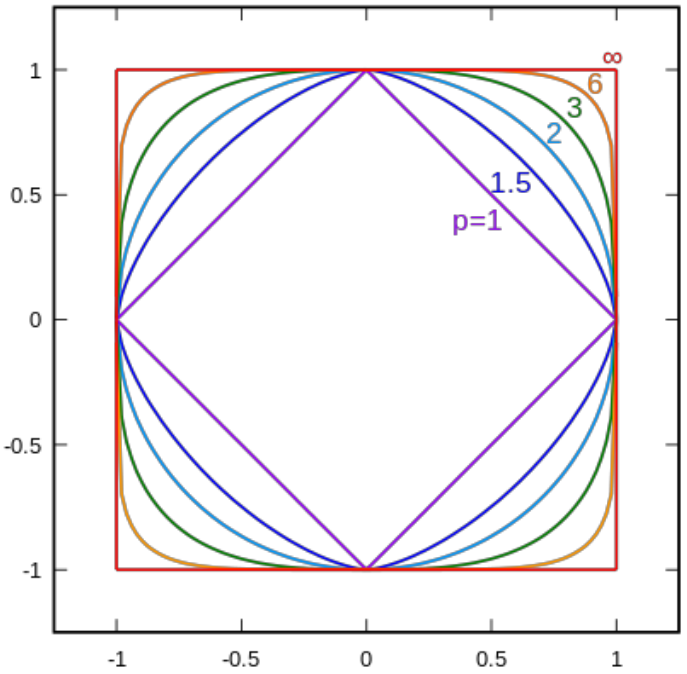
$$L_2: \; d_{ii'}(2) = (|x_{i1} - x_{i'1}|^2 + |x_{i2} - x_{i'2}|^2)^{1/2}$$



Source:http://www.econ.upf.edu/~michael/stanford/maeb5.pdf

Feature spaces and proximity measures

- Let *x,y* in [-1,1]
- For L1 norm
  - $|(x,y)|_1=1$ => x+y=1
  - If x=1, y=0
  - If x=0.8, y=0.2
  - …
- For L2 norm
  - $(x^2+y^2)^{1/2}=1$
  - It is circle
- …



Unit Circle for different Lp-distances

Source:https://de.wikipedia.org/wiki/P-Norm

Normalization

- Attributes with large ranges outweigh ones with small ranges
  - e.g. income [10K-100K]; age [10-100]
- To balance the "contribution" of an attribute *A* in the resulting distance, the attributes are scaled to fall within a small, specified range.
- min-max normalization: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - e.g. normalize age=30 in [0-1], when min=10,max=100. new_age=((30-10)/(100-10))*(1-0)+0=2/9
- z-score normalization also called zero-mean normalization
  - After zero-mean normalizing each feature will have a mean value of 0

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

e.g. normalize 70,000 iff μ=50,000, σ=15,000.
new_value = (70,000-50,000)/15,000=1.33