Dist & Similarity

8/21

2 types:
Numerical
Categorical

Numeric

- integers, reals, complex

E.g.

Age: domain(Age) = $\mathbb{N}$ ← natural #'s (non-neg ints)
Height: domain(Height) = $\mathbb{R}^+$ ← positive real #'s

Subtypes:

discrete: finite (or countably inf)

continuous: they take a real value

binary: take value $\{0, 1\}$

Categorical

- values come from a set of symbols
  Domain(Education) = {Highschool, BS, MS, PhD}
  Domain(Vehical type) = {car, truck, suv}

Subtypes

- **nominal**: unordered (only equality tests make sense)
- **ordinal**: values are ordered (equality and inequality)

Data matrix P (assume all values are numeric)

$\Rightarrow$ each row can be thought of as d-dim point

e.g. $x_i = (x_{i1}, x_{i2}, \cdots, x_{id}) \in \mathbb{R}^d$

↳ often as col vector $x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix}$

Let e_j be the j^{th} std basis vector

↳ $(000\cdots010\cdots0)$

↑ all 0s except j^{th} entry is 1

$$x_i = x_{i1} e_1 + x_{i2} e_2 + \cdots + x_{id} e_d = \sum_{j=1}^{d} x_{ij} e_j$$

5.9	3.0	4.2	1.5	Iris-versicolor
6.9	3.1	4.9	1.5	Iris-versicolor
6.6	2.9	4.6	1.3	Iris-versicolor
4.6	3.2	1.4	0.2	Iris-setosa
6.0	2.2	4.0	1.0	Iris-versicolor
4.7	3.2	1.3	0.2	Iris-setosa
6.5	3.0	5.8	2.2	Iris-virginica
5.8	2.7	5.1	1.9	Iris-virginica

consider this row

$x_2 = (6.9 \quad 3.1 \quad 4.9 \quad 1.5 \quad \text{Iris-versicolor})$

"project" to first 3 attribs

$\hat{x}_2 = (6.9 \quad 3.1 \quad 4.9) \in \mathbb{R}^3$ (x_{21} x_{22} x_{23})

$= 6.9 \begin{pmatrix}1\\0\\0\end{pmatrix} + 3.1 \begin{pmatrix}0\\1\\0\end{pmatrix} + 4.9 \begin{pmatrix}0\\0\\1\end{pmatrix} = \sum_{j=1}^{3} x_{2j} \cdot e_j$

e_1 e_2 e_3

simly for cols vectors $\mathbb{R}^n$

$x_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$

as d col vecs n

$\begin{pmatrix} | & | \\ x_1 & x_2 \cdots x_d \\ | & | \end{pmatrix}$ $x_j \in \mathbb{R}^n$

$D = \begin{pmatrix} x_{11} & x_{12} \cdots x_{1d} \\ x_{21} & x_{22} \cdots x_{2d} \\ \vdots & \ddots \\ x_{n1} & x_{n2} \cdots x_{nd} \end{pmatrix}$ an elt of $\mathbb{R}^{n \times d}$ $=$

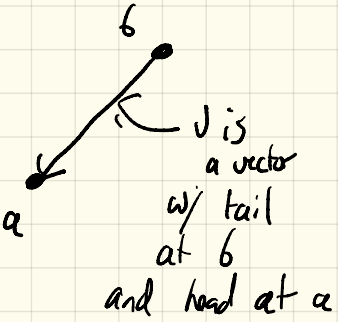
as n row vectors $x_i^T \in \mathbb{R}^d$

$\begin{pmatrix} — & x_1^T & — \\ — & x_2^T & — \\ & \vdots & \\ — & x_n^T & — \end{pmatrix}$ $=$ $\begin{pmatrix} | & | \\ x_1 & x_2 \cdots x_d \\ | & | \end{pmatrix}$

Dist & Angles

Let $a, b \in \mathbb{R}^m$

$$a = (a_1 \; a_2 \; \cdots \; a_m)^T$$
$$b = (b_1 \; b_2 \; \cdots \; b_m)^T$$



Subtract points $\quad a - b = \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ \vdots \\ a_m - b_m \end{pmatrix} = v$

dot product

$$a \cdot b = \sum_{i=1}^{m} a_i b_i$$

$p \in \mathbb{R}$ w/ $p \geq 1$

L_p-norm $\quad \|a\|_p = \left(\sum_{i=1}^{m} |a_i|^p \right)^{1/p}$

L_∞-norm $\quad \|a\|_\infty = \max_i a_i$

L_p-dist $\quad \|a - b\|_p = \left(\sum_{i=1}^{m} |a_i - b_i|^p \right)^{1/p}$

Note: $\|a\| = \|a\|_2$

$p = 2$ is Euclidean dist

Interpret geometrically

Consider points at dist 1 from origin

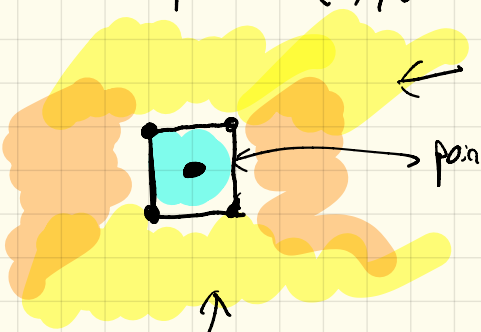
case $p=2$ find all points $(x,y) \in \mathbb{R}^2$ dist 1 from origin

$$\sqrt{(x-0)^2+(y-0)^2} = 1$$

$$\Leftrightarrow (x-0)^2 + (y-0)^2 = 1^2$$

circle is all points w/ L_2 dist $= 1$

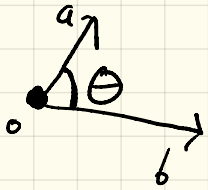
$p=\infty$ all points $(x,y) \in \mathbb{R}^2$ $\|(x,y)\|_\infty = 1$



Cos sim

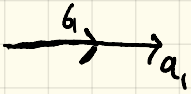
$$\frac{a \cdot b}{\|a\| \|b\|}$$

geometrically

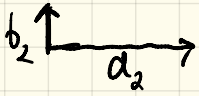


$$a \cdot b = \|a\| \|b\| \cos\theta$$

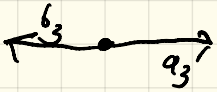
cos sim is cos of angle between 2 vecs



cos sim a_1 and b_1 1

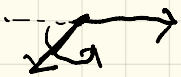


cos sim a_2 & b_2 is 0



cos sim a_3 & b_3 is -1

cos sim monotonically decrease from 1 to -1



cos sim monotonically increase from -1 to 1

orthogonal

iff $a \cdot b = 0 \iff \cos\theta = 0$