Linear Algebra

CSCI 2820

Lecture 5

Prof. Alexandra Kolla

Alexandra.Kolla@Colorado.edu
ECES 122

Today

- Distance
- Standard Deviation
- Angles (if time)
- Examples

Refresher on Norms/Distance

- Suppose that |x> and |y> are Boolean n-vectors, which means that each of their entries is either 0 or 1. What is their distance || |x> - |y>|| ?

$|x\rangle = (1, 0, 0, 1)$ Q: $\| |x\rangle - |y\rangle \| = ?$

$|y\rangle = (0, 1, 1, 1)$

$\| |x\rangle - |y\rangle \|^2 = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2 + (x_4-y_4)^2}^{\,2}$

$= \sqrt{1^2 + 1^2 + 1^2 + 0}^{\,2} = 3$

Refresher on Norms/Distance

$$\mathbf{rms}(x) = \sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}} = \boxed{\frac{\|x\|_2}{\sqrt{n}}}.$$

- Is RMS(*) a norm?

↳ Does it satisfy all 4 of norm conditions

YES! by def (almost)

Distance, examples

- How similar are two text documents?

$|x\rangle$, $|y\rangle$ with entries the frequencies of words

$\| |x\rangle - |y\rangle \|$ represents how similar/different they are

- Nearest neighbor

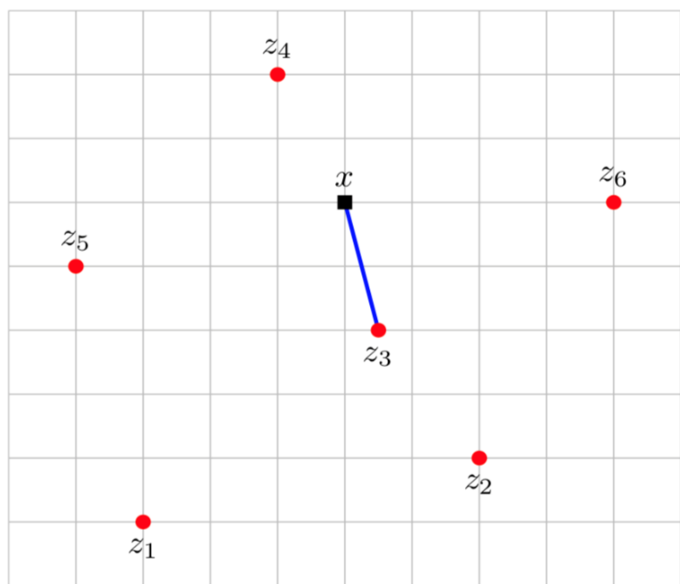
$\vec{z}_1, \dots, \vec{z}_m$ n-vectors

$\vec{x}$.

we say that $\vec{z}_j$ is nearest neighbor of $\vec{x}$

if $\|\vec{x} - \vec{z}_j\| \le \|\vec{x} - \vec{z}_i\|$, $i = 1, \dots, m$

Distance,examples



Distance, examples

	Veterans Day	Memorial Day	Academy Awards	Golden Globe Awards	Super Bowl
Veterans Day	0	0.095	0.130	0.153	0.170
Memorial Day	0.095	0	0.122	0.147	0.164
Academy A.	0.130	0.122	0	0.108	0.164
Golden Globe A.	0.153	0.147	0.108	0	0.181
Super Bowl	0.170	0.164	0.164	0.181	0

Q: What can we say about ||x-y|| vs ||y-x|| from this table?

What units to choose?

$\|\vec{x}-\vec{y}\|^2 = (x_1-y_1)^2 + \cdots + (x_n-y_n)^2$

$\vec{x}$ feature vector, each entry might have different units

eg (sq.ft, # bedrooms)

change to thousands of sq.ft

$\vec{x} = (1400, 2)$, $\vec{y} = (1500, 2)$, $\vec{z} = (1400, 4)$

$\|\vec{x}-\vec{y}\| > \|\vec{x}-\vec{z}\|$

Standard Deviation

n- vector $\vec{x}$

Def: de-meaned vector $\tilde{x} = \vec{x} - \text{avg}(\vec{x})\vec{1}$

$\text{avg}(\vec{x}) = \frac{x_1 + \cdots + x_n}{n} = \langle \frac{\vec{1}}{n} | \vec{x} \rangle$

$\text{avg}(\tilde{x}) = 0$

Q. when is $\tilde{x} = \vec{0}$, $\vec{x} = (\alpha, \alpha, \cdots, \alpha)$

$\text{avg}(\vec{x}) = \alpha$

$\tilde{x} = (\alpha, \alpha, \cdots, \alpha) - \alpha \cdot (1, \cdots, 1)$

$= (\alpha, \alpha, \cdots, \alpha) - (\alpha, \cdots, \alpha) = (0, \cdots, 0)$

$$\boxed{\text{std}(\vec{x}) = \sqrt{\frac{(x_1 - \text{avg}(\vec{x}))^2 + \cdots + (x_n - \text{avg}(\vec{x}))^2}{n}}}$$

$\rightarrow \text{rms}(\tilde{x}) \stackrel{\text{def}}{=} \frac{\|\tilde{x}\|}{\sqrt{n}}$

Q: when is $\text{std}(\vec{x}) = 0$?

$\therefore \quad \text{std}(\vec{x}) = \frac{\| \vec{x} - \frac{\langle 1, x \rangle}{n} \cdot \vec{1} \|}{\sqrt{n}}$

Standard Deviation

$$\vec{x} = (1, -2, 3, 2)$$

$$avg(\vec{x}) = \frac{1-2+3+2}{4} = 1$$

$$\tilde{x} = \vec{x} - avg(\vec{x})\vec{1} = (0, -3, 2, 1)$$

$$std(\vec{x}) = \frac{\|\tilde{x}\|}{\sqrt{4}} = \sqrt{\frac{9+4+1}{4}} = \sqrt{\frac{14}{4}} = 1.872$$

notation (mainly probability)

$$\mu = \frac{\langle 1, x \rangle}{n}, \quad \sigma = \frac{\|\vec{x} - \mu\vec{1}\|}{\sqrt{n}}, \quad \sigma^2 \sim E(x-\mu)^2$$

$\parallel$ $avg(\vec{x})$ $E(x)$

$\parallel$ $std(\vec{x})$

Standard Deviation

Claim: $\operatorname{rms}(\vec{x})^2 = \operatorname{avg}(\vec{x})^2 + \operatorname{std}(\vec{x})^2$

$$\Rightarrow \operatorname{std}^2(\vec{x}) = \operatorname{rms}(\vec{x})^2 - \operatorname{avg}(x)^2$$

$$= \frac{1}{n}\left\|\vec{x} - \frac{\langle 1,x\rangle}{n}\vec{1}\right\|^2 = \frac{1}{n}\left(\vec{x} - \frac{\langle 1,x\rangle}{n}\vec{1}\right)^T\left(\vec{x} - \frac{\langle 1,x\rangle}{n}\vec{1}\right)$$

$$= \frac{1}{n}\left(\langle x,x\rangle - 2\frac{\langle 1,x\rangle}{n}\langle x,1\rangle + \left(\frac{\langle 1,x\rangle}{n}\right)^2\langle 1,1\rangle\right)$$

$$= \frac{1}{n}\left[\langle x,x\rangle - \frac{2}{n}\langle 1,x\rangle^2 + n\left(\frac{\langle x,1\rangle}{n}\right)^2\right]$$

$$= \underbrace{\frac{1}{n}\langle x,x\rangle}_{\operatorname{rms}(\vec{x})^2} - \underbrace{\left(\frac{\langle 1,x\rangle}{n}\right)^2}_{(\operatorname{avg}(\vec{x}))^2}$$

$$\left(\frac{\|\vec{x}\|}{\sqrt{n}}\right)^2$$

Standard Deviation examples

n-vector $\vec{x}$ time series of daily avg. temp

$avg(\vec{x})$ = average temp of location

$std(\vec{x})$ = "how much does temp varies from avg?"

Standard Deviation/Chebyshev

Assume k is the number of entries of a vector $\vec{x}$ that satisfy $|x_i - \text{avg}(\vec{x})| \geq a$

then

$$\text{std}(\vec{x})^2 = \frac{(x_1-\text{avg}(\vec{x}))^2+\cdots+(x_n-\text{avg}(\vec{x}))^2}{n}$$

$$\geq \frac{k \cdot a^2}{n}$$

$$\Rightarrow \frac{k}{n} \leq \left(\frac{\text{std}(\vec{x})}{a}\right)^2,$$

how big can k be if k entries have: $|x_i - \text{avg}(\vec{x})| > 3\,\text{std}(\vec{x})$?

$$\frac{k}{n} \leq \left(\frac{\text{std}(\vec{x})}{3\,\text{std}(\vec{x})}\right)^2 \Rightarrow k/n \leq 1/3^2 = 1/9$$

Q: How many entries of the vector $\vec{x}$ can deviate from the mean $\text{avg}(\vec{x})$ by ≥ 3 standrad deviations?

in general, no more than $1/w^2$ fraction of entries can deviate more than $w\,\text{std}(\vec{x})$ from $\text{avg}(\vec{x})$

Standard Deviation/Chebyshev

Properties: (1) Adding constant doesn't matter: for any $\vec{x}$, any α:

$$\text{std}(\vec{x} + \alpha\vec{1}) = \text{std}(\vec{x})$$

(2) Multiply by a scalar.

any $\vec{x}$, α. $\quad \text{std}(\alpha\vec{x}) = |\alpha|\,\text{std}(\vec{x})$

- Standardization: $\tilde{x} = \vec{x} - \text{avg}(\vec{x})\vec{1}$

standardized version of vector $\vec{x}$:

$$\vec{z} = \frac{(\vec{x} - \text{avg}(\vec{x})\vec{1})}{\text{std}(\vec{x})}$$