DS2500
2/7-Fri :)

Admin

- HW3 due 9pm
- exam #1 $\boxed{2/14}$
- mini presos
  - 2/10 signups
  - 2/23 slides due
  - 2/24, 2/25 prez
- HW2 graded!

Agenda

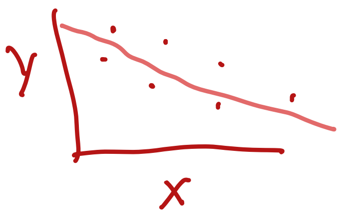
1. normalizing data
2. variance, std dev
3. Python

1. Normalizing Data

↳ tuesday: correlation, linear regression } adm rate, tuition

adm rate → x (ind)
tuition → y (dep)

- given x, what is y?
- given a new x, what would y be?



conc: adm rate, tuition highly correlated

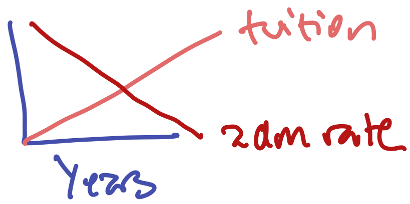
↳ could end up here if values up and down

↳ but both are consistently ↑ or ↓ so they may also correlate with time

Another ~ time series data

- sequence of observations assoc. with points in time
  ↳ successive, equally spaced
- time series analysis: extracting predictions and filling in gaps

Expectation



time: ind variable
tuition, adm: dep variable

↳ But! :·· O

peryear: tuition, adm rates } features

2014 vs. 2015?
2020 vs. other year?
2008 vs. 2024
or
2020 vs. 2024
{euclidean distance} !

Euclidean distance assumes that: unit change in any direction is equally significant

ex 2014 vs. 2015

	2014	2015
Adm	32.22	28.48
tuition	60290	62800

$$\delta(2014, 2015) = \sqrt{(x-x')^2 + (t-t')^2} = 2510$$

2014 ✓ 2015

Drop in adm rate? 2014 vs. 2016

adm rate = 18
tuition = 62800

$$\sqrt{(32.22-18)^2 + (60290-62800)^2}$$

2 510.04

2014 vs. 2016

Rise in tuition? 2014 vs. 2016

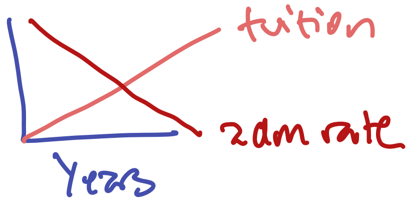
adm rate = 28.48
tuition = 78 202

17 912

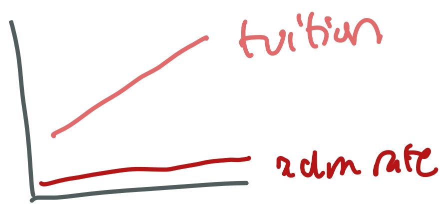
2014 vs. 2016

problem #1: one feature has huge impact over the other

expectation



→ problem #2: communication



To solve:
tuition: 50,000 to 78,000
adm : 8 to 33

Solution → normalization!
- max = 1
- min = 0

→ smallest val in population: 0
largest val in pop : 1

for every x_i ... x_{max}, x_{min}

everything else scaled

$$x_{scaled} = \boxed{\frac{x_i - x_{min}}{x_{max} - x_{min}}}$$

x_i vs x_{max}

x_i vs. x_{min}

↳ feature scaling

ex pop = 1, 2, 3, 4, 5

min 1

max 5

$$\frac{3-1}{5-1} = \frac{2}{4} = .5$$

min/max normalization:

↳ 0 to 1 range

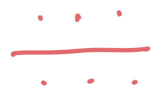
downside: no units

suppress outliers

2. Variance + Std Deviation

↳ How different from the mean?

low var



high var

- variance: avg squared difference from mean σ^2

  used for: comparing 2 datasets (end up w/bigger #s)

- std dev: avg difference from mean σ

  used for: one dataset

  (keeps units you started with)

$$\boxed{\sigma^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2}$$

$\sigma = \sqrt{\sigma^2}$

(basically a mean, but we divide by $N-1$)

(Bessel's, to correct for not having every data point)

lst2 = [-12, 30, 50, -55, -75, 80]

σ^2(lst2) $\bar{x} = 3$

3708

variance: lst 2 varies a lot more than lst 1

lst1 = [-1, 3, 5, -3, 6, 8]

σ^2(lst1) $\bar{x} = 3$

18

std dev

σ(lst2) = 60.89

↳ on avg, how far from mean

σ(lst1) = 4.24

↳ on avg, how far from mean

In Python: statistics.variance

Outlier: > 2 std devs from mean