Lesson Plan 3/8

Thursday, March 8, 2018 3:11 PM

Admin
- Grades are coming, we swear…

HDSC
- Visual explanation of stats and probability: http://students.brown.edu/seeing-theory/

Project 1
- Questions/comments?
- Operating on all pairs of rows from two tables

Data transformations

Handling missing data

Example:

- dates: difference in days
- numeric: $d = |x_i - x_j|$

$s = e^{-d}$

- cat similarity $s = 0$ if $v_1 = v_2$

$s(v_1, v_2) = \begin{cases} \\ 0 \quad \text{o.w} \end{cases}$

Similarity matrix

S_{ij} : similarity (i, j)

① Transform so that it satisfies

- central tendency
- spread

Can we perform transformations

- C.E. = 0
- spread = 1

Removes units

-> Standardization

Given data $x_1, \ldots, x_n$

$\mu = \bar{x}$ [illegible]

$$z_i = \frac{x_i - \bar{x}}{sd(x)}$$ ← data units / ← data units

Q1) $\bar{z}$:

$sd(z)$:

$$\bar{z} = \frac{1}{n}\sum_i z_i \qquad \text{(def'n)}$$

$$= \frac{1}{n}\sum_i \left(\frac{x_i - \bar{x}}{sd(x)}\right) \qquad \text{(plug-in)}$$

$$= \frac{1}{n}\sum_i \frac{x_i}{sd(x)} - \frac{1}{n}\sum_i \left(\frac{\bar{x}}{sd(x)}\right)$$

$$= \frac{1}{sd(x)} \left[\frac{1}{n} \sum_i x_i - \frac{1}{n} \sum_i \bar{x} \right]$$

$$= \frac{1}{sd(x)} \left\{ \bar{x} - \frac{1}{n} \cdot (n\bar{x}) \right\}$$

$$= 0$$

$sd(z) = 1$

Centering : $z_i = (x_i - \bar{x})$ $\bar{z}$? $sd(z)$?

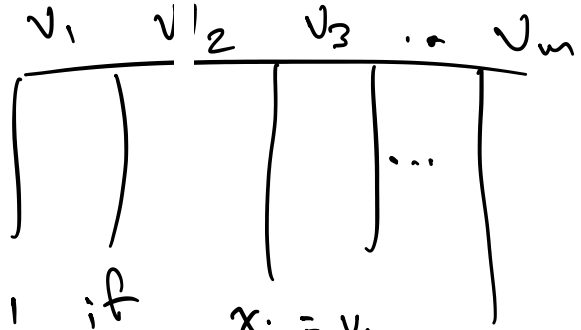
Scaling : $z_i = \frac{x_i}{sd(x)}$ $\bar{z}$? $sd(z)$?

Categorical → numeric

One-hot-encoding

→ Given categorical variable x can take values $\{v_1, v_2, \ldots, v_m\}$

⟹



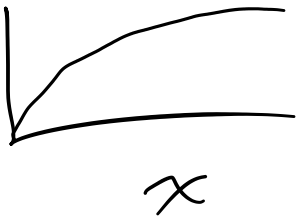
$$v_{i1} = \begin{cases} 1 & \text{if } x_i = v_1 \end{cases}$$

$$v_{ij} = \begin{cases} 1 & \text{if } x_i = v_j \\ 0 & \text{o.w.} \end{cases}$$

Reducing data skew with logarithmic transform

$$z_i = \log(x_i)$$

Only works if $x_i > c$



shifted log

$$z_i = \log(x_i - \min(x_i))$$

signed ~~shifted~~ log transform

$$z_i = \text{sign}(x_i) * \log(|x_i| + 1)$$

$$\text{sign}(x_i) = \begin{cases} +1 & \text{if } x_i > 0 \\ 0 & \text{if } x_i = 0 \\ -1 & \text{if } x_i < 0 \end{cases}$$

Handling missing data

→ missing @ random

vs. " systematically

→ imputation & its effect on central tendency & spread