Lesson Plan 3/6

Tuesday, March 6, 2018 5:26 AM

Admin:
- Midterm grading delay… coming soon
- HW2 grades posted tomorrow

HDSC:
- The amazing life and contributions of John W. Tukey: https://www.stat.berkeley.edu/~brill/Papers/life.pdf

Project 1:
- Questions/comments?

EDA (cont'd)
- Range
- Central Tendency
  - Why does the mean matter?
- Spread
  - Variance
  - IQR
- Outliers
- Skew
- Covariance and Correlation

Exploratory Data Analysis

- Central Tendency
- Spread

- Skew
- Outlier

Central Tendency

- median
- mean (average)

Notation:

n entities

p attributes

Single attribute

x_i : i^{th} observation $i \in [1, n]$

$x_{(1)}$: rank - smallest value

$x_{(n)}$: largest value

→ $x_{(\frac{n}{2})}$: median (half of the values $\leq x_{(\frac{n}{2})}$

Sample mean :

→ $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

✱ Mean is the optimal value of a distance optimization

problem

Intuition: central tendency: a value "close" to a lot of the data.

① Define close:

$$(x_i - \mu)^2 \quad \text{(squared distance)}$$

② "a lot of the data"

$$RSS(\mu) = \frac{1}{2n} \sum^{n} (x_i - \mu)^2$$

$i=1$

Residual sum of squares

* The sample mean $\bar{x}$ minimizes RSS(μ) over all values of μ

$$\bar{x} = \arg\min_{\mu} \frac{1}{2n} \sum_{i=1}^{n} (x_i - \mu)^2$$

→ Need to show

$$\frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{minimizes} \quad RSS(\mu)$$

First derivative test

First derivative

$$f'(x_0) = 0 \implies$$

$$x_0 = \arg\min_x f(x)$$

$$\frac{d}{d\mu} RSS(\mu) = \frac{d}{d\mu} \frac{1}{2n} \sum (x_i - \mu)^2$$

$$= \frac{1}{2n} \sum \frac{d}{d\mu} (x_i - \mu)^2 \quad \text{(sum rule)}$$

$$= \frac{1}{2n} \sum 2 (x_i - \mu) \frac{d}{d\mu} (x_i - \mu) \quad \text{(exponent + chain rule)}$$

$$= \frac{1}{2n} \sum 2 (x_i - \mu) * (-1)$$

$$= \frac{1}{n}\sum_i \mu - \frac{1}{n}\sum_i x_i$$

$$= \mu - \frac{1}{n}\sum_i x_i \quad \forall$$

$$\frac{d}{d\mu} RSS(\mu) = 0 \quad \Longrightarrow$$

$$\mu - \frac{1}{n}\sum_i x_i = 0 \quad \Rightarrow$$

$$\hat{\mu} = \frac{1}{n}\sum_i x_i = \bar{x} \qquad \blacksquare$$

Estimates are minimizers of optimization problems over data

* median(x) minimizes

$$sad(m) = \sum_i |x_i - m|$$

Sample Variance:

$$var(x) = \frac{1}{n} \sum_i (x_i - \bar{x})^2 = RSS(\bar{x})^*$$

note: $\frac{1}{n-1} \sum_i (x_i - \bar{x})^2$

standard deviation

$$\sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2}$$

$\text{Var}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Rank-based spread:

IQR: inter-quartile range

$[x_{(\frac{n}{4})}, x_{(\frac{3n}{4})}]$ (half the data within this range)

Outliers:

values that are unusually "far away" from center

$$outliers_{sd}(x) = \{ x_i \mid |x_i| > \bar{x} + k * sd(x) \}$$

$$k = 1.5$$

$$k = 3$$

$$outliers_{IQR}(x) = \{ x_i \mid x_i < x_{(\frac{n}{4})} - k * IQR(x) \text{ or } x_i > x_{(\frac{3n}{4})} + k * IQR(x) \}$$

Skew

Is the bigger spread below us.
above central tendency

$x_{(\frac{n}{2})} - x_{(\frac{n}{4})}$ lower spread

$x_{(\frac{3n}{4})} - x_{(\frac{n}{2})}$ higher spread

---

central: median / mean — optimality of these estimates

spread: variance
standard deviations
IQR

skew : IQR

outliers: sd rule

IQR rule

Covariance

Correlation

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Unit-less

$$cor(x,y) = \frac{cov(x,y)}{sd(x)\,sd(y)} \qquad [-1, 1]$$