Statistics for data science

Tuesday 22nd November, 2016

Statistics, what is it good for?

A data science perspective

Data $\longrightarrow$ Knowledge

Statistics, what is it good for?

A data science perspective

Data $\longrightarrow$ Knowledge

Knowledge is a statistical entity

Statistics, what is it good for?

A data science perspective

Data $\longrightarrow$ Knowledge

Knowledge is a statistical entity

- there is randomness in the data
  - a telephonic survey

Statistics, what is it good for?

A data science perspective

Data $\longrightarrow$ Knowledge

Knowledge is a statistical entity

- there is randomness in the data
  - a telephonic survey
- not enough data to achieve certainty
  - choosing the best restaurant when there are very few reviews

Data $\longrightarrow$ Knowledge

Knowledge is a statistical entity

- there is randomness in the data
  - a telephonic survey
- not enough data to achieve certainty
  - choosing the best restaurant when there are very few reviews
- the knowledge required is of a statistical nature
  - What are my odds when betting on a soccer match

Statistics, what is it good for?

Statistics is an important component of a data scientist's toolbox

Statistics, what is it good for?

Statistics is an important component of a data scientist's toolbox

- "raw" statistics - "getting a feel of the data"

data exploration

Statistics, what is it good for?

Statistics is an important component of a data scientist's toolbox

- "raw" statistics - "getting a feel of the data"

  data exploration

- many data science tools are of a statistical nature

Statistics, what is it good for?

Statistics is an important component of a data scientist's toolbox

- "raw" statistics - "getting a feel of the data"

  data exploration

- many data science tools are of a statistical nature
- a knowledge of statistics is essential
  - choose the right tool
  - correctly interpret results

Structured Vs. Unstructured data

Structured Data

data is organized in well defined fields ...

Structured Vs. Unstructured data

Structured Data

data easily fits in a relational database - tables

Examples?

Structured Vs. Unstructured data

Structured Data

Name	**Height** (H)	**Weight** (W)
Prince Humperdinck	6’	166lb
Inigo Montoya	5’3”	142lb
Princess Buttercup	5’4”	136lb

$\vdots$

Structured Vs. Unstructured data

Unstructured Data

data is not organized in well defined fields...

Structured Vs. Unstructured data

Unstructured Data

data does not easily fit in a relational database - tables

what's easily? examples?

Structured Vs. Unstructured data

Unstructured Data

Structured Vs. Unstructured data

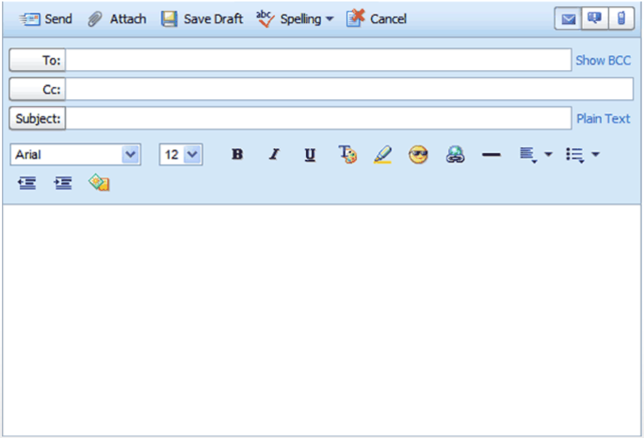
Semi-structured Data

Data is partially structured and partially unstructured

examples?

Structured Vs. Unstructured data

Semi-structured Data



Data Types

- Numerical - discrete/continuous

Data Types

- Numerical - discrete/continuous
  - Height
  - Temprature
  - Number of children

Data Types

- Numerical - discrete/continuous
  - Height
  - Temprature
  - Number of children
- Categorical

Data Types

- Numerical - discrete/continuous
  - Height
  - Temprature
  - Number of children
- Categorical
  - Gender
  - Nationality

Data Types

- Numerical - discrete/continuous
  - Height
  - Temprature
  - Number of children
- Categorical
  - Gender
  - Nationality
- Ordinal

Data Types

- Numerical - discrete/continuous
  - Height
  - Temprature
  - Number of children
- Categorical
  - Gender
  - Nationality
- Ordinal
  - 1st, 2nd, 3rd...
  - low, medium, high

Data Types

- Arbitrary (string)
  - natural language - quotes, reviews, phrases
  - hashes

Data Types

- Arbitrary (string)
  - natural language - quotes, reviews, phrases
  - hashes
- Time series
  - Navigation data (Waze)
  - Stock market data

A "practical" example

Physical measurements of N test subjects

Name	Height (H)	Weight (W)
Prince Humperdinck	6'	166lb
Inigo Montoya	5'3"	142lb
Princess Buttercup	5'4"	136lb

$\vdots$

A "practical" example

Physical measurements of N test subjects

- What can you say about height?
- What can you say about weight?

A "practical" example

IsraelTech challenge

Physical measurements of N test subjects

- mean

μ_X

$$E\left[X\right] = \sum_x x \cdot Pr(x)$$

$$E\left[X\right] = \int_x x \cdot f(x)dx$$

- variance

σ_X^2

$$V\left(X\right) = E\left[\left(X - E\left[X\right]\right)^2\right]$$

A "practical" example

Physical measurements of N test subjects

- What can you say about height?
- What can you say about weight?
- How would you assess the connection between the two?

A "practical" example

Physical measurements of N test subjects

- Covariance

$$cov\left(X,Y\right) = E\left[\left(X - E\left[X\right]\right)\left(Y - E\left[Y\right]\right)\right]$$

A "practical" example

Physical measurements of N test subjects

- Covariance

$$cov\left(X,Y\right)=E\Big[\big(X-E\left[X\right]\big)\big(Y-E\left[Y\right]\big)\Big]$$

- Covariance of height and weight

$$cov(H,W)=6.27$$

A "practical" example

Physical measurements of N test subjects

News from the citadel...

they've invented the metric system

A "practical" example

Physical measurements of N test subjects

News from the citadel...

Name	**Height** (H)	**Weight** (W)
Prince Humperdinck	183cm	75.3kg
Inigo Montoya	160cm	64.41kg
Princess Buttercup	162.5cm	61.69kg

$\vdots$

What happens to the connection between height and weight?

A "practical" example

Physical measurements of N test subjects

- Covariance of height and weight

$$cov(H, W) = 87.1125$$

Discussion

- What's the problem with using covariance to assess the connection between height and weight?

Discussion

- What's the problem with using covariance to assess the connection between height and weight?
- We need a measure that is invariant to a change of units (scale)

Discussion

- What's the problem with using covariance to assess the connection between height and weight?
- We need a measure that is invariant to a change of units (scale)
- How about normalizing the covariance?

Pearson Correlation

$\rho_{X,Y}$

informally: normalized covariance

$$corr\left(X,Y\right)=\frac{cov\left(X,Y\right)}{\sigma_X\cdot\sigma_Y}$$

$$=\frac{E\Big[\big(X-E\left[X\right]\big)\big(Y-E\left[Y\right]\big)\Big]}{\sqrt{E\Big[\big(X-E\left[X\right]\big)^2\Big]}\sqrt{E\Big[\big(Y-E\left[Y\right]\big)^2\Big]}}$$

Pearson Correlation

- $-1 \leq \rho_{X,Y} \leq 1$
- high direct correlation $\rho_{X,Y} \approx 1$
- high inverse correlation $\rho_{X,Y} \approx -1$

Pearson Correlation

$\boxed{\rho_{X,Y}}$

Examples

- direct correlation
  - X number of days it $\boxed{rains}$ in the morning
  - Y number of days sidewalks are $\boxed{wet}$ at noon
- inverse correlation
  - X number of days it $\boxed{rains}$ in the morning
  - Y number of days sidewalks are $\boxed{dry}$ at noon

A word about correlation and dependence

statistical dependence...

A word about correlation and dependence

Israeltech challenge

Does a lack of correlation imply independence

$$\rho_{X,Y} = 0 \overset{?}{\Rightarrow} X \perp\!\!\!\perp Y$$

A word about correlation and dependence

Does a lack of correlation imply independence

$$\rho_{X,Y} = 0 \overset{?}{\Rightarrow} X \perp\!\!\!\perp Y$$

No , Correlation $=$ Linear dependence

X, X^2

A word about correlation and dependence

Does a lack of correlation imply independence

$$\rho_{X,Y} = 0 \overset{?}{\Rightarrow} X \perp\!\!\!\perp Y$$

No , Correlation $=$ Linear dependence

X, X^2

A word about correlation and dependence

Does correlation imply dependence

$$\rho_{X,Y} \in \{-1, 1\} \overset{?}{\Rightarrow} X \not\perp\!\!\!\perp Y$$

A word about correlation and dependence

Israeltech challenge

Does correlation imply dependence

$$\rho_{X,Y} \in \{-1, 1\} \overset{?}{\Rightarrow} X \not\perp Y$$

Yes

A word about correlation and dependence

Does correlation imply dependence

$$\rho_{X,Y} \in \{-1, 1\} \overset{?}{\Rightarrow} X \not\perp Y$$

Yes , Linear dependence = a private case of dependence

A word about samples

- statistical operators - mean, variance,...

A word about samples

- statistical operators - mean, variance,...

$$E[X] = \sum_{x} x \cdot Pr(x)$$

A word about samples

- statistical operators - mean, variance,...

$$E\left[X\right] = \sum_{x} x \cdot Pr\left(x\right)$$

- how are these computed from a sample?

A word about samples

- statistical operators - mean, variance,...

$$E\left[X\right] = \sum_{x} x \cdot Pr\left(x\right)$$

- how are these computed from a sample?
- sample operators - sample mean, sample variance, ...
- all based on averaging

A word about samples

$$X = \{x_1, x_2, \ldots, x_n\}$$

Sample Mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A word about samples

$$X = \{x_1, x_2, \ldots, x_n\}$$

Sample Variance

$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2$$

A word about samples

$$X = \{x_1, x_2, \ldots, x_n\}$$

Sample covariance

$$cov\left(X, Y\right) = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)$$

$$= \frac{1}{n}\left(X - \overline{X}\right)^T \left(Y - \overline{Y}\right)$$

recap

- statistics is important
- knowledge is statistical
- data comes in diverse forms
- simple relations between data elements - covariance, correlation