Master studije
GEOINFORMATIKA

Mašinsko učenje - vežbe 2

Tatjana Jakšić Krüger

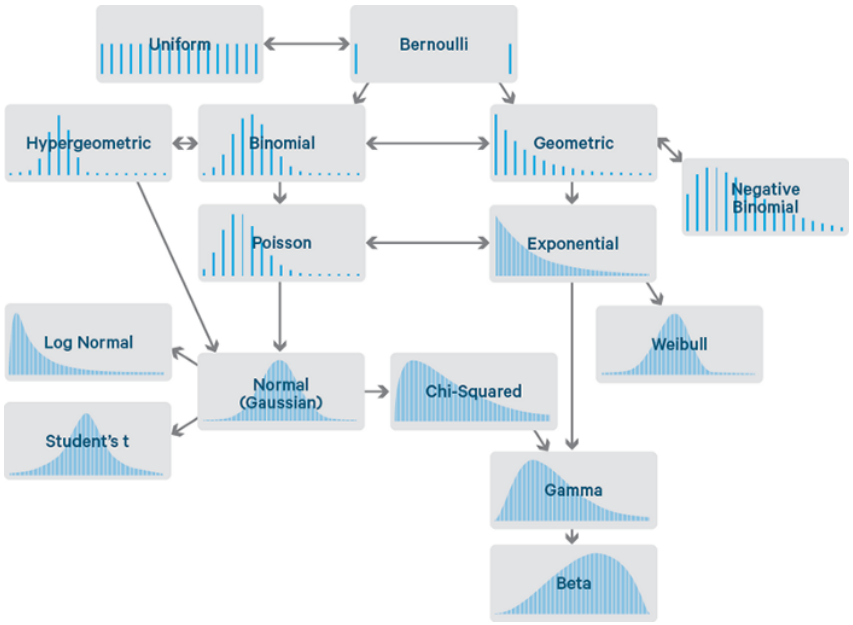
tatjana@turing.mi.sanu.ac.rs

29. oktobar 2021.

Građevinski fakultet Univerziteta u Beogradu

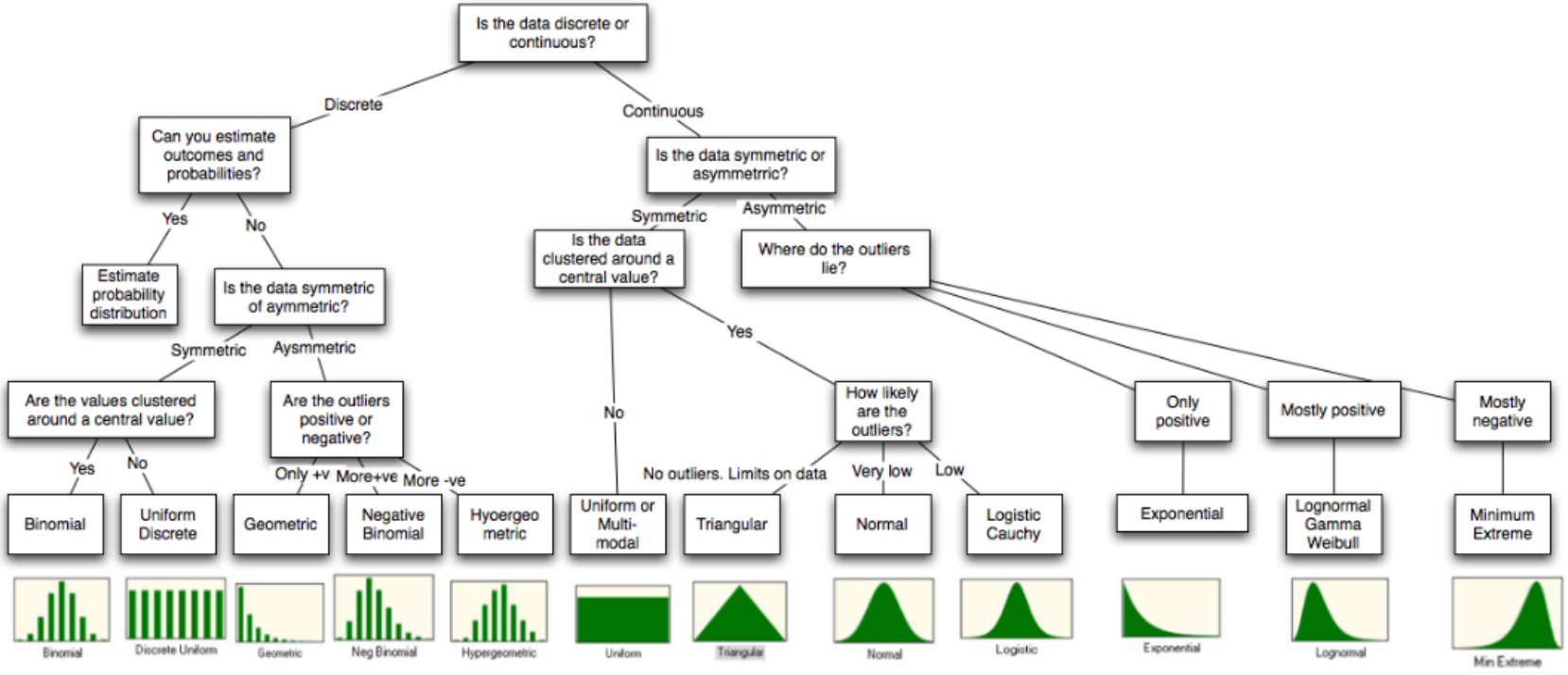
Raspodela podataka

- Prikazujemo raspodelu (distribuciju) podataka u jednom uzorku.
- Histogram.



Izvor https://medium.com/swlh/understanding-probability-distribution-b5c041f5d564

Raspodela podataka



Izvor `http://people.stern.nyu.edu/adamodar/New_Home_Page/StatFile/statdistns.htm`

Transformacija i normalizacija podataka

- Normalizacija na [0,1] interval.

$$z_i = \frac{x_i - min_{i \in [1,n]}(x_i)}{max_{i \in [1,n]}(x_i) - min_{i \in [1,n]}(x_i)}$$

- Konvertovanje u z-score vrednosti.

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

- Logistička transformacija.

$$z_i = \frac{1}{1 + \exp(-x_i)}$$

- Log transformacija.

$$z_i = log(x_i)$$

- Rotacija, translacija, druga preslikavanja.

Transformacija i normalizacija podataka

- Eliminisati neiskorićene kolone.
- Promeniti tip podataka. Npr. dugačke nazive konvertovati u integer zapis.
- Nedostajući podaci (Na, NaN).
- Eliminisati karaktere: linije, zagrade, itd.
- Konvertovati kategoričke promenljive u numeričke.
- Vreme zapisati kao realan broj.

Algoritmi za koje vrsimo transformacije

- Algoritmi za klasifikaciju.
- Logstička regresija.
- Neuralne mreže, itd.

Korisni linkovi

- Ukratko o R-u: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- R-studio: https://www.rstudio.com/products/rstudio/download/#download
- Članci: www.r-bloggers.com.
- Literature za pripremu podataka: Jason Brownlee, "Data Cleaning, Feature Selection, and Data Transforms in Python".

Hvala na pažnji.

Sada idemo na prikaz komandi.

Da li imate pitanja?