

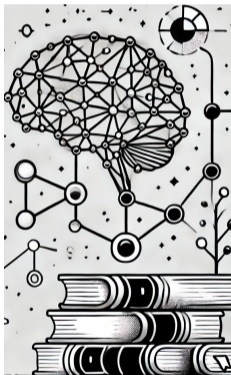
Kauzalita a metódy strojového učenia

Lukáš Lafférs

Katedra matematiky, Univerzita Mateja Bela

Vedatour 2024

Dnes



Algoritmy strojového učenia sa ukázali ako skvelý nástroj na predikciu!

Vedia nám však pomôcť aj s odhadovaním kauzálnych vzťahov??

Kauzalita

Prečo sa to stane?

Vyžaduje porozumenie.

Kus ťažšie.

Predikcia

Čo sa stane?

Ak funguje, tak funguje, **no a čo.**

Kus jednoduchšie.

Michal absolvoval rekvalifikačný kurz. Pomohlo to?

Mohli by sme ho porovnať s Jozefom, ktorý ho neabsolvoval.



Michal

- 27 rokov
- ženatý
- z Brezna
- stredná škola
- vie anglicky
- zdravý
- vod. preukaz typu B
- býva s 3 ďalšími ľuďmi
- ...

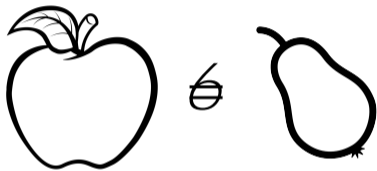
Jozef

- 53 rokov
- slobodný
- z Myjavy
- VŠ
- vie nemecky
- mal autonehodu
- vod. preukazy typu B,C
- stará sa o rodičov
- ...



Prekvapivo. Michal a Jozef sú rôzni.

Michal \neq Jozef





Michal

- 27 rokov
- ženatý
- z Brezna
- stredná škola
- anglicky
- zdravý
- vod. preukaz typu B
- býva v dome s 3 ďalšími ľuďmi
- ...

Peter

- 29 rokov
- slobodný
- z Podbrezovej
- stredná škola
- anglicky, francúzsky
- zdravý
- vod. preukaz typu B
- býva s 2 ľuďmi a so psom
- ...



Prekvapivo. Aj Michal a Peter sú rôzni.

Ale o dosť podobnejší.

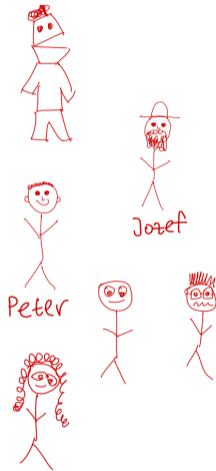
Michal Peter



Párovanie

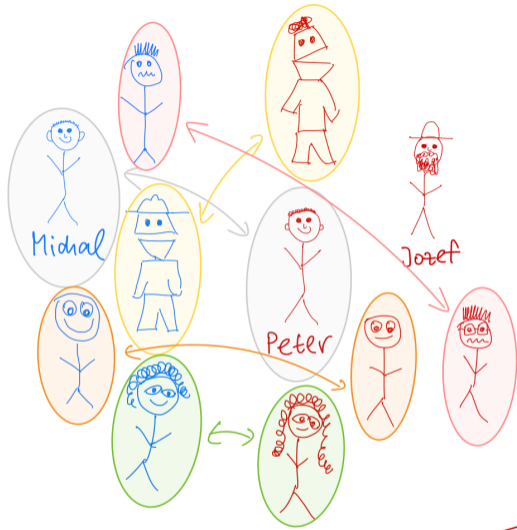


KURZ



Jozef

~~KURZ~~



KURZ

~~KURZ~~

Takýchto **Petrov** je však málo.

A nie každý má svoju "dvojičku" ako **Michal**.

Limitácie párovania



Princ Charles

- Muž
- Narodený 1948
- Vyrastal v UK
- Dvakrát ženatý
- Býva na zámku
- Bohatý
- Slávny

Ozzy Osbourne

- Muž
- Narodený 1948
- Vyrastal v UK
- Dvakrát ženatý
- Býva na zámku
- Bohatý
- Slávny

Charles é Ozzy

O Michalovi vieme **veľa** informácií.

Ale absolventov kurzu je **málo**.

- vek
- slobodný/á
- počet detí
- mesto
- typ vzdelania
- oblasť vzdelania
- znalosť cudzích jazykov
- zdravotné znevýhodnenie
- vod. preukaz
- história zamestnania
- typ predošlej práce
- dĺžka predošlého zamestnania
- klasifikácia predošlého zamestnania
- počet členov v domácnosti

- bariéry zamestnanosti
- národnosť
- sociálne dávky
- zdravotné poistenie
- zdravotné znevýhodnenie
- osamelý občan
- ochota vzdelávať sa
- ochota dochádzať za prácou
- účasť na predošlom kurze
- záujem o prácu o neúplný úväzok
- záujem o prácu v zahraničí
- úroveň segregácie
- vzdialenosť od krajského mesta
- vzdialenosť od Bratislavy
- ...

Tradičné štatistické metódy nevedia pracovať s mnohorozmernými dátami.

Čo s tým?

Priemerný efekt absolvovania kurzu na mzdu

$$= E \text{ MZDA}(\text{kurz}) - \text{MZDA}(\cancel{\text{kurz}})$$

m

 Predikovať priamo

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

Dáta

 m
!

mzda

 p

 Preváhovanie

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

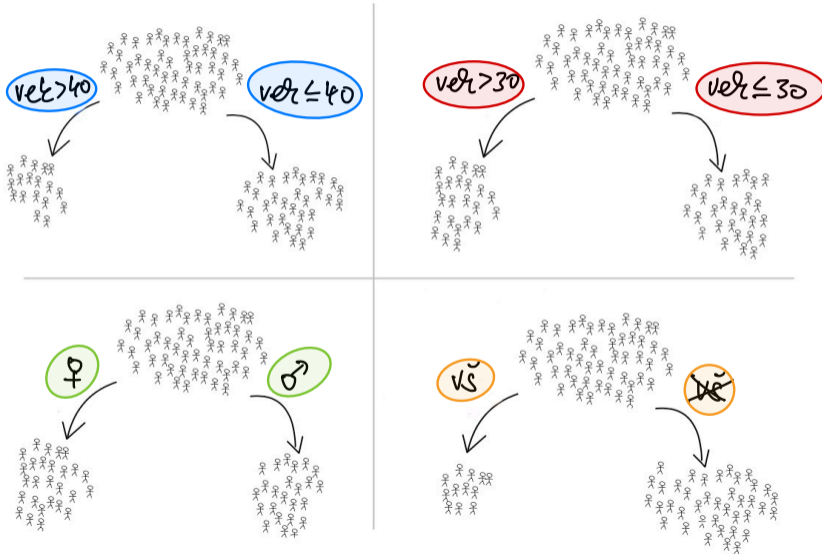
Dáta

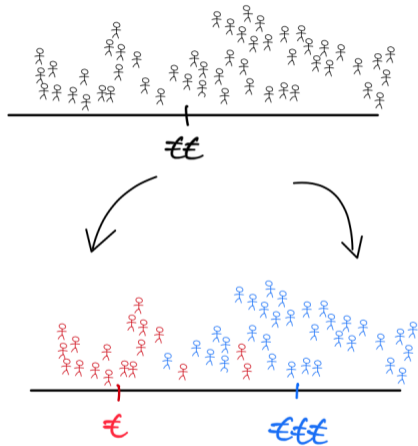
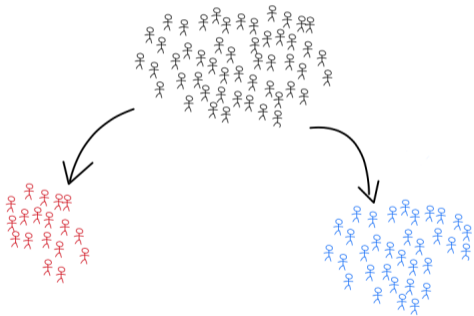
 p
!

ide na kurz

Aj m aj p odhadneme pomocou MACHINE LEARNING metód strojového učenia.

Príklad algoritmu strojového učenia



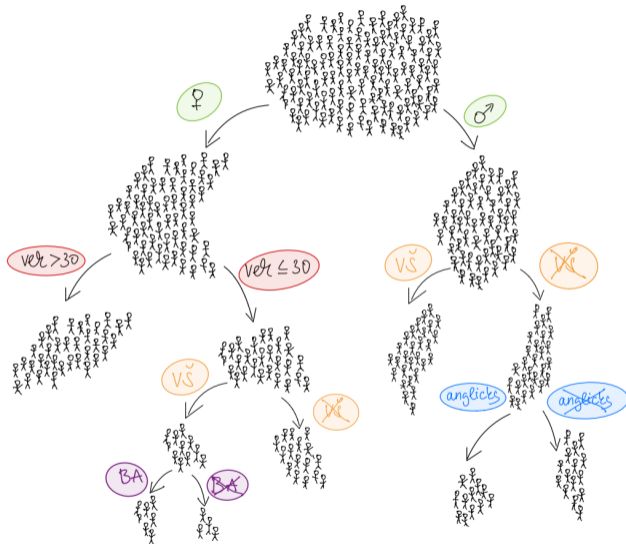


Ako nariasť strom?

Dobrá áno/nie otázka?

Vieme to odmerať?

Ako veľký strom?

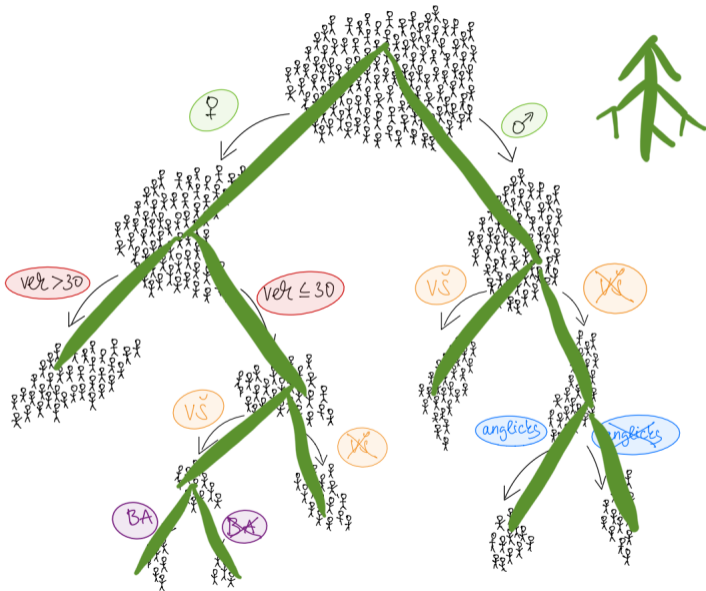


Ako nariasť strom?

Dobrá áno/nie otázka?

Vieme to odmerať?

Ako veľký strom?



Náhodný les

Pozdravujem vás, lesy, hory,
z tej duše pozdravujem vás!

P. O. Hviezdoslav



P.O.H. v lese.

(Nie až tak) náhodný les

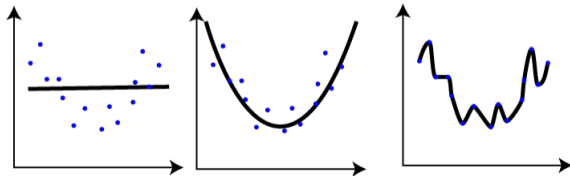


(Už celkom) náhodný les



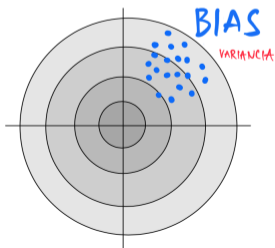
- Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Nevýhody

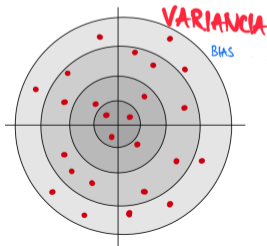


PREDIKČNÁ CHYBA = NÁHODNÁ CHYBA + NENÁHODNÁ CHYBA

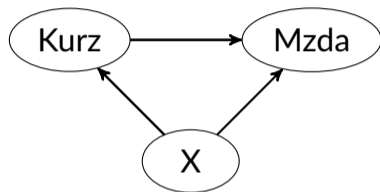
= NÁHODNÁ CHYBA + $\underbrace{\text{BIAS} + \text{VARIANCIA}}_{\text{NENÁHODNÁ CHYBA}}$



!



Naspäť k príkladu



X - informácie o veku, pohlaví, vzdelaní, zručnostiach...

$$= E \text{ MZDA}(\text{kurz}) \quad \text{MZDA}(\cancel{\text{kurz}})$$

- Pearl, J. (2009). Causality. Cambridge university press.
- Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge university press.

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

m
!

m

efekt na mzdu

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

p
!

p

efekt na mzdu

Priama aplikácia m alebo p (prevážení) na odhad povedie k **BIASu**.

Obe m aj p sú zlé: m \nearrow a p \nearrow

Double Machine Learning

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

m, p
!

$m; p$
efekt na mzdu

$m; p$ je fajn.

$m; p$ '!

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, volume 21, pp. C1–C68.
- Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429), 122-129.
- Van Der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1).

Double Machine Learning

- vek
- pohlavie
- vzdelanie
- zručnosti
- ...

veľadimenzionálne
dáta (neprijemné)

μ, π
→

- Neyman ortogonalita
- cross-fitting

$\Delta_{\mu, \pi}$ je **fajn.**

$\Delta_{\mu, \pi}$

efekt na mzdu



- asymptoticky normálny
- neprišľený
- s konečnou variáciou
- $o_p(n^{-1/2})$ konvergenie

Zhrňme si to

- Kauzalita je náročná
- S veľa informáciami ešte viacej

BIAS vs VARIANCIA

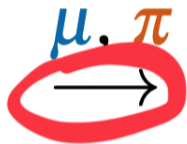
- ML metódy sú skvelé na predikciu
- Na odhadovanie parametrov nie až tak

m \nearrow a p \nearrow

- Skombinujeme dve slabé ML metódy $m;p$
- Dostaneme **odhad** s dobrými vlastnosťami

$m;p$ \nearrow

Rozšírenia

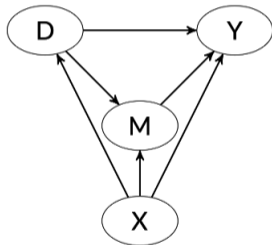


$\Delta_{\mu, \pi}$
efekt

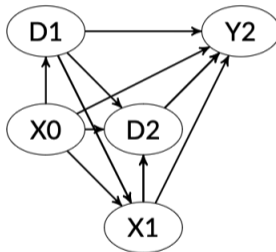
???



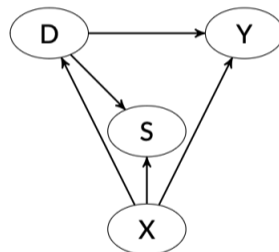
Mediačná analýza



Dynamické efekty



Výberová vzorka



- Mediačná analýza (H. Farbmacher, M. Huber, H. Langen, LL, M. Spindler)
- Dynamické efekty (H. Bodory, M. Huber, LL)
- Výberová výchylka (M. Bia, M. Huber, LL)

🍏 \notin 🍐 Párovanie 🍏 🍏

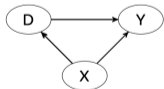


BIAS vs VARIANCA

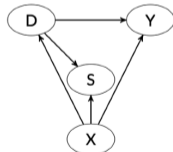
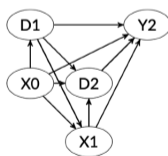
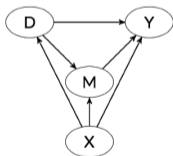
m \neq
 p \neq

Double ML \neq

$m; p$ \neq



Rozšírenia



Ďakujem za pozornosť.

www.lukaslauffers.com

#UMBmath