

ZÁPADOČESKÁ UNIVERZITA V PLZNI

OBOROVÝ PROJEKT IS

KIV/OPIS

Propensity Score Matching

Autor:
Marek Lovčí

22. března 2021



Obsah

1	Zadání	2
2	Vypracování	2
2.1	Statistické párování	2
2.1.1	Kovariát (covariate)	3
2.2	Ohodnocení podobnosti subjektů	3
2.2.1	Redukce biasu	4
2.3	Ověření hypotéz	4
3	Závěr	6

1 Zadání

1. Řešení statistického párování + PSM.
2. Pochopení PSM, tzn. jak matematicky funguje, jaké má předpoklady/limitace. V jakých situacích se (ne)hodí.
3. Vyzkoušení na ostrých datech.

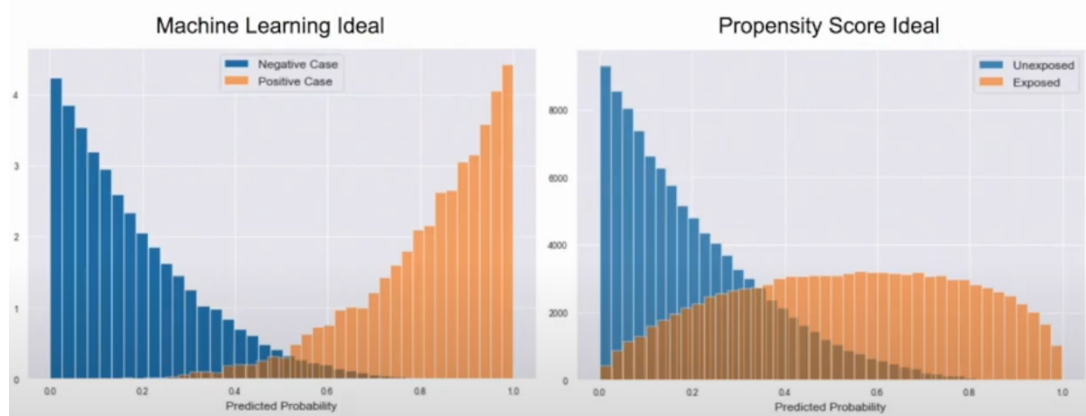
2 Vypracování

2.1 Statistické párování

Párování (v angl. „matching“) je statistická technika používaná k vyhodnocování efektu léčby porovnáním léčených a neléčených pacientů. Byla vyvinuta s cílem provádět observační studie stejně jako randomizované studie. Cílem párovacích metod je nalézt pro každého léčeného pacienta jednoho nebo více neléčených pacientů s podobným stavem oproti kterým lze evaluovat efekty léčby. Premisou takového postupu je umožnit odhadnout reálné účinky léčby a současně redukovat vzniklý statistický bias.

Jak již bylo zmíněno, hlavní myšlenkou je spojit jednotlivé pozorované subjekty ve skupině podstupující léčbu $A = 1$ s podobnými jedinci v kontrolní skupině $A = 0$ (obecně nepodstupující léčbu, či podstupující léčbu jinou). Podobnost jedinců je určena skrze matici příznaků (čili kovariátů) X . Takovým přístupem získáme zdání kauzality dostatečné v takové míře, že je lze za kauzalitu považovat.

Jakmile jsou data spárována, lze s nimi zacházet jako s randomizovanou studií. Nevýhodou přístupu je, že obvykle nelze pracovat s celou maticí příznaků. Práce s kompletní maticí příznaků je vhodná pro úlohu klasifikace, kdy se snažíme jednoznačně odlišit jednu skupinu od druhé. To ale není pro párování příliš praktické. V této úloze naopak hledáme matici příznaků takovou, které nám umožní najít znatelný překryv. Právě v překryvu dvou skupin lze hledat entity vhodné pro párování.



Obrázek 1: Klasifikace vs PSM

Při provádění metod statistického párování postupujeme tedy následovně:

1. data rozdělíme na dvě skupiny, kontrolní a experimentální; kontrolní je obvykle větší;
2. vybereme příznaky, které budou použity pro výpočet skóre podobnosti;

3. vypočteme skóre podobnosti (logistickou regresí, či jinou metodou);
4. dle skóre z předešlého kroku spárujeme prvky z kontrolní a experimentální skupiny (například pomocí algoritmu pro hledání nejbližšího souseda).

Výběr příznaků pro metodu PSM je volbou analytika v závislosti na konkrétní situaci. Jak již bylo zmíněno výše, je nutné zvolit takové příznaky, které umožní najít jistý překryv kontrolní a léčené skupiny.

Při výpočtu skóre náchylnosti je výchozí volbou *logistická regrese*. Použití jiných metod záleží (stejně jako u *výběru příznaků*) pouze na analytikovi a dané situaci. V článku „Propensity score estimation: machine learning and classification methods as alternatives to logistic regression“ z roku 2010 popisují autoři použití i neuronových sítí, SVM a rozhodovacích stromů. Nejčastější alternativní volbou by byl pravděpodobně Bayesův klasifikátor. Ten podle článku „Bayesian propensity score analysis for observational data“ z roku 2009 bude lepší volbou obzvláště v případech, kdy se v datech nachází nějaká míra nejistoty.

Při párování se pro každý prvek experimentu hledá prvek v kontrolní skupině. Implicitně se předpokládá, že kontrolní skupina je početnější, jelikož léčba je obvykle nákladná nebo nemoc vzácná. Otázka nastává, co dělat s případy, kdy je kontrolní skupina menší než léčená (což nastává i v naší situaci). V takových případech je metoda PSM stále validní možností, protože úloha není nutně o hledání párů dat, ale spíše o hledání dvou srovnatelných skupin dat. Toho i v tomto „inverzním“ případě dosáhneme (ani se nemění směr párování). Dle dostupných informací na fórech je však rozumné uvážit (a nastudovat) i další metody, jako např. *Coarsened Exact Matching*, které v takových případech mohou být někdy vhodnější.

2.1.1 Kovariát (covariate)

Kovariáty jsou charakteristikami účastníků experimentu (jejich vlastnosti, vyjma vlastní léčby), tedy jedná se o matici příznaků X . Mějme experiment zkoumající vliv antihypertenziv (léků snižujících krevní tlak) při léčbě posttraumatické stresové poruchy (PTSP) vyskytující se u válečných veteránů. Míra dávky léku je v tomto kontextu léčbou, zatímco věk a další biologické rysy (váha, výška) jsou faktory ovlivňující výkonnost léčby, jsou tedy kovariáty léčby.

2.2 Ohodnocení podobnosti subjektů

Abychom mohli realizovat cíl *statistického párování* potřebujeme nejdříve nadefinovat nějakou metriku, dle které půjde zhodnotit podobnost jednoho subjektu k subjektu druhému. Asi nejjednodušší metrikou jsou *propensity scores*, v překladu skóre náchylnosti či sklonu.

Již jsme definovali experimentální množinu subjektů $A = 1$ a kontrolní skupinu $A = 0$. „Propensity score“ π pro subjekt i definujeme jako pravděpodobnost příslušnosti subjektu do experimentální skupiny.

$$\pi_i = P(A = 1 | X_i)$$

Uvedme příklad, jestliže je skóre subjektu 0.3, pak to znamená, že na základě matice vektoru příznaků má 30% šanci, že spadá do experimentální skupiny (často označováno jako skupina *léčba*). Toto skóre je možné získat jakýmkoliv klasifikačním algoritmem (samozřejmě upraveným tak, aby nevracel jen třídu,

ke skteré si myslí, že prvek náleží, ale i regresní složku, tedy procentuální příslušnost). Tady lze využít např. *logistickou regresi*.

2.2.1 Redukce biasu

Přes pečlivý výběr konkrétních příznaků, které budou pro párování použity, se pravděpodobně nepovede vybrat příznaky zcela vyvážené (nezatížené biasem). Nevyváženost příznaků bohužel v této metodě způsobuje, že nedojde k dostatečnému podobnostnímu překryvu párovaných skupin. Potřebujeme tedy výpočet, díky kterému zjistíme které příznaky způsobují imbalance. To lze zjistit pomocí ‚standardized mean difference‘ (*smd*), tedy pomocí *standardizovaných odchylek od průměru*. Výpočet *smd* je rozdílem středních hodnot mezi dvěma skupinami, vyděleným sdruženou směrodatnou odchylkou.

$$smd = \frac{\bar{X}_t - \bar{X}_c}{\sqrt{\frac{(s_t^2 + s_c^2)}{2}}}$$

Spočítáme *smd* pro každý příznak. Pro určení, jestli je příznak vyvážený, nebo ne, se použije následující pravidlo:

- $smd < 0.1$ ideální stav,
- $smd \in (0.1 - 0.2)$, hraniční stav,
- $smd > 0.2$, vážně nevyvážený případ.

V ideálním případě jsou hodnoty *smd* rovny prvnímu případu, tedy do 0.1. Mnohdy tomu ale tak není, tento problém má několik řešení.

První řešení je odstranit entity s extrémními hodnotami *propensity score*. Odstraníme tedy subjekty, jejichž skóre je menší než minimum z *léčené* skupiny a jejichž skóre je větší než maximum z kontrolní skupiny a *vice versa*.

Druhé řešení je vzít v úvahu pouze takové dvojice subjektů (z kontrolní a experimentální skupiny), u nichž je rozdíl podobnostní metriky menší než specifikovaná mez δ .

Třetí řešení je použito v kódu dodaného spolu s tímto textem. Jedná se o řešení, kdy subjekt z experimentální skupiny je spárován s několika sousedy z druhé skupiny (kNN).

2.3 Ověření hypotéz

V rámci implementace PSM byli na datech ověřeny 2 hypotézy. První hypotézou H_1 je, že provedení léčebných postupů vede ke snížení NIHSS sledovaných pacientů. Hypotézy byly definovány následovně:

- H_0 : provedení léčebných postupů (Trombectomie, Trombolýza, Bridge) *nevede* ke snížení NIHSS sledovaných pacientů;
- H_A : provedení léčebných postupů *vede* ke snížení NIHSS sledovaných pacientů.

Druhou ověřovanou hypotézou H_2 bylo, že u pacientů ve věku do 65 let provedení léčebných postupů vede ke snížení NIHSS většímu než u pacientů nad 65 let.

- H_0 : u pacientů ve věku do 65 let provedení léčebných postupů (Trombectomie, Trombolýza, Bridge) *nevede* ke snížení NIHSS ve srovnání s pacienty nad 65 let;
- H_A : u pacientů ve věku do 65 let provedení léčebných postupů *vede* ke snížení NIHSS sledovaných pacientů ve srovnání s pacienty nad 65 let.

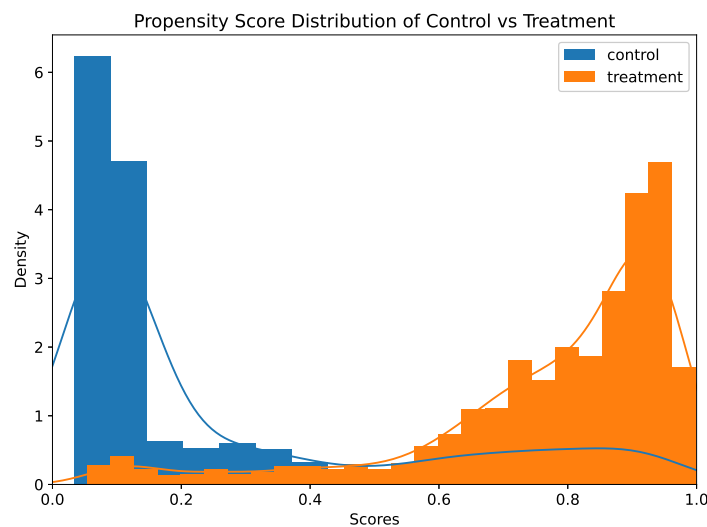
V obou případech se podařilo T-testem zamítnout nulovou hypotézu na úrovni hladiny $\alpha = 5 \%$ a přijmout hypotézu alternativní.

	Statistika	p-value	Velikost kontrolní skupiny	Velikost exponované skupiny
H_1	-7.99	2.6×10^{-15}	1312	1578
H_2	-5.63	2.3×10^{-8}	385	1193

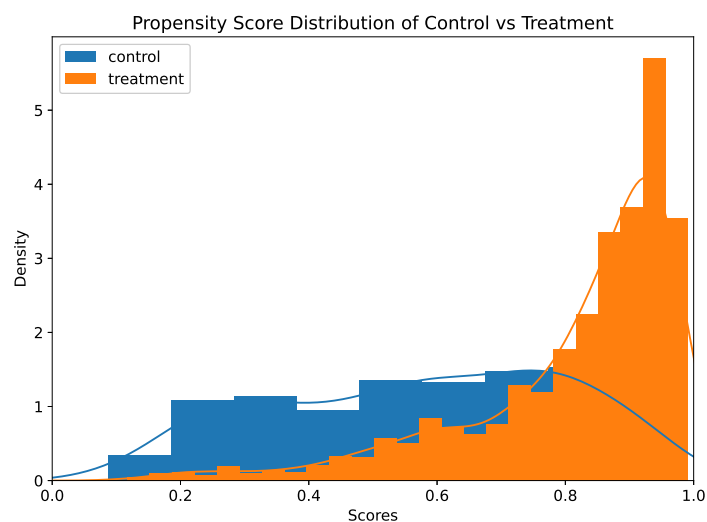
Dle výsledků statistik z tabulky výše můžeme vyvodit, že léčení pacienta (první hypotéza) vede ke znatelnému snížení NIHSS (znaménko mínus ve statistice lze v tomto případě interpretovat jako snížení NIHSS). U druhé hypotézy můžeme soudit obdobně.

Datasetu obsahuje celkem 2890 záznamů. V první hypotéze uvažujeme celý dataset rozdělený do dvou skupin, přičemž kontrolní skupina obsahuje 1312 záznamů a exponovaná skupina (léčená) obsahuje celkem 1578 záznamů (66 trombektomií, 1251 trombolýz a 261 bridge), pro kontrolu $66 + 1251 + 261 = 1578$ a $1312 + 1578 = 2890$. V druhé hypotéze uvažujeme redukovaný dataset, pouze léčených pacientů (celkem 1578 záznamů), rozdělený podle věku pacientů (nad 65 let včetně v počtu 1193 a pod 65 let v počtu 385 pacientů).

Při párování se v uvažovaných hypotézách, kvůli dominanci exponované skupiny, multiplikují prvky z kontrolních skupin. Párování proběhne vždy v celém oboru léčené skupiny, je tedy párováno 1578 dvojic pro H_1 a 1193 dvojic pro H_2 .



Obrázek 2: Graf Propensity Scores pro H_1



Obrázek 3: Graf Propensity Scores pro H_2

3 Závěr

Ačkoliv je PSM velice efektivní technikou, tak v některých publikacích není doporučována, konkrétně v článku „Why Propensity Scores Should Not Be Used for Matching“. Autoři v tomto článku vysvětlují, kdy je technika PSM používána a proč je tak používána nesprávně. Prezентují techniky CEM (Coarsened Exact Matching) a MDM (Mahalanobis Discrepancy Matching), které představují alternativu k PSM dávající v určitých případech smysluplnější výsledky než PSM. V těchto případech se však jedná o „pouhou“ změnu definice podobnosti.

Vcelku se jedná o velice zajímavý statistický nástroj, v principu i snadno pochopitelný.