SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Filip Milinković

METODE STROJNOG UČENJA ZA KAUZALNU ANALIZU PODATAKA

Diplomski rad

Voditelj rada:
prof. dr. sc. Tomislav Šmuc

Zagreb, Rujan, 2017.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Kauzalna analiza dugo je prisutna u području statistike, no primjene u području strojnog učenja novijeg su datuma. Većina metoda za izbor značajki *(engl. feature selection)* ne bavi se određivanjem kauzalnih veza između mjerenih veličina *(engl. feature)* i ciljne veličine *(engl. target)*, nego su usredotočene na stvaranje što boljeg modela za predviđanje vrijednosti ciljne veličine.
U ovom radu razmatrati ćemo situaciju u kojoj poznavanje kauzalnih veza omogućuje točniji izbor značajki.
Cilj ustanovljavanja kauzalnih veza je predviđanje posljedica **akcija** i **manipulacija**, što se značajno razlikuje od predviđanja na temelju **opservacija** u kojem podrazumijevamo ne izvođenje eksperimenata odnosno intervencija na sustav koji promatramo. S druge strane akcije odnosno manipulacije predstavljaju promjene u sustavu.
Miješanje opservacijskih i intervencijskih predviđanja često dovodi do pogrešnih zaključaka. Uzmimo za primjer da je uočena korelacija između vremena provedenog u krevetu i smrtnosti. U ovoj situaciji je pogrešno zaključiti da ćemo provođenjem manje vremena u krevetu smanjiti svoj rizik od smrti. Vjerojatan kauzalni model jest da *bolest* uzrokuje povećanje u *vremenu u krevetu* i *smrtnosti*. Ovaj primjer pokazuje da korelirana značajka (*vrijeme u krevetu*) može poslužiti za predviđanje vrijednosti cilja (*smrtnost*) ako je sustav *stacionaran* (nema promjena u distribuciji niti jedne od promatranih varijabli), no ne omogućuje nam predviđanje posljedica u slučaju intervencija (*e.g. tjeranje* ljudi da provode više ili manje vremena u krevetu) pa se time jasno vidi razlika između **korelacije** i **kauzalnosti**.

Poglavlje 1

Ciljevi izbora značajki i kauzalne analize

1.1 Ciljevi izbora značajki

U problemu nadziranog izbora značajki, promatramo skupinu slučajnih varijabli $\mathbf{X} = [X_1, X_2, ..., X_n]$ i ciljnu slučajnu varijablu Y. Općenito, cilj izbora značajki jest **pronaći što manji podskup od X kojim možemo "dobro" procijeniti Y**.
Konkretnije, želimo postići slijedeće:

- **Predviđanje:** Postizanje što veće moći predviđanja vrijednosti varijable Y izbacivanjem značajki koje su nevažne iz $\mathbf{X}$ i smanjenjem dimenzionalnosti od $\mathbf{X}$
- **Učinkovitost:** Smanjenje potrošnje memorije, vremena treniranja i vremena procesiranja smanjivanjem količine i "cijene" potrebnih podataka
- **Razumijevanje podataka:** Prepoznavanje značajki koje su korelirane uz ciljnu veličinu Y

Ne postoji univerzalna metoda za rješavanje ovog problema, što je lako za razumijeti. Naime, podaci se pojavljuju u različitim oblicima i količinama. Osim toga, u raznim slučajevima razlikuju se ciljevi: ponekad je važnije napraviti brži algoritam, a ponekad točniji.

1.2 Ciljevi kauzalne analize

U kauzalnoj analizi zadan je skup slučajnih varijabli $\mathbf{X} = [X_1, X_2, ..., X_N]$ i zajednička distribucija $P(\mathbf{X})$. Pojedina varijabla može i ne mora biti istaknuta kao ciljna: svaka može

biti shvaćena kao cilj i kao značajka. Cilj kauzalne analize je otkriti kauzalne veze između varijabli iz jednog ili više od slijedećih razloga

- **Predviđanje:** Predviđanje budućih vrijednosti neke od varijabli bez vanjskih utjecaja na sustav (isto kao i u izboru značajki)
- **Proturječno predviđanje *(engl. counterfactual prediction)*:** Nakon uočavanja pojedinog rezultata, predviđanje do kakvog bi rezultata došlo da je napravljena drugačija akcija
- **Manipulacija:** Predviđanje posljedica pojedinih akcija, odnosno vanjskih promjena na sustav.
- **Razumijevanje podataka:** Utvrđivanje strukture modela iz kojeg podaci dolaze.

1.3 Usporedba izbora značajki i kauzalne analize

Jedna očita razlika između izbora značajki i kauzalne analize je činjenica da je u izboru značajki istaknuta varijabla Y, no i kauzalnu analizu možemo lako usredotočiti na jednu varijablu koju bi zvali Y ili redom varijable X_i smatrati ciljnima. Usporedbe radi, istaknut ćemo jednu varijablu Y.
Zajednički ciljevi izbora značajki i kauzalne analize su:

- Postizanje što boljeg **predviđanja.** U kauzalnoj analizi, postoji bitna razlika između predviđanja i manipulacija. U predviđanju se pretpostavlja da testni podaci dolaze iz iste distribucije $P(\mathbb{X}, Y)$ kao i podaci za treniranje, dok se u manipulaciji pretpostavlja da će distribucija biti promjenjena nekim vanjskim utjecajem.
- **Razumijevanje podataka.** I izbor značajki i kauzalna analiza pokušavaju pronaći "važne" faktore ili varijable za ciljnu varijablu. Razlika je u tome što u izboru značajki varijablu smatramo "važnom" za ciljnu ako između njih postoji korelacija, što u kauzalnoj analizi nije dovoljno.

Razlike između izbora značajki i kauzalne analize su:

- Kauzalna analiza se ne dotiče problema **učinkovitosti** u smislu vremena izvođenja. Ipak, u praksi je većina algoritama za kauzalnu analizu izvediva samo u slučajevima kad se radi o malom broju varijabli pa se metode izbora značajki koriste kako bi se izvukle najvažnije.
- Izbor značajki se ne dotiče **proturječnosti.**

Možemo zaključiti da je kauzalna analiza profinjenija od izbora značajki s obzirom da je cilj otkriti mehanizme sustava a ne samo statističke zavisnosti. Također, kauzalna analiza bi trebala omogućiti bolje razumijevanje stvarne strukture podataka, kao i mogućnost predviđanja posljedica manipulacija odnosno promjena u distribuciji.

1.4 Nekoliko primjera

Prije nego uđemo u formalne detalje, navesti ćemo dva primjera koji ilustriraju na koji način kauzalna analiza može poboljšati mogućnost predviđanja.

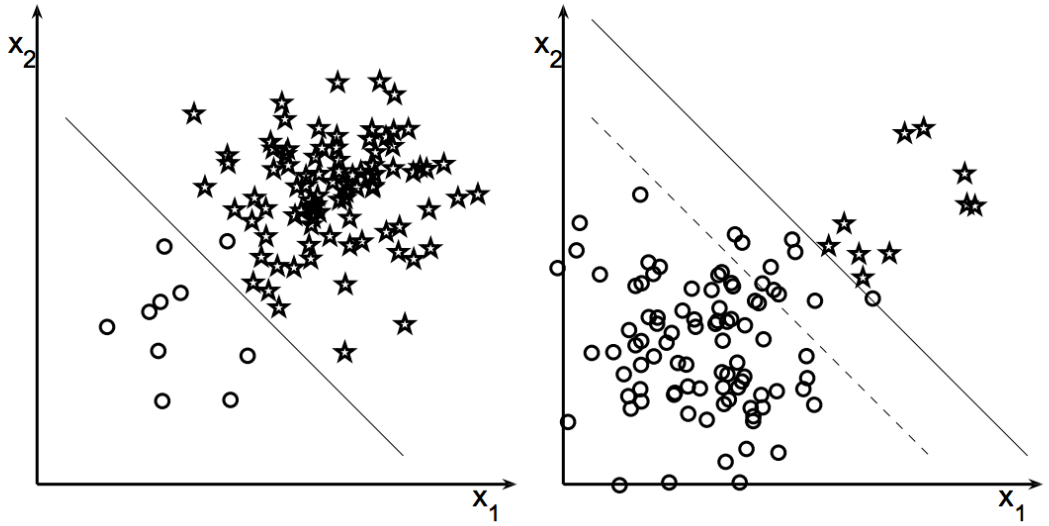
Primjer 1.4.1. *Promjena u distribuciji*

U strojnom učenju često pretpostavljamo da je niz slučajnih varijabli nezavisan i jednoliko distribuiran s distribucijom $P(\mathbf{X}, Y)$. Bez pretpostavki o mehanizmu kojim podaci nastaju, Bayesov teorem daje nam dvije ekvivalentne mogućnosti za modeliranje $P(\mathbf{X}, Y)$:

$$P(\mathbf{X}, Y) = P(\mathbf{X}) * P(Y|\mathbf{X}) = P(Y) * P(\mathbf{X}|Y)$$

S druge strane, ove dvije varijante nisu ekvivalente ako ih interpretiramo kao proces kojim nastaju podaci. U prvom slučaju (oznaka $\mathbf{X} \rightarrow Y$) uzorak $\mathbf{x}_i$ generiran je iz distribucije $P(\mathbf{X})$ i od njega distribucijom $P(Y|\mathbf{X})$ nastaje uzorak y_i, primjerice deterministčkim mehanizmom ili primjenom neke funkcije f i dodavanjem šuma ϵ: $y_i = f(\mathbf{x}_i) + \epsilon$. Drugi slučaj (oznaka $Y \rightarrow \mathbf{X}$) je analogan.
Prilikom ustanovljavanja kauzalnih veza, obično pretpostavljamo da se mehanizmi dobivanja podataka ne mijenjaju, to jest u slučaju $\mathbf{X} \rightarrow Y$, $P(Y|\mathbf{X})$ se ne mijenja, a u slučaju $Y \rightarrow \mathbf{X}$, $P(\mathbf{X}|Y)$ se ne mijenja. Ova pretpostavka zapravo govori "isti uzrok rezultira istim posljedicama". Uočimo da je ova pretpostavka slabija od nezavisnosti i jednolike distribuiranosti podataka koja kaže da distribucija $P(\mathbf{X}, Y)$ ostaje ne promijenjena.
Uzmimo za primjer problem klasifikacije, u kojem je Y varijabla klase (poprima vrijednosti +1 ili -1) i varijable X_1 i X_2 dvije neprekidne slučajne varijable koje su uvjetovane klasom Y, odnosno $Y \rightarrow \mathbf{X}$.
Cilj je iz testnog seta podataka odrediti distribuciju $P(Y|\mathbf{X})$, no može se utvrditi da je u ovom slučaju bolje odrediti distribuciju $P(\mathbf{X}|Y)$ i iskoristiti činjenicu da je $P(Y|\mathbf{X}) \sim P(\mathbf{X}|Y) * P(Y)$, gdje $\sim$ znači *proporcionalno*. U tom slučaju, ako dođe do promjene distribucije $P(Y)$ možemo je ponovo procijeniti i lako pomaknuti granicu odlučivanja.
Na slici 1.1 prikazani su na taj način generirani podaci. U trening setu Y je generirana prema distribuciji $P(Y = 1) = 0.9, P(Y = -1) = 0.1$, a $\mathbf{X} \sim N(\mu_{\pm}, \sigma)$, gdje je $\sigma = 0.75$ za obje klase a $\mu_+ = [0.8, 0.8], \mu_- = [0.2, 0.2]$.
U testnom setu uzeli smo promjenjenu distribuciju $P(Y)$, takvu da je sada $P(Y = 1) = 0.1, P(Y = -1) = 0.9$ dok je $P(\mathbf{X}|Y)$ ostala ne promijenjena. Iscrtkana linija predstavlja

granicu odlučivanja koju bismo dobili da smo trenirali direktno $P(Y|\mathbf{X})$ dok puna linija predstavlja, na ranije opisan način, pomaknutu granicu odlučivanja. Jasno vidimo da ovakav pristup daje točnije rezultate.



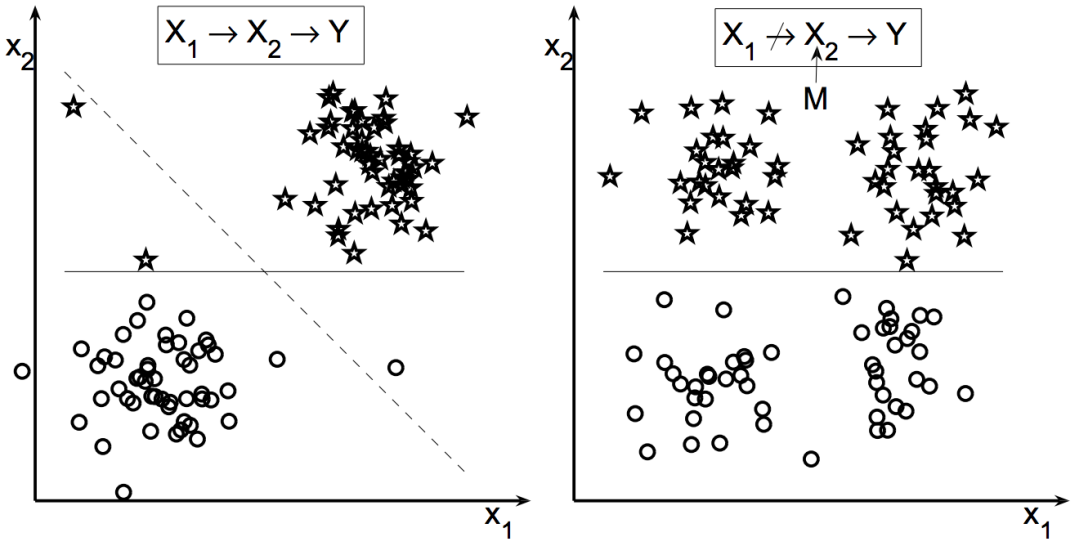
Slika 1.1: Problem klasifikacije

Primjer 1.4.2. *Manipulacija*

U ovom primjeru prikazat ćemo situaciju u kojoj su ponovo Y kategorijska, a X_1 i X_2 neprekidne slučajne varijable, no ovaj put X_1 i X_2 nisu nezavisne već vrijedi $X_1 \to X_2 \to Y$. Radi jednostavnosti dati ćemo varijablama neko značenje: primjerice neka je X_1 broj sati koji radnik u nekoj kompaniji provede obavljajući fizičke poslove, X_2 duljina rukava majice radnika, a Y indikator vide li se radniku laktovi. U tom slučaju pretpostavljamo da je radnicima koji provedu više vremena obavljajući fizičke poslove toplije pa nose kraću odjeću.
Na slici 1.2 prikazani su slučajno generirani podaci gdje su zvijezdom označeni radnici kojima se laktovi ne vide, a krugom oni kojima se vide.
Analitičar koji ne zna prirodu podataka lako bi mogao zaključiti da su i X_1 i X_2 relevantne i da je proces tipa $Y \to \mathbf{X} = [X_1, X_2]$ i u tom slučaju zaključiti da je iscrtkana linija optimalna granica odlučivanja, dok je jasno da Y direktno ovisi isključivo o X_2.
Pretpostavimo sada da je analitičar koji ne poznaje prirodu podataka u mogućnosti provesti eksperiment takav da "natjera" svakog zaposlenika da se zamijeni za majicu sa drugim slučajno odabranim zaposlenikom. U ovom slučaju više ne vrijedi $X_1 \to X_2$, nego smo sada u situaciji $M \to X_2 \to Y$ gdje je M slučajna permutacija majica. Podaci dobiveni na ovaj način prikazani su u slici 1.2 b). Sada se jasno vidi da prva hipoteza $Y \to \mathbf{X}$ nije istinita kao i da je očito $X_2 \to Y$ jer i nakon manipulacije postoji zavisnost između X_2 i Y. Također možemo zaključiti da X_2 nije uzrok X_1, pošto nakon manipulacije više ne postoji zavisnost između te dvije varijable pa je jedino objašnjenje za zavisnost prije manipulacije $X_1 \to X_2$.
Ovaj kauzalni model utvrđen manipulacijom kaže da je samo vrijednost X_2 potrebna za utvrđivanje vrijednosti Y, to jest da su X_1 i Y nezavisne uvjetno na X_2. U ovom slučaju X_2 je direktan uzrok, a X_1 indirektan i nije potreban za predviđanje vrijednosti Y ukoliko je poznata vrijednost X_2. Bez utvrđivanja kauzalnog modela početni model (iscrtkana linija) bi dovodio do pogrešnih predviđanja.

Iz ovog primjera proizlazi nekoliko zaključaka:

- Samo iz opservacija nije uvijek moguće utvrditi sve kauzalne veze.
- Predviđanja na temelju pogrešnog kauzalnog modela ($Y \to \mathbf{X}$) mogu se jako razlikovati od optimalnih (iscrtkana linija umjesto pune).
- Poznavanje prirode podataka može biti od velike pomoći. U tom slučaju bi lako mogli doći do točnog zaključka samo iz opservacija.
- Činjenica da postoji korelacija između X_1 i Y ne znači nužno da je X_1 koristan podatak za predviđanje Y, kao na primjer u slučaju kad je poznat X_2.



Slika 1.2: Rezultat manipulacije

- Manipulacije omogućuju utvrđivanje kauzalnih veza u situaciji kada se one ne mogu utvrditi isključivo iz opservacija.

Poglavlje 2

Osnovni pojmovi u kauzalnoj analizi

U ovom poglavlju ćemo prvo definirati neke pojmove važne za klasičan izbor značajki, a nakon toga analogne pojmove nužne za kauzalnu analizu. U nastavku podrazumijevamo da imamo podatke u obliku slučajnih vektora $\mathbf{X} = [X_1, X_2, ..., X_n]$ i ciljnu slučajnu varijablu Y koji dolaze iz distribucije $P(\mathbf{X}, Y)$

2.1 Individualna relevantnost značajke

Definicija 2.1.1. *Neka su A i B slučajne varijable i $\mathbf{C}$ skup slučajnih varijabli. Za A i B kažemo da su **nezavisne uvjetno na $\mathbf{C}$** i pišemo $A \perp B \mid C$ ako vrijedi*

$$P(A, B|C) = P(A|C) * P(B|C),$$

za sve vrijednosti A, B i $\mathbf{C}$.
U slučaju da je $\mathbf{C}$ prazan skup, kažemo da su A i B nezavisne i pišemo $A \perp B$

Definicija 2.1.2. *Za značajku X_i kažemo da je **individualno irelevantna za Y** ako za svaki $\mathbf{V}^{\setminus i}$ skup značajki koji ne sadrži X_i vrijedi*

$$P(X_i, Y|\mathbf{V}^{\setminus i}) = P(X_i|\mathbf{V}^{\setminus i}) * P(Y|\mathbf{V}^{\setminus i}),$$

za sve moguće vrijednosti $\mathbf{X}$ i Y

Definicija 2.1.3. *Za značajku X_i kažemo da je **jako relevantna za Y** ako postoje vrijednosti x, y i $\mathbf{v}$ za koje vrijedi $P(X_i = x, \mathbf{X}^{\setminus i} = \mathbf{v}) > 0$ takve da*

$$P(Y = y \mid X_i = x, \mathbf{X}^{\setminus i} = \mathbf{v}) \neq P(Y = y \mid \mathbf{X}^{\setminus i} = \mathbf{v})$$

Definicija 2.1.4. *Za značajku X_i kažemo da je* ***slabo relevantna za Y*** *ako nije jako relevantna i postoje podskup značajki $\mathbf{V}^{\setminus i}$ i vrijednosti x, y i* $\mathbf{v}$ *za koje vrijedi $P(X_i = x, \mathbf{V}^{\setminus i} = \mathbf{v}) > 0$ takve da*

$$P(Y = y \,|X_i = x, \mathbf{V}^{\setminus i} = \mathbf{v}) \neq P(Y = y \,|\mathbf{V}^{\setminus i} = \mathbf{v})$$

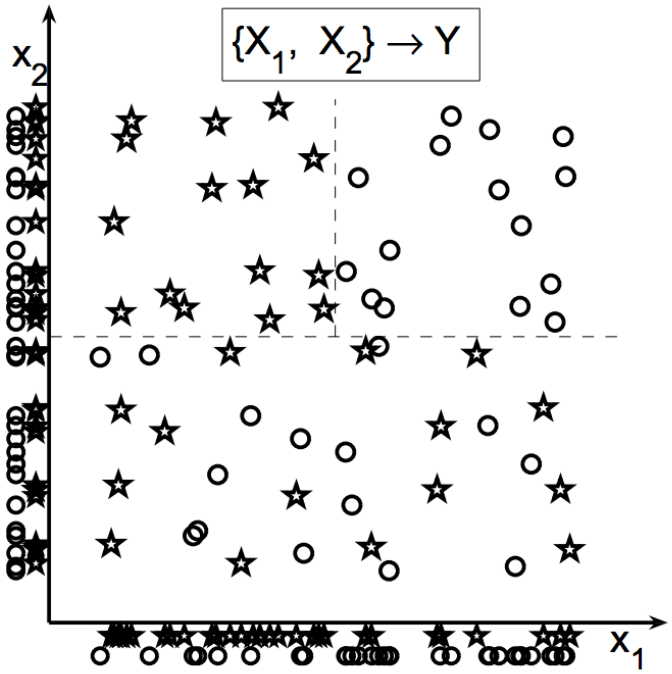
Definicije relevantnosti na ovaj način su opće prihvaćene, a prvi puta su uvedene u [6] Uočimo da iz

$$P(Y|X_i, \mathbf{V}^{\setminus i}) = P(Y|\mathbf{V}^{\setminus i})$$

primjenom Bayesovog pravila slijedi

$$P(X_i, Y|\mathbf{V}^{\setminus i}) = P(X_i \,|\mathbf{V}^{\setminus i}) * P(Y|\, \mathbf{V}^{\setminus i}),$$

pa zaključujemo da je svaka značajka X_i individualno irelevantna za Y, jako relevantna za Y ili slabo relevantna za Y.
Također, napomenimo da ovako definirana relevantnost ne mora nužno značiti da je značajka korisna za predviđanje vrijednosti Y. Ova činjenica je trivijalna za slabo relevantne značajke, ali vrijedi čak i za jako relevantne. Pogledajmo slijedeći primjer:



Slika 2.1: Relevantna ali ne i korisna značajka

Y je kategorijska varijabla koja poprima vrijednosti $+1$ (kružić) odnosno -1 (zvijezda). Iz grafa je jasno da je X_2 jako relevantna za Y. Naime, vrijedi

$$P(Y|\, X_1, X_2) \neq P(Y|\, X_1),$$

ali X_2 ne doprinosi kvaliteti predviđanja: greška od 25% ne može biti popravljena poznavanjem vrijednosti X_2

Iako je kauzalnost svima blizak pojam iz svakodnevice, nije jednostavno dati općenitu definiciju. Da bismo to mogli, ograničiti ćemo razmatranje na dobro definirane sustave koji su izolirani od okoline i nad kojima se mogu vršiti eksperimenti odnosno **manipulacije**. U nekim slučajevima eksperimentiranje nije moguće iz praktičnih ili etičkih razloga, ali u teoriji se može što je važno za opći koncept kauzalnosti.

2.2 Vjerojatnosna kauzalnost

U nekim ne deterministčkim sustavima kao primjerice Markovljevim lancima, isti uzrok može dovesti do različitih posljedica i obratno, različiti uzroci mogu dovesti do iste posljedice. Ovo znači da se ne možemo držati ideje "isti uzrok rezultira istim posljedicama" bez uvođenja vjerojatnosti.

Definicija 2.2.1. ***Manipulacija*** *s oznakom $do(X_i)$ je vanjska intervencija na sustav koja mjenja distribuciju slučajne varijable X_i iz njezine prirodne distribucije u distribuciju nezavisnu od sustava.*

Tipična manipulacija prikazana je u primjeru 1.4.2, gdje su se zaposlenici izmijenili za majice i na taj način duljina rukava majice koju su nosili nije bila ona koju bi nosili "prirodno".

Definicija 2.2.2. *Za slučajnu varijablu (značajku) X_i kažemo da je* ***individualno kauzalno relevantna*** *za (ciljnu) slučajnu varijablu Y ako postoje manipulacija $do(.)$ i vrijednosti x i y za koje vrijedi $P(do(X_i) = x) > 0$ takve da:*

$$P(Y = y \mid\ do(X_i) = x) \neq P(Y = y)$$

Uočimo da uz ovu definiciju nismo u potpunosti opisali kauzalnu relevantnost. Također, iz ove definicije je jasna razlika između kauzalnosti i korelacije [11] Uzmimo primjerice da X_1 i X_2 imaju geometrijsku razdiobu s parametrom $p > 0$ i definirajmo

$$Y = \begin{cases} 1, & X_1 + X_2 \text{ paran} \\ 0, & X_1 + X_2 \text{ neparan} \end{cases}$$

U ovom slučaju niti X_1 niti X_2 nisu individualno kauzalno relevantne za Y, ne postoji manipulacija jedne od varijabli kojom bi promijenili distribuciju od Y. S druge strane, takva zajednička manipulacija na $\mathbf{X}$ postoji.
U nastavku sa $\mathbf{X}^{\setminus i}$ označavamo skup svih slučajnih varijabli iz $\mathbf{X}$ osim X_i

Definicija 2.2.3. *Za slučajnu varijablu (značajku) X_i kažemo da je **jako kauzalno relevantna** za (ciljnu) slučajnu varijablu Y ako postoje manipulacija do(.) i vrijednosti x, y i* $\mathbf{v}$ *za koje vrijedi* $P(do(X_i) = x, do(\mathbf{X}^{\setminus i}) = \mathbf{v}) > 0$ *takve da*

$$P(Y = y \mid do(X_i) = x, do(\mathbf{X}^{\setminus i}) = \mathbf{v}) \neq P(Y = y \mid do(\mathbf{X}^{\setminus i}) = \mathbf{v})$$

Definicija 2.2.4. *Za slučajnu varijablu (značajku) X_i kažemo da je **slabo kauzalno relevantna** za (ciljnu) slučajnu varijablu Y ako nije jako kauzalno relevantna i postoje manipulacija do(.), podskup $\mathbf{V}^{\setminus i} \subset \mathbf{X}^{\setminus i}$ i vrijednosti x, y i* $\mathbf{v}$ *za koje vrijedi* $P(do(X_i) = x, do(\mathbf{V}^{\setminus i}) = \mathbf{v}) > 0$ *takve da*

$$P(Y = y \mid do(X_i) = x, do(\mathbf{V}^{\setminus i}) = \mathbf{v}) \neq P(Y = y \mid do(\mathbf{X}^{\setminus i}) = \mathbf{v})$$

Ovim pojmovima je dobro definirana kauzalna relevantnost, no u praksi se baš i ne koriste. Da bi se utvrdilo za neku značajku da nije kauzalno relevantna, trebalo bi proći po svim skupovima $\mathbf{V}^{\setminus i}$, po svim mogućim vrijednostima x, y i $\mathbf{v}$ i, možda i najmanje realno, po svim mogućim kombinacijama manipulacija $do(X_i)$ i $do(\mathbf{V}^{\setminus i})$.

2.3 Kauzalne Bayesove mreže

Kao i do sada, u ovoj sekciji ćemo slučajne varijable označavati velikim tiskanim slovima (X,Y,Z), a realizacije (vrijednosti) malim (x,y,z). Ciljna varijabla je Y, dok su ostale (značajke) X_i.
Za definiciju Bayesove mreže koristit ćemo **usmjeren aciklički graf** (UAG). Prisjetimo se, radi se o usmjerenom grafu koji ne sadrži cikluse. Ako postoji direktan brid od A do B tada kažemo da je A roditelj od B i da je B dijete od A. Ako postoji put od A do B tada kažemo da je A predak od B i da je B potomak od A.

Definicija 2.3.1. *Neka je* $\mathbf{X}$ *skup diskretnih slučajnih varijabli i P zajednička vjerojatnosna distribucija od* $\mathbf{X}$ *. Neka je* $\mathcal{G}$ *usmjeren aciklički graf takav da su njegovi čvorovi u jedanna-jedan korespondenciji sa elementima od* $\mathbf{X}$ *i takav da vrijedi:*

$$(\forall A \in \mathbf{X})(\forall B \in \mathbf{X})(B \text{ nije potomak od } A) \Rightarrow (B \perp A \mid \mathbf{C})$$

gdje je $\mathbf{C}$ *skup svih roditelja od A.*
Tada uređenu trojku $(\mathbf{X}, \mathcal{G}, P)$ *nazivamo **(diskretna) Bayesova mreža**.*

Drugim rječima, uvjet Bayesove mreže jest da za svaku slučajnu varijablu $A \in \mathbf{X}$ vrijedi da je ona nezavisna od svake druge varijable $B \in \mathbf{X}$ koja joj nije potomak, uvjetno na svoje roditelje. Ovaj uvjet se još naziva i **Markovljevo svojstvo**.

Definicija 2.3.2. ***Kauzalna Bayesova mreža*** *[8] je Bayesova mreža* $(\mathbf{X}, \mathcal{G}, P)$ *za koju vrijedi* $(\forall A \in \mathbf{X})(\forall B \in \mathbf{X})$ *Postoji direktan brid od A do B u* $\mathcal{G} \Rightarrow$ *A uzrokuje B*

Prikaz mreže pomoću usmjerenog acikličkog grafa je koristan jer se lako mogu pročitati nezavisnosti u mreži. Osim toga, obično se podrazumijeva još jedan uvjet koji omogućava da se iz grafa mogu pročitati i zavisnosti.

Definicija 2.3.3. *Za Bayesovu mrežu* $(\mathbf{X}, \mathcal{G}, P)$ *kažemo da je* ***vjerodostojna*** *ako vrijedi*

$$(\forall A \in \mathbf{X}, \forall B \in \mathbf{X}, \forall C \subset \mathbf{X}), A \not\perp_{\mathcal{G}} B \mid C \Rightarrow A \not\perp_P B \mid C$$

Ako Bayesova mreža zadovoljava i uvjet vjerodostojnosti, onda graf $\mathcal{G}$ sigurno točno prikazuje sve zavisnosti i nezavisnosti između varijabli.

2.4 Učenje kauzalne strukture u mreži

Cilj kauzalne analize jest iz zadanih podataka utvrditi strukturu grafa Bayesove mreže, što je moguće napraviti direktno iz opservacija (bez manipulacija i eksperimenata). Opisat ćemo jednu od metoda koje ovo omogućuju, a koja se sastoji od niza testiranja uvjetnih nezavisnosti među varijablama.
Uzmimo zbog jednostavnosti sustav u kojem imamo samo tri slučajne varijable A, B i C. Pogledajmo sve mogućnosti različitih (do na imena varijabli) usmjerenih acikličkih grafova:

1. Potpuno nepovezani graf: A, B, C
2. Jedan brid: $A \rightarrow C, B$ ili $A \leftarrow C, B$
3. **Lanac**: $A \rightarrow C \rightarrow B$ ili $A \leftarrow C \leftarrow B$
4. **Vilica**: $A \leftarrow C \rightarrow B$
5. **V-struktura**: $A \rightarrow C \leftarrow B$
6. Potpuno povezani graf: $A \rightarrow C \rightarrow B$, $A \rightarrow B$

Svaki od ovih grafova odgovara jednoj vjerodostojnoj kauzalnoj Bayesovoj mreži:

1. Potpuno nepovezani graf: $A \perp B$, $B \perp C$ i $C \perp A$

2. Jedan brid: $A \perp B$ i $B \perp C$
3. **Lanac**: $A \perp B \mid C$
4. **Vilica**: $A \perp B \mid C$
5. **V-struktura**: $A \perp B$ ali $A \not\perp B \mid C$
6. Potpuno povezani graf: nema nezavisnosti

Uz ovako ispisane nezavisnosti lako uočavamo:

- U slučaju lanca i jednog brida smjerovi bridova se mogu okrenuti bez promjene uvjetnih nezavisnosti
- Vilica i lanac imaju iste uvjetne nezavisnosti
- U potpuno povezanom grafu bridovi mogu biti okrenuti u bilo kojem smjeru (dok god graf ostaje acikličan)

Zaključujemo da su samo nepovezani graf i V-struktura jednoznačno određeni uvjetnim nezavisnostima. Ovo svojstvo V-strukture omogućava ustanovljavanje kauzalnih veza među varijablama.

Poglavlje 3

Relevantnost značajki u Bayesovim mrežama

Ranije smo definirali što znači da je pojedina značajka slabo odnosno jako relevantna i što znači da je slabo odnosno jako kauzalno relevantna. U Bayesovoj mreži jako relevantne značajke prepoznajemo kao one koje su bridom povezane s ciljnom značajkom, no, kao što smo napomenuli ranije, to ne znači nužno da su one i jako kauzalno relevantne. U ovom poglavlju ćemo razmotriti različite slučajeve koji se mogu dogoditi i utvrditi u kojoj situaciji možemo za značajku zaključiti da je jako kauzalno relevantna.
U nastavku pretpostavljamo da je $\mathbf{X} \cup Y$ skup svih varijabli u razmatranju i $\mathbf{V}$ podskup od $\mathbf{X}$

3.1 Markovljev pokrivač

Definicija 3.1.1. *Za podskup* $\mathbf{M}$ *od* $\mathbf{X}$ *kažemo da je* ***Markovljev pokrivač*** *[7] od* Y *ako za svaki* $\mathbf{V}$ *podskup od* $\mathbf{X}$ *vrijedi da su* Y *i* $\mathbf{V}\backslash\mathbf{M}$ *nezavisne uvjetno na* $\mathbf{M}$*, to jest:*

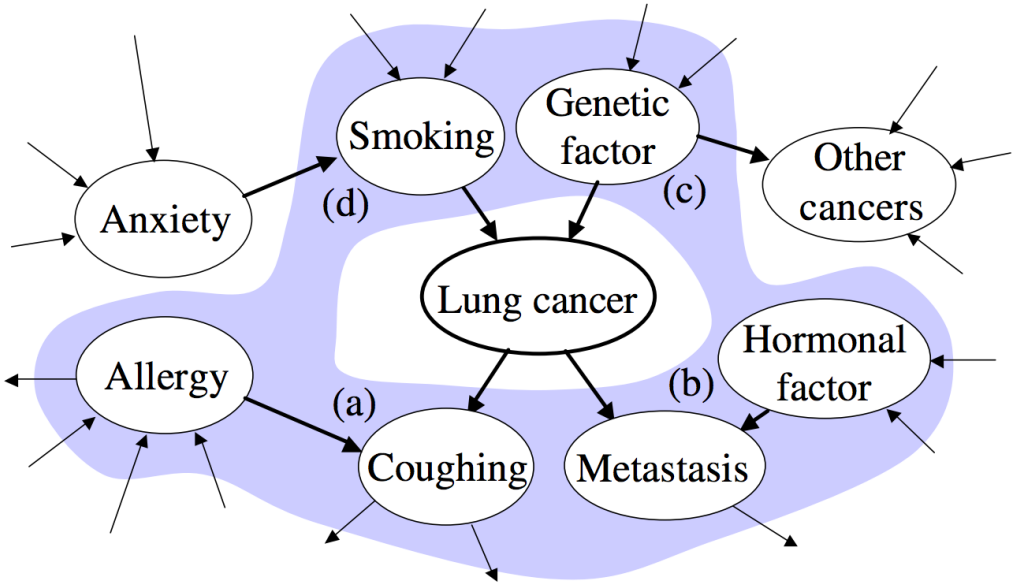
$$(\forall \mathbf{V} \subset \mathbf{X}) P(Y, \mathbf{V}\backslash\mathbf{M} \mid \mathbf{M}) = P(Y, \mid \mathbf{M}) * P(\mathbf{V}\backslash\mathbf{M} \mid \mathbf{M})$$

Uočimo da iz ovoga slijedi

$$P(\mathbf{V}\backslash\mathbf{M} \mid \mathbf{M}) > 0 \Rightarrow P(Y \mid \mathbf{V}\backslash\mathbf{M}, \mathbf{M}) = P(Y \mid \mathbf{M})$$

Općenito, Markovljev pokrivač nije jedinstven, ali u slučajevima kad kauzalna Bayesova mreža zadovoljava Markovljevo svojstvo i uvjet vjerodostojnosti jest. U tom slučaju Markovljev pokrivač od Y sadrži direktne uzroke (roditelje), direktne posljedice (djecu) i roditelje direktnih posljedica (supružnike). Zanimljivo je da ne sadrži direktne posljedice direktnih uzroka (braću) niti direktne uzroke direktnih uzroka (praroditelje).

Da bismo dobili bolje razumijevanje ovog pojma pogledajmo primjer sa slike u kojem je ciljna značajka“rak pluća”:



Slika 3.1: Markovljev pokrivač

U slučaju kad znamo vrijednosti direktnih roditelja, vrijednosti indirektnih roditelja ne daju dodatnu informaciju. Na primjeru sa slike, “anksioznost” *(engl. anxiety)* može povećati “pušenje” *(engl. smoking)*, ali ne utječe na “rak pluća” *(engl. lung cancer)* direktno pa je stoga u svrhu predviđanja dovoljno znati informaciju o pušenju. Slično tome, posljedica direktnog uzroka kao “ostali rakovi” *(engl. other cancers)* ne daje dodatnu informaciju u slučaju kad je “genetski faktor” *(engl. genetic factor)* poznat.
Direktne posljedice su uvijek korisne za predviđanje, no njihova moć predviđanja može se povećati poznavajući direktne uzroke direktnih posljedica. Primjerice, “alergija” *(engl. allergy)* može uzrokovati “kašljanje” *(engl. coughing)* neovisno o raku pluća. Važno je znati informaciju o alergiji jer time možemo objasniti da kašljanje nije uzrokovano rakom pluća. *Supružnici* koji nisu direktno povezani s ciljem individualno ne doprinose predviđanju. Trebaju imati zajedničko *dijete* da bi postali prediktivni.
Sada smo u stanju povezati koncept relevantnosti sa Markovljevim pokrivačem u vjerodostojnoj distribuciji:

- **Irelevantnost:** Značajka je irelevantna ako ne postoji put izmežu nje i cilja u grafu.
- **Jaka relevantnost:** Značajka je jako relevantna za ciljnu značajku Y ako se nalazi u Markovljevom pokrivaču od Y.
- **Slaba relevantnost:** Značajka je slabo relevantna za ciljnu značajku Y ako postoji put između nje i Y, ali se ne nalazi u Markovljevom pokrivaču od Y.

Prva tvrdnja preslikava definiciju 2.1.2 u nepovezanost značajke sa ciljnom, a slijedi direktno iz Markovljevog svojstva grafa. Druga tvrdnja povezuje definiciju 2.1.3 sa Markovljevim pokrivačem. Naime, samo jako relevantne značajke ne mogu biti izbačene bez gubitka prediktivne sposobnosti od $\mathbf{X}$, pa je skup jako relevantnih značajki $\mathbf{M}$ dovoljan za predviđanje Y. Za proizvoljne vrijednosti $\mathbf{v}$ ostalih značajki iz $\mathbf{X}\backslash\mathbf{M}$ vrijedi $P(Y \mid \mathbf{M}) = P(Y \mid \mathbf{M}, \mathbf{X} \backslash \mathbf{M} = \mathbf{v})$ pa je prema definiciji 3.1.1 $\mathbf{M}$ markovljev pokrivač.
Kako je markovljev pokrivač jedinstven u vjerodostojnim distribucijama, sada je jasno da je i skup jako relevantnih značajki jedinstven u vjerodostojnim distribucijama.

3.2 Karakteriziranje značajki odabranih klasičnim metodama

U ovoj sekciji ćemo pokazati da finijom analizom u smislu kauzalnih zavisnosti možemo iskoristiti pojam relevantnosti značajki. Ograničavamo analizu na značajke koje su u direktnoj blizini ciljne, dakle Markovljev pokrivač i neke značajke “blizu” Markovljevog pokrivača. Ovo nam omogućava da analizu svedemo na nekoliko slučajeva:

- Direktan uzrok (roditelj)
- Nepoznati direktan uzrok (zajednički roditelj koji nedostaje može dovesti do tumačenja “brata” kao roditelja)
- Direktan potomak (dijete)
- Nepoznati direktan potomak (zajedničko dijete koje nedostaje može dovesti do tumačenja “supružnika” kao roditelja)
- Ostali članovi Mrakovljevog pokrivača (“supružnici”)
- Šumovi (varijable koje se nalaze u Markovljevom pokrivaču, ali ne želimo ih imati kao dio sustava, npr. greška u mjerenju)

3.3 Strukture *lanca*, *vilice* i *V-strukture*

Pogledajmo prvo značajke direktno vezane za roditelje od ciljne, dakle praroditelje i braću. Oni se pojavljuju u strukturama *lanca* i *vilice* u kojima su zavisnosti oblika $Y \not\perp X_1$, $Y \perp X_1 \mid X_2$.

Kao što smo utvrdili prije, značajke praroditelji i braća nisu dio Markovljevog pokrivača pa nisu korisne za predviđanje, no tvrdimo da su bez obzira na to vrijedne spomena. Konkretno, u slučajevima kada je ne moguće utjecati na direktne uzroke (u primjeru raka pluća "pušenje") može biti korisno pronaći indirektne uzroke (utjecanjem na "anksioznost" možemo utjecati na "pušenje"). Posljedice direktnih uzroka ("braća") mogu biti zanimljive iz drugog razloga: u našem primjeru nije moguće mjeriti "genetski faktor", ali poznavajući vrijednost "ostalih rakova" možemo ga procijeniti.

Problem je to što iz zavisnosti nije moguće zaključiti radi li se o strukturi "lanca" ili "vilice" pa je nemoguće zaključiti radi li se o praroditelju ili bratu. Ovaj se problem ponekad može riješiti poznavajući veze s drugim varijablama, prirode podataka ili eksperimentima. Ako stvarni roditelji nisu poznati, praroditelji postaju najdirektniji uzrok, no lako je moguće zamijeniti brata za praroditelja i protumačiti njega kao uzrok.

Na primjer, u stvarnosti je na temelju korelacije utvrđeno da pušenje uzrokuje rak pluća i stoga su uvedene restrikcije vezane za pušenje na javnim mjestima. Nakon toga su proizvođači duhana tvrdili da to nije nužno istina, nego je moguće i da postoji i zajednički roditelj (npr. genetski faktor) koji uzrokuje i sklonost pušenju i povećanu vjerojatnost oboljevanja od raka pluća, odnosno da su pušenje i rak pluća "braća". Do danas nije utvrđeno da takav faktor doista postoji.

Zavisnosti oblika $(X_2 \perp Y, X_2 \not\perp Y | X_1)$ su karakteristične za **V-strukture**. I djeca i supružnici se pojavljuju u tim strukturama i nalaze se u Markovljevom pokrivaču od ciljne varijable pa su prema tome jako relevantne. Valja naglasiti da nisu "kauzalno" relevantne, u smislu da manipulacije nad njima neće rezultirati promjenom distribucije ciljne varijable, no korisne su za izradu predviđanja kao što je pokazano ranije na primjeru alergije i kašljanja. Osim toga korisni su za eksperimentiranje, primjerice, činjenica da se pacijentu nakon liječenja raka pluća (manipulacije) smanjilo kašljanje može biti indikator o uspješnosti liječenja.

Poglavlje 4

Metode kauzalne analize

Dugo je vladalo uvjerenje da se kauzalne veze mogu ustanoviti isključivo manipulacijama odnosno eksperimentima, no u zadnje vrijeme dosta istraživanja je posvećeno utvrđivanju kauzalnih veza iz opserviranih podataka, to jest, podataka prikupljenih iz sustava bez planiranih eksperimenata i vanjskih utjecaja na sustav.

4.1 Definiranje cilja i pretpostavki

Učenje Bayesove mreže $(\mathbf{X}, \mathcal{G}, P)$ iz podataka sastoji se od dva podzadatka: učenje strukture grafa $\mathcal{G}$ i učenje vjerojatnosne distribucije P. Nama je od interesa učenje strukture grafa $\mathcal{G}$.

Definicija 4.1.1. *Za skup slučajnih varijabli* $\mathbf{X} = \{X_1, X_2, ..., X_n\} \subseteq \mathbf{S}$ *kažemo da je* ***kauzalno samodovoljan*** *ako niti jedan skup od dvije ili više varijabli iz* $\mathbf{X}$ *nema zajedničkog roditelja iz* $\mathbf{X} \setminus \mathbf{S}$

U nastavku ćemo promatrati skup slučajnih varijabli $\mathbf{X}$ uz slijedeće pretpostavke:

- Skup slučajnih varijabli $\mathbf{X}$ je kauzalno samodovoljan
- Postoji dovoljno velik broj podataka da bi se statističkim testovima mogle utvrditi uvjetne zavisnosti i nezavisnosti u distribuciji iz koje podaci potječu
- Proces iz kojeg potječu podaci može se predstaviti kauzalnom Bayesovom mrežom

4.2 Algoritam prototipne kauzalnosti

Osnovna metoda kauzalne analize naziva se **algoritam prototipne kauzalnosti** ***(engl. prototype causality algorithm)*** ili skraćeno **PC algoritam**. Pod gore navedenim pretpos-

tavkama, algoritam je dokazano dobar u smislu da uspješno utvrđuje strukturu Bayesove mreže iz koje potječu podaci.

PC Algoritam:
Inicijaliziraj sa potpuno povezanim ne usmjerenim grafom.

1. Za svaki par varijabli A i B testiraj nezavisnost uvjetno na svaki podskup varijabli od $\mathbf{X} \setminus \{A, B\}$ (uključujući i prazan skup). Ako postoji $\mathbf{V} \subseteq \mathbf{X} \setminus \{A, B\}$ takav da su A i B nezavisne uvjetno na $\mathbf{V}$, ukloni brid između A i B

2. Usmjeri bridove unutar V-struktura (to jest struktura oblika $A \rightarrow C \leftarrow B$) koristeći slijedeći uvjet:
Ako postoje bridovi između A, C i C,B, ali ne između A, B tada vrijedi $A \rightarrow C \leftarrow B$ ako i samo ako ne postoji podskup $\mathbf{V} \subseteq \mathbf{X} \setminus \{A, B\}$ koji sadrži C takav da je $A \perp B \,|V$

3. Usmjeri ostale bridove dok god je moguće koristeći slijedeća dva pravila:

   - Ako $A \rightarrow B \rightarrow ... \rightarrow C$ i $A - C$ tada $A \rightarrow C$
   - Ako $A \rightarrow B - C$ tada $B \rightarrow C$ (u suprotnom bi se radilo o V-strukturi koja bi bila prepoznata u koraku 2)

4.3 Algoritmi indukcije Markovljevog pokrivača

Iz prošle sekcije jasno je da je moguće primijeniti PC algoritam i iščitati Markovljev pokrivač ciljne varijable Y, no ako se radi o velikom sustavu sa desecima tisuća varijabli i još više kauzalnih veza, primjena PC ili sličnog algoritma koji otkriva cijelu mrežu postaje nepraktična. U tim slučajevima koristimo algoritme lokalne kauzalne analize kojima je cilj strukturu samo dijela mreže koji je u neposrednoj blizini od ciljne varijable.
Zadržat ćemo se na dva efikasna algoritma koji su nedavno predloženi. **HITON** [2] algoritam i **MMMB** [10] algoritam. Oba pronalaze direktne bridove uz Y i bridove do potencijalnih supružnika (ne usmjerene) na isti način kao i PC algoritam, osim što HITON i MMMB kreću sa praznim grafom za razliku od PC-a koji kreće sa potpuno povezanim. Osim što traže samo direktne bridove i supružnike, što je već veliko ubrzanje u odnosu na izgradnju cijele mreže, HITON i MMMB algoritmi koriste veliki broj heuristika kako bi ubrzali pretraživanje što se u praksi pokazalo vrlo efikasno. Algoritmi se razlikuju u heuristikama kojima ograničavaju pretraživanje *(rezanje grana)* i tzv. *provjeri zdravlja (engl. sanity check)*. S ovim prilagodbama rješavaju brzo i efikasno probleme i sa 10^5 varijabli. U objavljenim eksperimentima, HITON i MMMB su se pokazali boljima od ostalih predoženih algoritama, ali HITON eliminira neke pogrešno otkrivene *(engl. false positive)* varijable koje MMMB prihvaća i dizajniran je specifično za problem klasifikacije.

Ovi lokalni algoritmi nude značajnu prednost nad globalnim (npr. PC algoritam) u slučajevima kad se radi o tzv. *rijetkim* grafovima u kojima su pojedine regije puno gušće po broju bridova od ostalih. Globalni algoritmi u tom slučaju zahtjevaju izrazito veliku količinu podataka koja često nije dostupna, a inače stvaraju mnogo grešaka koje se onda propagiraju i u ostale regije u mreži. S druge strane lokalni algoritmi brzo i precizno rješavaju problem s rijetkim regijama pa su u takvim slučajevima uspješniji i u konstruiranju kompletnih mreža od PC algoritma [4].
S obzirom na sve do sada rečeno, lokaliziranje pretrage za direktnim bridovima je vrlo korisno, ali izrazito kompleksno algoritamski. Grubo rečeno, kada izrađujemo skup djece/roditelja oko Y često ćemo ubaciti i varijable X_i koje nisu direktno vezane za Y, ali jesu za X_j koje su vezane za Y. Srećom ovaj problem se uspješno rješava ponavljanjem cijelog postupka s tim da uzmemo $Y = X_i$ za varijable X_i za koje mislimo da bi mogle biti vezane za Y.

Poglavlje 5

Primjena kauzalne analize

U slijedećem poglavlju pokazat ćemo primjenu i rezultate kauzalne analize. U tu svrhu koristit ćemo softver **Tetrad** [3], koji sadrži implementacije algoritama koje smo spominjali i pruža pregledno grafičko sučelje, a za simulaciju podataka koristit ćemo softver **R** [9]. U R-u generirane podatke spremamo u *csv* datoteku koju proslijeđujemo Tetradu, koji tada obavlja kauzalnu analizu. Za svaki test (primjerice nezavisnosti u sklopu PC algoritma) uzimamo razinu značajnosti $\alpha = 0.01$

5.1 Nekoliko jednostavnih primjera

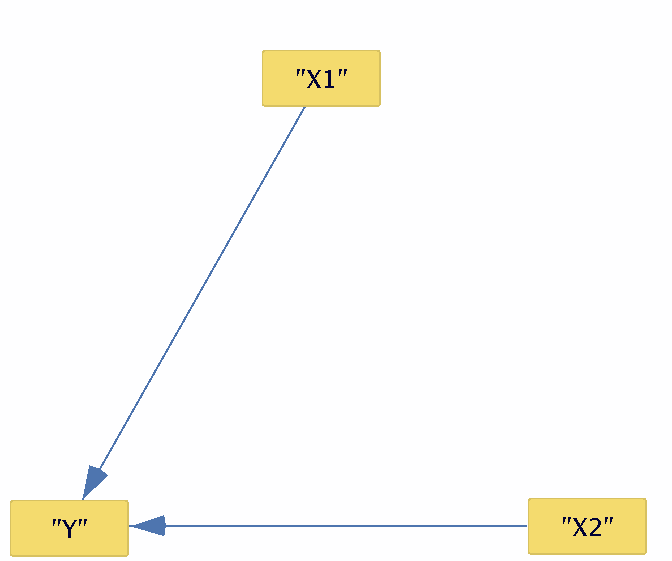
Primjer 5.1.1. *Jednostavan zbroj*

U ovom primjeru simuliramo sustav u kojem imamo dvije nezavisne varijable $X_1 \sim N(0,1)$ i $X_2 \sim N(5,5)$ i varijablu $Y = X_1 + X_2$. Za broj podataka uzimamo $N = 1000$.

```
N <- 1000
simData <- matrix(0,N,3)
simData[,1] <- rnorm(N,0,1)
simData[,2] <- rnorm(N,5,5)
simData[,3] <- simData[,1] + simData[,2]
tmp = c("X1","X2","Y")
write.table(simData,'data1.csv',row.names=FALSE,col.names=tmp,sep=",")
```

Simulacija 5.1: R kod za simulaciju podataka

Iz slike 5.1 vidimo da PC algoritam uspješno pronalazi kauzalne veze između ovih podataka



Slika 5.1: Jednostavan zbroj

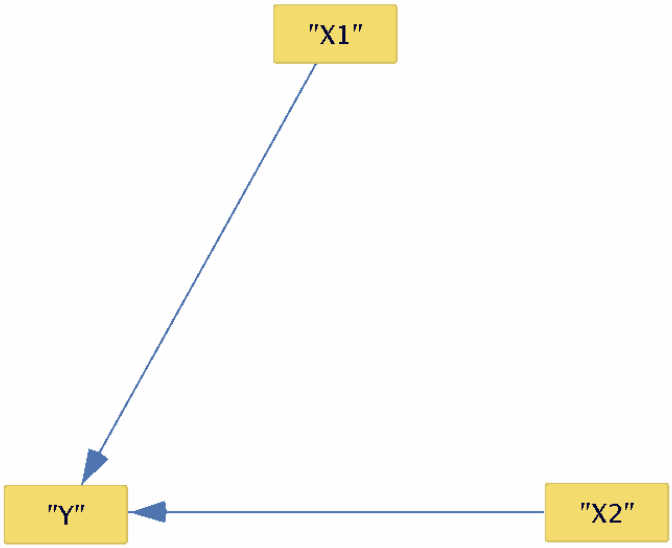
Primjer 5.1.2. *Zbroj sa šumom*

Ovaj primjer je vrlo sličan prethodnom: sustav se sastoji od dvije nezavisne varijable $X_1 \sim N(0,1)$ i $X_2 \sim N(5,5)$ i varijable šuma $W \sim N(0,1)$ koju ne spremamo u podatke i ne predajemo Tetradu. Definiramo varijablu $Y = X_1 + X_2 + W$. Očekujemo da će bez obzira na šum biti prepoznate kauzalne veze $X_1 \to Y$ i $X_2 \to Y$. Za broj podataka ponovo uzimamo $N = 1000$.

```
N <- 1000
simData <- matrix(0,N,3)
simData[,1] <- rnorm(N,0,1)
simData[,2] <- rnorm(N,5,5)
simData[,3] <- simData[,1] + simData[,2] + rnorm(N,0,1)
tmp = c("X1","X2","Y")
write.table(simData,'data2.csv',row.names=FALSE,col.names=tmp,sep=",")
```

Simulacija 5.2: R kod za simulaciju podataka

Vidimo da je PC algoritam i u ovom slučaju dao očekivani rezultat: šum nije stvorio problem u prepoznavanju kauzalnih veza.



Slika 5.2: Zbroj sa šumom

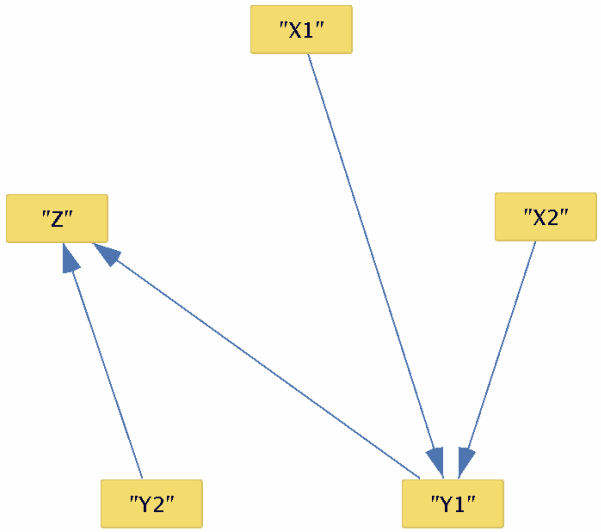
Primjer 5.1.3. *Kompleksniji sustav*

U ovom primjeru promatrat ćemo nešto kompleksniji sustav: imamo varijable $X_1 \sim N(0,1)$ i $X_2 \sim N(5,5)$ i varijablu $Y_1 = X_1 + X_2 + W_1$, gdje je W_1 varijabla šuma ($W_1 \sim N(0,1)$). Nadalje imamo varijablu $Y_2 \sim N(-5,2)$ i varijablu $Z = Y_1 + Y_2 + W_2$, gdje je $W_2 \sim N(0,1)$ varijabla šuma (nezavisna od W_1). Kako se sustav sastoji od dvije *V-strukture*, očekujemo da će PC algoritam prepoznati sve kauzalne veze koje postoje u sustavu.

```
N <- 2000
simData <- matrix(0,N,5)
simData[,1] <- rnorm(N,0,1)
simData[,2] <- rnorm(N,5,5)
simData[,3] <- simData[,1] + simData[,2] + rnorm(N,0,1)
simData[,4] <- rnorm(N,-5,2)
simData[,5] <- simData[,3] + simData[,4] + rnorm(N,0,1)
tmp = c("X1","X2","Y1","Y2","Z")
write.table(simData,'data3.csv',row.names=FALSE,col.names=tmp,sep=",")
```

Simulacija 5.3: R kod za simulaciju podataka

Na slici 5.3 vidimo da je i u ovom slučaju PC algoritam prepoznao sve kauzalne veze kao što smo i očekivali. Uočimo da smo u ovom primjeru za broj podataka uzeli $N = 2000$. Naime, za $N = 1000$ PC algoritam je davao netočne rezultate. Kako vidimo da je testiranjem sa $N = 2000$ rezultat dobar, zaključujemo da u prvom slučaju nije bila zadovoljena pretpostavka o dovoljnom broju podataka. Općenito, teško je reći koji je broj podataka *dovoljan* da bi algoritam uspješno pronašao sve uvjetne zavisnosti odnosno nezavisnosti među varijablama.



Slika 5.3: Kompleksniji sustav

5.2 Potencijalni problemi PC algoritma

U svakom od primjera koje smo naveli do sad, sustav se sastojao od *V-struktura*, koje omogućuju PC-algoritmu pronalaženje kauzalnih veza kao što smo opisali u sekciji 4.2, no to ne mora uvijek biti slučaj.
Promotrimo slijedeći primjer:

Primjer 5.2.1. *Sustav sa zajedničkim roditeljem*

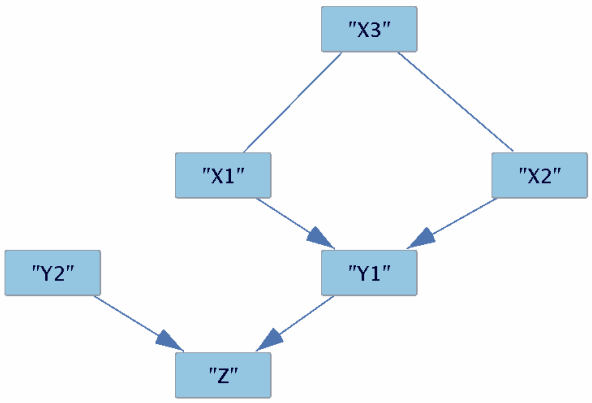
Neka su varijable zadane na slijedeći način:
$W_1, W_2, W_3, W_4 \sim N(0, 1)$ nezavisne varijable šuma,
$X_3 \sim N(2, 2)$,

$X_1 = 2 * X_3 + W_1$,
$X_2 = 5 * X_3 + W_2$,
$Y_1 = X_1 + X_2 + W_3$,
$Y_2 \sim N(-5, 2)$,
$Z = Y_1 + Y_2 + W_4$
U tom slučaju, varijable X_1 i X_2 su obje direktna posljedica varijable X_3, no PC-algoritam opisan u 4.2 to ne može ustanoviti. Naime, promatrajući samo uvjetne zavisnosti odnosno nezavisnosti nije moguće definitivno utvrditi kauzalne veze unutar trojke (X_1, X_2, X_3), jer je moguće da se ladi o strukturi *lanca* ili *vilice*. U ovom slučaju da bismo sa sigurnošću utvrdili točne kauzalne veze nužno je poznavanje prirode podataka ili mogućnost manipulacije.

```
N <- 200000
simData <- matrix(0,N,6)
simData[,6] <- rnorm(N,2,2)
simData[,1] <- 2*simData[,6] + rnorm(N,0,1)
simData[,2] <- 5*simData[,6] + rnorm(N,0,1)
simData[,3] <- simData[,1] + simData[,2] + rnorm(N,0,1)
simData[,4] <- rnorm(N,-5,2)
simData[,5] <- simData[,3] + simData[,4] + rnorm(N,0,1)
tmp = c("X1","X2","Y1","Y2","Z","X3")
write.table(simData,'data4.csv',row.names=FALSE,col.names=tmp, sep=",")
```

Simulacija 5.4: R kod za simulaciju podataka

Uočimo da smo u ovom primjeru uzeli $N = 200000$ zato da bi bili sigurni da se greška na koju želimo ukazati nije pojavila zbog nedovoljne količine podataka.



Slika 5.4: Zajednički roditelj

Iz slike 5.4 jasno vidimo da PC-algoritam nije utvrdio smjerove veza $X_3 - X_1$ i $X_3 - X_2$ kao što smo i očekivali. Zaključak je da samo iz opservacija nije uvijek moguće utvrditi sve kauzalne veze, najčešće one koje se nalaze na rubu mreže.
Drugi slučaj koji bi bilo dobro izdvojiti je kada nije zadovoljena pretpostavka o **samodovoljnosti** (definicija 4.1.1) sustava. Naime, algoritmi spomenuti u poglavlju 4 pretpostavljaju da je sustav samodovoljan. U slučaju da nije tako, može doći do raznih pogrešnih zaključaka kao što ćemo vidjeti u primjeru.

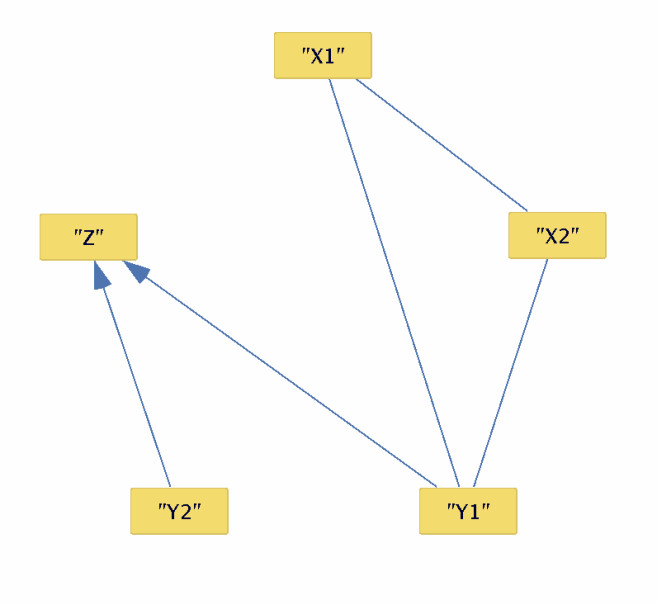
Primjer 5.2.2. *Ne samodovoljan sustav.*

Neka je sustav pod razmatranjem isti kao u prošlom primjeru, dakle:
$W_1, W_2, W_3, W_4 \sim N(0,1)$ nezavisne varijable šuma,
$X_3 \sim N(2,2)$,
$X_1 = 2 * X_3 + W_1$,
$X_2 = 5 * X_3 + W_2$,
$Y_1 = X_1 + X_2 + W_3$,
$Y_2 \sim N(-5,2)$,
$Z = Y_1 + Y_2 + W_4$,
ali u ovom slučaju nećemo zapisati podatke o X_3. Simulaciju ponovo izvodimo za $N = 200000$.

```
N <- 200000
simData <- matrix(0,N,5)
confounder <- rnorm(N,2,2)
simData[,1] <- 2*(confounder) + rnorm(N,0,1)
simData[,2] <- 5*(confounder) + rnorm(N,0,1)
simData[,3] <- simData[,1] + simData[,2] + rnorm(N,0,1)
simData[,4] <- rnorm(N,-5,2)
simData[,5] <- simData[,3] + simData[,4] + rnorm(N,0,1)
tmp = c("X1","X2","Y1","Y2","Z")
write.table(matrix, 'data5.csv', row.names=FALSE, col.names=tmp, sep=",")
```

Simulacija 5.5: R kod za simulaciju podataka

Rezultate dobivene PC-algoritmom vidimo na slici 5.5.



Slika 5.5: Nepoznati zajednički roditelj

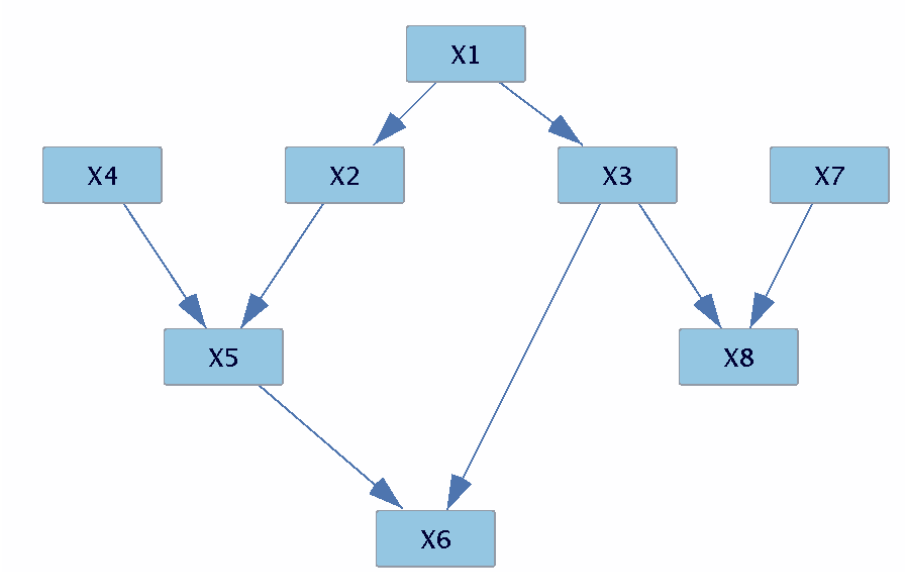
U ovom slučaju nisu utvrđeni niti smjerovi veza $X_1 - Y_1$ i $X_2 - Y_1$. Razlog tome jest to što bez poznavanja varijable X_3 iz prošlog primjera testiranjem dobivamo zavisnost između X_1 i X_2 pa struktura unutar trojke (X_1, X_2, Y_1) više nije *V-struktura* i algoritam ne može utvrditi smjerove niti jedne od veza. Za rješavanje ovakvih problema koriste se neki sofisticiraniji algoritmi od PC-algoritma, kao što je FCI-algoritam *(engl. Fast Causal Inference algorithm)*

5.3 Primjena lokalnih algoritama

Za početak, simulirat ćemo mali sustav u na kojem ćemo primijeniti MMMB algoritam spomenut u poglavlju 4 implementiran u paketu **MXM** (https://CRAN.R-project.org/package=MXM) programskog jezika R [9]. Algoritam radi na način da pronalazi direktne uzroke i posljedice ciljne varijable i njih ubacuje u skup varijabli koje čine Markovljev pokrivač. Nakon toga isti postupak ponavlja za varijable koje su već u Markovljevom pokrivaču i testira jesu li njihovi direktni uzroci i posljedice "supružnici" ciljne varijable. Ako jesu dodaje ih u Markovljev pokrivač a u suprotnom ih odbacuje.

Primjer 5.3.1. *Primjena MMMB algoritma na simuliranim podacima.*

Promatramo sustav prikazan slikom 5.6.



Slika 5.6: Simulirani sustav

Sve slučajne varijable u sustavu su neprekidne normalno distribuiranie ili linearna kombinacija svojih roditelja i bijelog šuma ($N(0,1)$ distribucija).

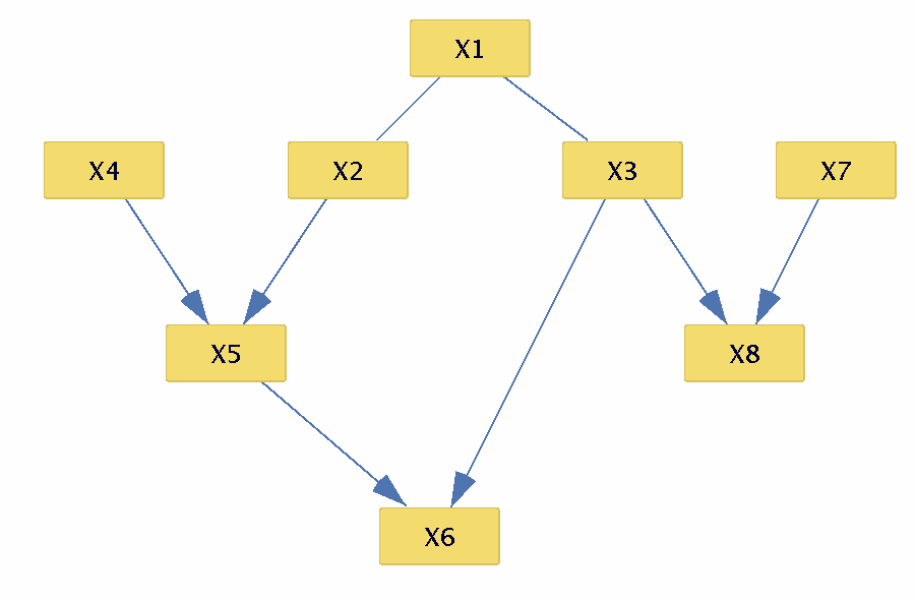
```
N<-10000
data <- matrix(0,N,8)
data[,1] <- rnorm(N,5,1)
data[,2] <- 2* data[,1] + rnorm(N,0,1)
data[,3] <- 3* data[,1] + rnorm(N,0,1)
data[,4] <- rnorm(N,-5,1)
data[,5] <- 2* data[,2] + 1.5* data[,4] + rnorm(N,0,1)
data[,6] <- -2* data[,5] + 4* data[,3] + rnorm(N,0,1)
data[,7] <- rnorm(N,0,3)
data[,8] <- data[,3] + 1.5* data[,7] + rnorm(N,0,1)
target <- 1
write.table(data, 'data_mmmb.csv', row.names=FALSE, sep=",")
```

Simulacija 5.6: R kod za simulaciju podataka

Na sustav je prvo primijenjen PC algoritam, također iz **MXM** paketa programskog jezika R. Algoritam daje sljedeću matricu povezanosti:

$$M = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Što odgovara sustavu iz kojeg podaci potječu. Usmjereni graf dobiven PC algoritmom prikazan je na slici 5.7:



Slika 5.7: Simulirani sustav

Vidimo da se dogodio isti problem kao i u primjeru 5.2.1: PC algoritam ne može utvrditi smjerove veza $X_1 - X_2$ i $X_1 - X_3$, što smo mogli i očekivati.
Na ovom sustavu isprobali smo i MMMB algoritam na nekima od varijabli sa donekle razočaravajućom učinkovitošću. Za varijablu X_1 MMMB algoritam prepoznaje Markovljev pokrivač kao skup $\{X_2, X_3, X_5, X_6\}$, dok bi točno rješenje trebalo biti $\{X_2, X_3\}$. Ova greška može se donekle objasniti činjenicom da algoritam ne može otkriti smjerove veza $X_1 - X_2$ i $X_1 - X_3$. S druge strane za Markovljev pokrivač varijable X_8 algoritam vraća skup

$\{X_1, X_3, X_6, X_7\}$, a za varijablu X_5 skup $\{X_1, X_2, X_3, X_4, X_6\}$. Ovi rezultati nisu obećavajući, ali opet, grešku možemo opravdati time što nije moguće zaključiti odnose između X_1, X_2 i X_3.
Zaključujemo da na ovako malim sustavima MMMB algoritam ne daje najtočnija rješenja: ima isti problem kao i PC algoritam. S druge strane, sustav koji smo simulirali je relativno malen. U većim sustavima ovaj problem se ne bi događao za varijable koje nisu na rubu mreže. U našem simuliranom sustavu, gotovo sve varijable su na rubu mreže.

5.4 Primjena na stvarnim podacima

U svrhu primjene lokalnih algoritama koristit ćemo REGED *(engl. REsimulated Gene Expression Dataset)* [1] skup podataka koji je napravljen u svrhu natjecanja *Causation and Prediction Challenge*. Podaci su generirani iz modela treniranog na stvarnim podacima. Iz perspektive kauzalne analize, važno je razlučiti aktivnost kojeg gena uzrokuje rak pluća, a aktivnost kojeg je posljedica bolesti.
Imamo tri skupa podataka: REGED0, REGED1 i REGED2. Svaki od njih sastoji se od niza od 999 slučajnih varijabli (nema skrivenih varijabli niti nepoznatih podataka) i binarne ciljne varijable te istog trening uzorka duljine 500. Razlika je u tome što u skupu REGED0 testni uzorak dolazi iz iste distribucije kao i trening uzorak, dok su u skupu REGED1 neke varijable manipulirane u testnom uzorku i sudionicima natjecanja je poznato koje su to varijable, a u REGED2 je vrlo velik broj varijabli manipuliran.
Zbog ovako velikog broja varijabli te trening uzorka relativno male duljine, PC i slični algoritmi su neupotrebljivi. Također, skup podataka ne zadovoljava neke od pretpostavki nužnih za korištenje PC algoritam, kao što su **samodovoljnost** sustava i **vjerodostojnost** distribucije. Natjecatelji su na temelju trening uzorka gradili svoj model i pomoću njega stvarali predikcije za testni uzorak. Najviše korišteni pristupi bili su:

- **Utvrđivanje kauzalnih veza** u blizini ciljne varijable i stvaranje predviđanja na temelju vrijednosti varijabli blizu ciljne
- **Utvrđivanje Markovljevog pokrivača** i stvaranje predviđanja na temelju vrijednosti varijabli iz njega, bez pokušaja utvrđivanja kauzalnih veza među varijablama.
- **Izbor značajki**, odnosno utvrđivanje varijabli koreliranih sa ciljnom bez pokušaja utvrđivanja Markovljevog pokrivača ili kauzalnih veza.

Utvrđivanje kauzalnih veza svodi se na utvrđivanje lokalnog usmjerenog grafa u "susjedstvu" ciljne varijable pomoću niza testova uvjetne nezavisnosti. Ovisno o setu podataka, nije se koristio uvijek isti model. U slučaju REGED0, sve varijable za koje je utvrđeno da pridonose predviđanju korištene su, dok u slučaju REGED1 korištene samo one od njih

koje nisu manipulirane. U slučaju REGED2 korišteni su samo direktni uzroci ciljne varijable.

Utvrđivanje Markovljevog pokrivača može se protumačiti kao podzadatak prošlog pristupa pa je teško očekivati da će dati bolje rezultate. U slučaju kada je poznat cijeli Markovljev pokrivač, taj skup varijabli je dovoljan za predviđanje ciljne varijable u slučajevima kada nema manipulacija, dok je u slučaju manipuliranih podataka kao i u prošlom odjeljku nužno limitirati se na roditelje ciljne varijable, nemanipuliranu djecu i nemanipulirane supružnike. U slučaju REGED2 gdje nije poznato koje su varijable manipulirane, optimalno je rješenje u predviđanju koristiti samo roditelje. Zbog toga korištenje cijelog Markovljevog pokrivača za predviđanje u svim testnim uzorcima nije optimalno, ali bez obzira na to, neki su se od natjecatelja odlučili za baš ovaj pristup.

Izbor značajki se u prošlosti pokazao kao iznimno uspješna ideja u raznim problemima pa je stoga poznat širok spektar metoda koje rješavaju ovakav zadatak. U teoriji ovaj pristup nema nikakvo opravdanje iz aspekta kauzalnosti, osim toga što u nekim slučajevima mogu prilično točno procijeniti Markovljev pokrivač varijable. Neki natjecatelji su koristili prilagođene metode izbora značajki i dobili iznenađujuće dobre rezultate.

Iako je natjecanje završeno, još uvijek je dostupna web stranica za slanje rezultata pa smo pokušali usporediti predviđanja na temelju Markovljevog pokrivača (MMMB algoritam spomenut u poglavlju 4), jedne od sofisticiranijih klasičnih metoda izbora značajki: **algoritma slučajne šume** (*engl. Random Forest Algorithm*) i jednom jednostavnom metodom izbora značajki: **algoritam dobitka informacija** (*engl. Information Gain Algorithm*). U prvom slučaju smo uzeli značajke koje pripadaju Markovljevom pokrivaču i njih ubacili u model, dok smo u drugom i trećem uzeli "najvažnije" varijable koje je izbacio algoritam i njih koristili kao model za konstrukciju predviđanja. Predviđanja smo konstruirali linearnom regresijom.

Tablica 5.1: Indeksi izabranih varijabli

Information Gain	Random Forest	MMMB
21	21	26
83	83	83
102	251	102
126	277	251
251	305	312
321	321	321
344	344	344
362	362	409
409	409	410
410	410	425
425	425	453
453	453	457
457	471	471
471	495	495
504	504	556
556	556	561
593	571	593
594	593	594
601	594	712
739	601	739
757	739	825
804	825	897
825	930	930
930	939	939
939	983	983

MMMB algoritam je svrstao 25 varijabli u Markovljev pokrivač pa smo njih usporedili sa 25 najznačajnijih koje su odredili Random Forest odnosno Information Gain algoritmi. Uočimo da MMMB ima 18 zajedničkih varijabli i sa Information Gain i sa Random Forest algoritmom.

```
trainData <- read.table('reged0_text/reged0_train.data')
target <- read.table('reged0_text/reged0_train.targets')
target <- target$V1
result <- mmmb(target,trainData)
blanket <- result$mb
```

```
dftrain <- data.frame(trainData[,blanket])
fit <- lm(target ~., dftrain)

testData <- read.table('reged0_text/reged0_test.data')
dftest <- data.frame(testData[,blanket])
result <- predict(fit, dftest)
binaryResults <- array(-1,20000)
binaryResults[result > 0] <- 1
write.table(binaryResults, 'reged0_test.predict',row.names=FALSE, col.
    ↪ names=FALSE)
```

Simulacija 5.7: R kod za predikciju MMMB algoritmom

```
trainData <- read.table('reged0_text/reged0_train.data')
target <- read.table('reged0_text/reged0_train.targets')
target <- target$V1
ftrain <- data.frame(trainData)
res <- randomForest(target~., trainData)
blanket <- order(res$importance)[975:999]
dftrain <- data.frame(trainData[,blanket])
fit <- lm(target ~., dftrain)


testData <- read.table('reged0_text/reged0_test.data')
dftest <- data.frame(testData[,blanket])
result <- predict(fit, dftest)
binaryResults <- array(-1,20000)
binaryResults[result > 0] <- 1
write.table(binaryResults, 'reged0_test.predict',row.names=FALSE, col.
    ↪ names=FALSE)
```

Simulacija 5.8: R kod za predikciju Random Forest algoritmom

```
trainData <- read.table('reged0_text/reged0_train.data')
target <- read.table('reged0_text/reged0_train.targets')
target <- target$V1
ftrain <- data.frame(trainData)
res <- attrEval(target~., trainData,"InfGain")
blanket <- order(res)[975:999]
dftrain <- data.frame(trainData[,blanket])
fit <- lm(target ~., dftrain)

testData <- read.table('reged0_text/reged0_test.data')
dftest <- data.frame(testData[,blanket])
result <- predict(fit, dftest)
binaryResults <- array(-1,20000)
binaryResults[result > 0] <- 1
```

```
15 write.table(binaryResults, 'reged0_test.predict',row.names=FALSE, col.
    ↪ names=FALSE)
```

Simulacija 5.9: R kod za predikciju Information Gain algoritmom

Rezultati koje smo dobili bili su manje više očekivani: za početak, Random Forest algoritam daje bolje rezultate nego Information Gain na svakom od uzoraka. Nadalje, rezultati su vrlo bliski u sva tri slučaja, što smo očekivali jer se varijable izabrane ovim algoritmima u većoj mjeri poklapaju. Procjene na temelju Markovljevog pokrivača nešto su lošije na uzorcima REGED0 i REGED1, ali bolje u slučaju REGED2. Nismo očekivali veliku razliku jer nismo eliminirali manipulirane varijable iz predviđanja niti pokušavali utvrditi koje varijable iz Markovljevog pokrivača su direktni uzroci te pomoću njih radili predviđanja u slučaju REGED2. Ipak, činjenica da ovako sofisticiran algoritam izbora značajki daje bliske rezultate kao i Markovljev pokrivač daje nam do znanja da varijable iz Markovljevog pokrivača uistinu imaju značajnu prediktivnu moć.

	Information Gain	Random Forest	MMMB
REGED0	0.9720	0.9738	0.9684
REGED1	0.6723	0.6763	0.6475
REGED2	0.5441	0.5231	0.5500

Bibliografija

[1] *Causality Workbench*, http://www.causality.inf.ethz.ch/challenge.php?page=datasets, [Online; accessed 07-September-2017].

[2] Constantin F Aliferis, Ioannis Tsamardinos i Alexander Statnikov, *HITON: a novel Markov Blanket algorithm for optimal variable selection*, AMIA Annual Symposium Proceedings, sv. 2003, American Medical Informatics Association, 2003, str. 21.

[3] Carnegie Mellon University, Pittsburgh, PA, *The Tetrad Project*, http://www.phil.cmu.edu/tetrad/index.html.

[4] Nir Friedman, Iftach Nachman i Dana Peér, *Learning bayesian network structure from massive datasets: the «sparse candidate «algorithm*, Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1999, str. 206–215.

[5] Isabelle Guyon, Constantin Aliferis i André Elisseeff, *Causal feature selection*, Computational methods of feature selection (2007), 63–82.

[6] Ron Kohavi i George H John, *Wrappers for feature subset selection*, Artificial intelligence **97** (1997), br. 1-2, 273–324.

[7] Judea Pearl, *Causality: models, reasoning and inference*, Econometric Theory **19** (2003), br. 675-685, 46.

[8] Peter Spirtes, Clark N Glymour i Richard Scheines, *Causation, prediction, and search*, MIT press, 2000.

[9] The R Foundation for Statistical Computing, *R Project*, https://www.r-project.org/.

[10] Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov i Er Statnikov, *Algorithms for Large Scale Markov Blanket Discovery.*, FLAIRS conference, sv. 2, 2003, str. 376–380.

[11] Martin T Wells, *Computation, Causation and Discovery*, Journal of the American Statistical Association **95** (2000), br. 451, 1019–1019.

Sažetak

Cilj izbora značajki jest utvrđivanje podskupa varijabli $X_1, X_2, ...$ korisnih za predviđanje Y. Iz perspektive kauzalnih veza, smisao relevantnosti varijable može biti poboljšan. Konkretno, uzroci su bolje mete vanjskih utjecaja od posljedica: ako je X_i uzrok od Y, manipulacija X_i manifestirat će se na vrijednostima od Y, ali ne i ako je X_i posljedica od Y. U jeziku Bayesovih mreža, roditelji (uzroci), djeca (posljedice) i supružnici (drugi uzroci direktnih posljedica) su članovi Markovljevog pokrivača iz čega slijedi da su jako relevantni u smislu definicije 2.1.3, u vjerodostojnim distribucijama. Direktni uzroci su jako kauzalno relevantni. Supružnici nisu individualno relevantni u smislu definicije 2.1.2, ali roditelji i djeca jesu, u vjerodostojnim distribucijama. I uzroci i posljedice pridonose predviđanju Y, ali djeca se ponekad mogu objasniti drugim urzocima posljedica od Y (supružnicima od Y), pa se puna prediktivna moć djece ne može iskoristiti bez poznavanja vrijednosti supružnika. Uzroci i posljedice imaju različitu prediktivnu moć u slučaju kada dođe do promjene distribucije u sustavu, ovisno o promjeni. Konkretno, uzroci bi trebali imati veću moć predviđanja od posljedica ako se varijablama $X_1, X_2, ...$ doda nova nepoznata varijabla šuma. U slučaju da se nepoznati šum doda varijabli Y, varijable posljedice su bolji izbor. Nepoznate varijable kao što su greške u mjerenju i zajednički roditelji nekih od varijabli u sustavu mogu dovesti do potpunog neuspjeha kauzalne analize ako se zanemari njihovo potencijalno postojanje. Kauzalna analiza može pomoći u osmišljanju novih eksperimenata kojima bi se dodatno razjasnila relevantnost značajki.

Summary

Feature selection focuses on uncovering subsets of variables $X_1, X_2, ...$ predictive of a target Y. In light of causal relationships, the notion of variable relevance can be refined. In particular, causes are better targets of action of external agents than effects: if X_i is a cause of Y, manipulating it will have an effect on Y, not if X_i is a consequence (or effect). In the language of Bayesian networks, direct causes (parents), direct effects (children), and other direct causes of the direct effects (spouses) are all members of the Markov blanket. The members of the Markov blanket are strongly relevant in the sense of definition 2.1.3, in faithful distributions. Direct causes are strongly causally relevant. Spouses are not individually relevant in the sense of definition 2.1.2, but both parents and children are, in faithful distributions. Both causes and consequences of Y are predictive of Y , but consequences can sometimes be "explained away" by other causes of the consequences of Y . So the full predictive power of children cannot be harvested without the help of spouses. Causes and consequences have different predictive power when the data distribution changes, depending on the type of change. In particular, causal features should be more predictive than consequential features, if new unknown "noise" is added to the variables $X_1, X_2, ...$ If new unknown noise is added to Y however, consequential variables are a better choice. Unknown features, including possible artifacts or confounders, may cause the whole scaffold of causal feature discovery to fall apart if their possible existence is ignored. Causal feature selection method can assist the design of new experiments to disambiguate feature relevance.

Životopis

Rođen sam u Zagrebu, 01. 07. 1994. godine. U rujnu 2000. godine započinjem svoje školovanje u Osnovnoj školi Dragutina Tadijanovića. Nakon završene osnovne škole, u rujnu 2008. upisujem Zagrebačku XV. gimnaziju. Kroz osnovnu i srednju školu bio sam uglavnom odličan učenik, a 2008. i 2011. sam i sudjelovao na državnom natjecanju iz matematike. 2012. godine upisujem Prirodoslovno matematički fakultet Sveučilišta u Zagrebu.
Kroz studij sam prolazio s izvrsnim uspjehom, a od rujna 2014. uz studij i radim kao razvojni programer mobilnih aplikacija u firmi Implementacija Snova.
Tokom studija držao sam demonstrature iz kolegija Linearna algebra 1 i 2 i Diskretna matematika.
U akademskim godinama 2015./2016. i 2016./2017. bio sam dobitnik stipendije za izvrsnost Sveučilišta u Zagrebu.