Obravnava manjkajočih vrednosti in neenakomerne porazdelitve vrednosti ciljne spremenljivke

Ljupčo Todorovski
Univerza v Ljubljani, Fakulteta za upravo

Junij 2018

Pregled predavanja

Manjkajoče vrednosti

- Brisanje primerov in spremenljivk
- Nadomeščanje (*imputation*) manjkajočih vrednosti

Neenakomerne porazdelitve vrednosti ciljne spremenljivke

- Težava z neenakomernimi porazdelitvami
- Rešitev: podvzorčenje in prevzorčenje
- Metodi SMOTE in SMOTER

Brisanje primerov

Standardni prejem za manjkajoče vrednosti ciljne spremenljivke

Analiza popolnih primerov (*Complete-Case Analysis*)

Iz podatkovne množice odstranimo vse **primere** z manjkajočimi vrednostmi katerekoli spremenljivke.

Dve težavi

- Uvedemo predsodek v podatkih in naučenih modelih
- Podatkovna množica ima premalo primerov za učenje

Brisanje napovednih spremenljivk

Analiza razpoložljivih podatkov (*Available-Case Analysis*)

Iz podatkovne množice odstranimo vse **napovedne spremenljivke** z manjkajočimi vrednostmi.

Dve težavi

- Uvedemo predsodek v podatkih in naučenih modelih
- Odstranimo tudi **pomembne** napovedne spremenljivke

Nadomeščanje s povprečjem

Manjkajoče vrednosti spremenljivke nadomestimo z njenim povprečjem njenih znanih vrednosti; težava: zmanjšanje variance

Povprečje **numerične** spremenljivke X v podatkovni množici S

$$m_X = \frac{1}{|\{e \in S : X_e \in D_X\}|} \sum_{e \in S : X_e \in D_X} X_e$$

Povprečje **diskretne** spremenljivke X v podatkovni množici S

$$m_X = \arg\max_{v \in D_X} |\{e \in S : X_e = v\}|$$

Znana vrednost spremenljivke X, ki je najbolj zastopana v množici S

Nadomeščanje z novo vrednostjo in/ali spremenljivko

Diskretna spremenljivka X

Neznane vrednosti nadomestimo z novo vrednostjo $v \notin D_X$.

Numerična spremenljivka X

- Vpeljemo novo spremenljivko $M_X = I(X = NaX)$
- Neznane vrednosti X nadomestimo s povprečjem (ali 0)

Naključno nadomeščanje (dejanje iz obupa)

Vsako manjkajočo vrednost spremenljivke X nadomestimo z naključno izbrano znano vrednostjo X.

Nadomeščanje z najbližjimi sosedi

Za izbran primer z manjkajočimi vrednostmi

1. Poiščemo k najbližjih sosedov primera z znanimi vrednostmi
2. Izračunamo povprečne vrednosti spremenljivk v množici sosedov
3. Nadomestimo neznano vrednost X v izbranem primeru s izračunano povprečno vrednostjo X

Težava: nastavitev parametra k

Nadomeščanje z napovednimi modeli: iterativni postopek

Prva iteracija

Uporabimo enostavno metodo nadomeščanja.

Nadaljnje iteracije

1. Izberemo spremenljivko X (običajno izbiramo v nekem vrstnem redu)
2. Naučimo se napovedni model m za ciljno spremenljivko X, kjer vse druge spremenljivke napovedne
3. Nadomestimo neznane vrednosti X z napovedmi modela m

Ustavitvena pogoja

- Maksimalno število iteracij
- Nespremenjene vrednosti spremenljivk

Algoritem TDIDT obvladuje neznane vrednosti

Razdelitev množice na osnovi spremenljivke X

Primeri z neznano vrednostjo X se enakomerno porazdelijo med vsemi nasledniki vozlišča.

V čem je težava: predsodek napovednih modelov

Povečan predsodek

- Točne napovedi za pogosto opazovane vrednosti manjšajo napako
- Zato napovedi pogostih vrednosti ciljne spremenljivke bolj pomembne

Klasifikacija

Bolj pogosto napovedovanje večinskega razreda.

Regresija

Bolj točne napovedi za pogosto opazovane vrednosti.

Rešitev: vzorčenje podatkovne množice

Podvzročenje (*undersampling*)

Iz vzorca brišemo predstavnike večinskega razreda.

Prevzročenje (*oversampling*)

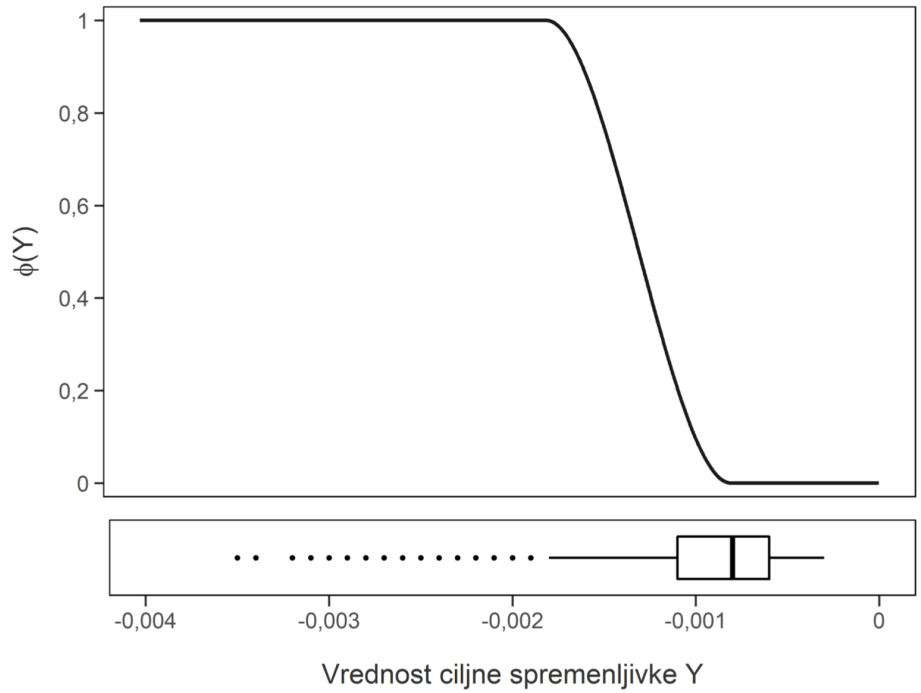
V vzorcu ponavljamo primere iz manjšinskega razreda.

Kaj pa v primeru regresije?

Regresija: funkcija pomembnosti

$$\phi : Y \to [0, 1]$$

- Primere z visoko pomembnostjo $\{(\boldsymbol{x}, y) \in S : \phi(y) > t\}$ prevzorčimo
- Primere z nizko pomembnostjo $\{(\boldsymbol{x}, y) \in S : \phi(y) \leq t\}$ podvzorčimo



SMOTE: osnovna ideja in parametri

SMOTE=*Synthetic Minority Oversampling TEchnique*

Kombinacija prevzorčenja in podvzorčenja

- Parameter *pre*: stopnja prevzorčenja v odstotkih
- Parameter *pod*: stopnja podvzorčenja v odstotkih
- Število najbližjih sosedov k, $k > pre/100$

Sintetični primeri

- Namesto ponavljanja primerov iz manjšinskega razreda
- Tvorimo nove sintetične primere

SMOTE: tvorjenje sintetičnih primerov

Stopnja prevzorčenja *pre*

- Za vsak primer iz manjšinskega razreda
- Ustvarimo $\lfloor pre/100 \rfloor$ sintetičnih primerov

Za vsak primer e iz manjšinskega razreda $e = (\boldsymbol{x}, y) \in S$

1. Iz množice k najbližjih sosedov primera e iz manjšinskega razreda, naključno izberemo $\lfloor pre/100 \rfloor$ primerov
2. Za vsakega izbranega soseda $e_n = (\boldsymbol{x_n}, y) \in S$ ustvarimo nov primer $e_s = (\boldsymbol{x_s}, y)$, kjer je

$$\boldsymbol{x_s} = \boldsymbol{x} + g \cdot (\boldsymbol{x_n} - \boldsymbol{x})$$

in je g naključno število iz intervala $[0, 1]$.

SMOTE: podvzorčenje in nastavitve parametrov

Stopnja podvzorčenja *pod*

Iz primerov večinskega razreda ohranimo $\lfloor pod/100 \rfloor$ krat manj primerov, kot je primerov v manjšinskem razredu po opravljenem prevzorčenju.

Običajne nastavitve parametrov

- *pre*: 100%, 200%, 300%
- *pod*: 100%, 200%
- k: 5

SMOTER: regresija

Manjšinski in večinski razred

- Na osnovi vrednosti funkcije koristnosti ϕ
- Večinski razred za primere, kjer $\phi(y) > t$
- Manjšinski razred za ostale primere

Vrednost y_s za sintetične primere

$$y_s = y + g \cdot (y_n - y)$$

Algoritmi in implementacije

SMOTE (Chawla in ost. 2002)

Implementacija v R: paket `smotefamily`

SMOTER (Torgo in ost. 2013)

Implementacija v R: paket `ubl`