PRIMENJENI ALGORITMI

Priprema za kolokvijum

Priprema podataka

- Analiza Data Frejma - **display(describe(df))** za ceo datafrejm ili **display(countmap(df[!, :NazivPromenljive]))** za distribuciju po kategorijama.
- Nedostajuci podaci - uklanjanju cele kolone sa **select!(df, Not(:NazivKolone))** ili izbacivanje redove iz frejma gde podaci nedostaju **dropmissing!(df, [:NazivKolone])**
- Zamena nedostajućih - prazna polja u kolonama sa kategoričkim podacima možemo zameniti vrednošću iz te kolone koja se najčešće pojavljuje uz pomoć funkcije **mode()**. Prazna numerička polja možemo popuniti sa srednjom vrednost iz cele kolone.
- Podaci koji štrče
- Korelacija i multikolinearnost

Linearna regresija - Korelacija

- Korelacija predstavlja zavisnost između dve promenljive. U slučaju linearne regresije, to su promenljive x i y
- Perasonov koeficjent korelacije predstavlja vrednost iz intervala [-1, 1] kojim se izražava zavisnost između promenljivih.
  - *Vrednost 1 govori nam da su vrednosti u savršenoj korelaciji, 0 da nisu u korelaciji, a –1 da su u savršenoj negativnoj korelaciji*
- Iako ne postoji precizno tumačenje vrednoszti ovog koeficijenta, može se uzeti:
  - *Za vrednosti veće od 0.9, postoji veoma jaka korelacija*
  - *Za vrednosti između 0.7 i 0.9, postoji jaka korelacija*
  - *Za vrednosti između 0.5 i 0.7, postoji umerena korelacija*
  - *Za vrednosti manje od 0.5, postoji slaba korelacija*
- U Juliji, u paketu Statistics postoji funkcija **cor()** za računanje koeficijenta korelacije

Linearna regresija - Ocena kvaliteta

- Koeficijent determinacije r^2
  - Za vrednosti veće od **0.9**, model je **jako dobar** za predvidjanje
  - Za vrednosti između **0.7 i 0.9**, model je **veoma dobar** za predvidjanje
  - Za vrednosti između **0.5 i 0.7**, model je **dobar** za predviđanje
  - Za vrednosti **manje od 0.5**, model **nije dobar** za predvidjanje
- Mere za grešku:
  - *Prosek absolutnih vrednosti greske:* $\boldsymbol{mean(abs(e_i))}$
  - *Prosečna relativna greška:* $\boldsymbol{mean(abs(e_i/y_i))}$
  - *Prosek kvadrata greške (MSE):* $\boldsymbol{mean(e_i^2)}$
  - *Koren proseka kvadrata greške* (RMSE) : $\boldsymbol{sqrt(mean(e_i^2))}$

Logistička regresija - Confusion matrix

- Neka smo izvršili predviđanje klasa za neki skup podataka. Tada definišemo:
  - *TP (True positives) – broj tačnih klasifikacija da je podatak iz klase 1 (1->1)*
  - *TN (True negatives) – broj tačnih klasifikacija da podatak nije iz klase 1 (0->0)*
  - *FP (False positives) – broj netačnih klasifikacija da je podatak iz klase 1 (0->1)*
  - *FN (False negatives) – broj netačnih klasifikacija da podatak nije u kalsi 1 (1->0)*
  - *P (Positives, P=TP+FN) – broj elemenata klase 1*
  - *N (negatives, N=TN+FP) - broj elemenata koji nisu u klasi 1*

Actual class \ Predicted class	P	N
P	TP	FN
N	FP	TN

Logistička regresija - ocena kvaliteta klasifikacije

- Kvalitet klasifikacije ocenjujemo pomoću:
  - ***accuracy*** *(preciznost)* = ***(TP+TN)/(TP+TN+FP+FN)*** *= (TP+TN)/(P+N), preciznost (kvalitet) klasifikacije*
  - ***sensitivity*** *(osetljivost, True positive rates)* = ***TP/(TP+FN)*** *= TP/P , procenat tačno klasifikovanih podataka klase 1*
  - ***specificity*** *(specifičnost, True negative rates)* = ***TN/(TN+FP)*** *= TN/N, procenat tačno klasifikovanih podataka klase 0*
- Objektivnu meru klasifikatora predstavlja **AUC**, on predstavlja površinu ispod **ROC** krive i ova vrednost treba da bude što bliža vrednosti 1. ROC kriva predstavlja odnos između osetljivosti (TP/P) i greške (FP/N)

Zadatak 1: Predviđanje odobravanja kredita

- U datoteci **„krediti.csv“** nalaze se podaci o klijentima banke i da li im je odobren kredit. Koristeći tehnike pripreme podataka i logističke regresije istrenirati model koji predviđa da li će nekom klijentu biti odobren kredit.
  - Ukoliko u nekoj koloni nedostaje manje od 40% podataka, zameniti ih odgovarajućim tehnikama zamene.
  - Proceniti kvalitet modela.

Zadatak 2: Predviđanje težina ribe

- U datoteci **„ribe.csv“** nalaze se podaci o ribama koje su prodavane u jednom marketu, podatke koje imamo o ribama su vrsta, tezina, vertikalna duzina, dijagonalna duzina, duzina, visina, sirina i cena. Uz pomoc linearne regresije neophodno je predvideti tezinu ribe na osnovu svih podataka.
  - Ukoliko u nekoj koloni nedostaje manje od 40% podataka, zameniti ih odgovarajućim tehnikama zamene.
  - Ukoliko u nekoj koloni fali vise od 50% ukloniti tu kolonu
  - Prilikom testiranja koristiti podelu 75:25
  - Proceniti kvalitet modela.