Istraživanje podataka 1 - vežbe 7, 2020.

1 Dimenziona redukcija skupa podataka

Neki algoritmi istraživanja podataka ne rade dobro sa podacima velike dimenzionalnosti, tj. sa podacima sa velikim brojem atributa. Da bi se rešio ovaj problem, u okviru preprocesiranja podataka može da se koristi neki metod za dimenzionu redukciju.

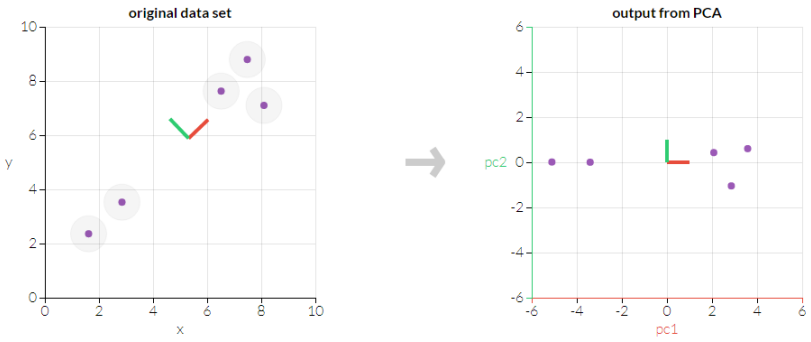
Metode za dimenzionu redukciju skupa podataka se primenjuju i pri vizuelizaciji podataka. Sa manjim brojem atributa je lakše i preglednije grafički prikazati skup.

1.1 Analiza glavnih komponenti - PCA (Principal Component Analysis)

Jedan od metoda za dimenzionu redukciju je analiza glavnih komponenti - PCA. Glavne osobine PCA su:

- Primenjuje se na numeričke podatke.
- Novi atributi, dobijeni primenom PCA, su nezavisni.
- Novi atributi su uređeni prema veličini varijanse skupa koju obuhvataju.
- **Prvi (novi) atribut** obuhvata **najveći** deo disperzije skupa.
- Svaki sledeći atribut obuhvata manji deo disperzije skupa koji nije pokriven prethodnim atributima.
- Transformacija skupa podataka se dobija korišćenjem matrice kovarijansi atributa skupa i njenih sopstvenih vrednosti.
- Svaki novi atribut je linearna kombinacija originalnog skupa. Težine linearne kombinacije $i.$ atributa su komponente $i.$ sopstvenog vektora.
- Varijansa $i.$ novog atributa je λ_i
- Novi atributi se zovu **glavne komponente**.
- Najčešće se je veliki deo disperzije skupa pokrivena malim brojem novih atributa u odnosu na ukupan broj atributa.

Na slici 1 je dat primer transformacije skupa primenom PCA.



Slika 1: Primena PCA Na grafiku sa originalnim skupom (levi grafik) crvenom bojom je označen pravac u kom se najviše prostiru podaci, na osnovu koga se pravi prva glavna komponenta (PC1). PC1 je označena na X osi na grafiku sa podacima nakon transformacije (desni grafik). Zelenom bojom je prikazan pravac koji obuhvata ostatak disperzije u skupu i na osnovu koga se pravi druga glavna komponenta (PC2). PC2 je prikazana na Y osi na grafiku sa podacima nakon transformacije (desni grafik).

Koraci pri upotrebi PCA:

- Standardizacija numeričkih atributa skupa podataka.
- Primena PCA na skup podataka. Ukupan broj novih atributa odgovara broju atributa originalnog skupa.
- Za svaki atribut (glavnu komponentu) se računa proporcija varijanse skupa koju taj atribut obuhvata.
- Uzima se prvih n glavnih komponenti (tj. atributa dobijenih nakon primene PCA) koji zajedno obuhvataju dovoljan deo varijanse za dalji rad (npr. 90%).
- Napomena: u daljem radu se ne koriste originalni atributi skupa, kao ni sve dobijene glavne komponente, već samo prvih n glavnih komponenti koje je korisnik izabrao.

1.2 PCA u biblioteci scikit-learn

Metod Analiza glavnih komponenti je implementiran klasom **sklearn.decomposition.PCA** čije osobine su:

- parametri
  - *n_components* - broj glavnih komponenti koje će biti sačuvane i korišćene u daljem radu
- atributi
  - *components_* - glavne komponente koje predstavljaju pravce maksimalne varijanse u podacima
  - *explained_variance_* -količina varijanse koju obuhvata svaka od glavnih komponenti
  - *explained_variance_ratio_* - procenat varijanse obuhvaćen svakom od glavnih komponenti.
    Napomena: ovaj atribut klase je bitan za izbor n glavnih komponenti za dalji rad. Bira se prvih n glavnih komponenti za koje važi da zajedno obuhvataju procenat varijanse koji je dovoljan za dalji rad.

- metode
  - *fit* - pravljenje modela za PCA na osnovu zadatog skupa
  - *fit_transform* - pravljenje modela i primena na zadatom skupu
  - *transform* - primena dimenzione redukcije na zadati skup
  - *inverse_transform* - transformacija u originalni skup

1.3 PCA u alatu IBM SPSS Modeler

U alatu IBM SPSS Modeler za analizu glavnih komponenti koristi se čvor **PCA/Factor**, koji je u paleti *Modeling*. PCA uzima u obzir samo numeričke atribute. Postavljanje parametara i čitanje rezultata:

- u odeljku *Model* za parametar *Extraction Method* potrebno je izabrati vrednost *Principal Components*
- u odeljku *Expert*, preko parametra *Maximum Number* postavlja se broj glavnih komponenti koje će biti izdvojene
- biranjem opcije *Run* pravi se model čiji je izlaz skup koji sadrži atribute iz originalnog skupa i atribute dobijene primenom PCA
- u okviru napravljenog modela u odeljku *Advanced* u tabeli *Total Variance Explained* u koloni *% of Variance* je prikazano za svaku glavnu komponentu koji procenat varijanse obuhvata. Kolona *Cumulative %* je korisna za odabir broja glavnih komponenti koje će se koristiti za dalji rad.

2 Prikaz grafika u programskom jeziku Python

Za grafičko predstavljanje podataka u programskom jeziku Python može da se koristi biblioteka *matplotlib*. U *matplotlib.pyplot* su implementirane funkcije korisne za crtanje. Neke od njih su:

- *figure* - za pravljenje nove slike (figure)
- *plot* - za crtanje tačaka pomoću zadatih markera i spajanje zadatih tačaka linijom. Neki argumenti:
  - *x* - vrednosti tačaka na x osi (prvi argument)
  - *y* - vrednosti tačaka na y osi (drugi argument)
  - *marker* - način označavanja tačaka (npr. x, *, o)
  - *color* - boja za crtanje
  - *label* - tekst koji se prikazuje u legendi
- *bar* - za crtanje stupčanog grafikona (bar chart). Neki argumenti:
  - *x* - vrednosti tačaka na x osi (prvi argument)
  - *height* -vrednosti tačaka na y osi (drugi argument)
  - *label* - tekst koji se prikazuje u legendi

- *scatter* - za pravljenje šeme sa raspršenim elementima. Neki argumenti:
  - x - vrednosti tačaka na x osi (prvi argument)
  - y -vrednosti tačaka na y osi (drugi argument)
  - *marker* - način označavanja tačaka (npr. x, *, o)
  - s - veličina tačaka
  - *color* - boja za crtanje
  - *label* - tekst koji se prikazuje u legendi
- *xlabel* - za postavljanje teksta na x osi
- *ylabel* - za postavljanje teksta na y osi
- *legend* - za prikaz legende na slici. Neki argumenti:
  - *loc* - lokacija za prikaz legende na figuri (vrednosti: best, upper right, upper left, lower right, right, center, ...)
- *show* - za prikaz slike

Primeri:

1. Primena PCA na skup iris
   Rešenje:
   - Python skript: **pca_iris.py**
   - IBM SPSS Modeler radni tok: **pca_iris.str**

   Zaključak: umesto 4 atributa, dovoljno je da koristimo 2 atributa, tj. prve dve glavne komponente nakon primene PCA.

2. Klasifikacija skupa podataka *car*

   Skup *car* sadrži podatke o automobilima. Atributi skupa su:
   - class - klasa automobila
   - cylinders - broj cilindara
   - displacement - zapremina motora
   - horsepower - konjska snaga
   - weight - težina
   - acceleration - ubrzanje

   Koristeći programski jezik Python i algoritam K najbližih suseda izvršiti klasifikaciju nad datim skupom. Atribut *class* je ciljni atribut. Primeniti unakrsnu validaciju za odabir optimalnih parametara.

   Za model napravljen sa optimalnim parametrima izdvojiti
   - preciznost na trening skupu;
   - preciznost na test skupu;

- matricu konfuzije za trening skup;
- matricu konfuzije za test skup;

Zatim primeniti PCA na skup i izvršiti klasifikaciju nad transformisanim skupom. Koji broj atributa ste izabrali nakon primene PCA i zašto? Uporediti model dobijen nad originalnim skupom sa modelom dobijenim nad podacima dobijenim nakon PCA transformacije.

Rešenje:

- Skup podataka: **car.csv**
- Python skript: **pca_car_class.py**
- Izlaz jednog pokretanja i odgovori na pitanja: **izlaz_i_komentari_primer2.txt**

3 Matrica cene u klasifikaciji

U praktičnoj primeni modela klasifikacije, pogrešna klasifikacija instanci jedne klase (ili jedne grupe klasa) može biti skuplja od druge. Npr. pogrešno klasifikovanje podnosioca zahteva za kredit sa visokim rizikom kao podnosioca zahteva sa niskim rizikom je za banku skuplje od klasifikovanja podnosioca zahteva sa niskim rizikom kao podnosioca zahteva sa visokim rizikom.

Matrica cena u klasifikaciji omogućava da korisnik definiše značaj za različite greške predviđanja i te vrednosti se uzimaju u obzir pri pravljenju modela. Matrica cena izgleda kao matrica konfuzije i prikazuje cenu za svaku moguću kombinaciju stvarne i dodeljene klase. Podrazumevano su sve cene pogrešne klasifikacije postavljene na 1, a cene dobre klasifikacije na 0. Podrazumevana matrica cena za 4 klase je data u tabeli 1.

Tabela 1: Podrazumevane vrednosti u matrici cena za 4 klase

		Dodeljena klasa			
		C_1	C_2	C_3	C_4
Stvarna klasa	C_1	0	1	1	1
	C_2	1	0	1	1
	C_3	1	1	0	1
	C_4	1	1	1	0

Cena se menja samo za pogrešna predviđanja, a za dobra predviđanja uvek ostaje 0. Npr. ako je za korisnika važnije da model dobro predviđa instance klase C_2 nego instance ostalih klasa, matrica cena bi mogla da bude kao u tabeli 2. Za jednu stvarnu klasu, cena pogrešnih predviđanja ne mora biti ista za sve dodeljene klase, npr. za instance klase C_2 cena dodele klasi C_3 može biti veća od cene dodele klasi C_4, kao u tabeli 3.
Matrica cena može da se zada pri korišćenju čvora C5.0 u alatu IBM SPSS Modeler, kao i nekih drugih čvorova za drveta odlučivanja.

Primena matrice cena u najvećem broju slučajeva dovodi do

- povećanja preciznosti za stvarne klase za koje se povećana cena (u primeru u tabeli 3 za klasu C_2) i smanjenja preciznosti za klase sa manjim cenama greške

Tabela 2: Primer matrice cena za 4 klase

		Dodeljena klasa			
		C_1	C_2	C_3	C_4
Stvarna klasa	C_1	0	1	1	1
	C_2	2	0	2	2
	C_3	1	1	0	1
	C_4	1	1	1	0

Tabela 3: Primer matrice cena za 4 klase

		Dodeljena klasa			
		C_1	C_2	C_3	C_4
Stvarna klasa	C_1	0	1	1	1
	C_2	2	0	4	3
	C_3	1	1	0	1
	C_4	1	1	1	0

- smanjenja ukupne preciznosti modela

Cene se biraju tako da se postigne odgovarajući kompromis između preciznosti modela i dobrog ponašanja za klasu/klase od interesa.

Primer:

3. U alatu IBM SPSS Modeler primeniti klasifikaciju nad skupom *bank.csv* korišćenjem C5.0. Ciljni atribut je oročena štednja.

   - Koji atributi su korišćeni pri pravljenju modela?
   - Komentarisati dobijen model.

   Rešenje:

   - Skup podataka: **bank.csv**
   - IBM SPSS Modeler radni tok: **bank_class.str**
   - Komentari:**primer3_komentari.pdf**