Istraživanje podataka

Vežbe 7

4. Mai 2018

Outline

Outline

Transformacija atributa

Numerički atributi

Sklariranje vrednosti u opseg $[0, 1]$ formulom

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Kategorički atributi

Za svaku kategoriju u kategoričkom atributu se pravi binarni atribut. U binarnom atributu za svaku kategoriju se instanciama te kategorije dodeljuje vrednost $\sqrt{\frac{1}{2}}$, a ostalim instancama 0.

Outline

Algoritam

Pronalazak klastera u algoritmu K-sredina je itrativni proces računanja centroida za svaki klaster i dodeljivanja instance klasteru.

Algoritam

Koraci

1. Računanje inicijalnih centroida za k klastera.
2. Svaku instancu dodeliti najbližem klasteru korišćenjem euklidskog rastojanja.
3. Za svaki klaster ažurirati centroid na osnovu dodeljenih instanci tom klasteru.
4. Ponavljati korake 2 i 3 dok se ne ispuni jedan od uslova:
   - Nijedan centroid se nije promenio u odnosu na prethodnu iteraciju.
   - Izvršen je maksimalan broj iteracija.

Inicijalni centroidi

Primena maxmin algoritma

1. Prva instanca u skupu se postavlja za centroid prvog klastera.
2. Za svaku instancu izračunati rastojanje do definisanih centroida klastera.
3. Pronaći najudaljeniju instancu od definisanih centroida i nju dodati kao novi centroid.
4. Ponavljati korake 2 i 3 dok se ne definiše k inicijalnih centroida.

Ažuriranje centroida klastera

Za svaki klaster C_j se centroid ažurira na kraju svake iteracije po formuli:

$$c_{qj} = \frac{\sum_{i=1}^{n_j} x_{qi}(j)}{n_j}$$

gde je

- n_j broj instanci u klasteru C_j
- $x_{qi}(j)$ je vrednost q. transformisanog atributa instance i koja je dodeljena klasteru C_j

Parametri u SPSS modeleru

- Broj klastera
- Maksimalan broj iteracija
- Tolerancija greške
  Ukoliko je za svaki klaster C_j u iteraciji i euklidsko rastojanje centroida u iteraciji i i centroida u iteraciji $i-1$ manje od zadate vrednosti tolerancije greške, vraća se dobijeni model.
- Vrednost sa kojom se kodira T u transformisanim binarnim atributima za kategoričke atribute.

Outline

Algoritam

- Model Kohonenovog algoritma je posebna vrsta modela neuronske mreže za učenje bez nadzora.
- Vrši prostorno organizovano klasterovanje, odnosno mapiranje osobina u 2D prostor u kome se slične instance grupišu.
- Kohonenova mreža modela ima dva sloja: ulazni i izlazni koji su u potpunosti povezani (svaki ulazni čvor sa svakim izlaznim) i svaka veza ima određenu težinu. Ulazni čvorovi odgovaraju atributima, a izlazni klasterima. Svaki klaster ima pridruženi centar (vektor težina između ulaznih čvorova i izlaznog čvora).

Algoritam

Pri obradi instanci skupa centri klastera se obrađuju slično kao kod algoritma K-means, stim što:

- klasteri su organiyovani prostorno u 2D mreži;
- pri dodavanju instance klasteru ažurira se njegov centar, ali i centri klastera u **susedstvu**.

Algoritam

Koraci

1. Mreža se inicijalizuje sa malim nasumice izabranim težinama.
2. Instance iz skupa se obrađuju nasumice izabranim redosledom.
3. Za svaku instancu se određuje najbliži klaster računanjem eyklidskog rastojanja između instance i centra klastera. Izlazni čvor koji odgovara najbližem klasteru se proglašava za **pobednika**.
   Centar pobednika se ažurira tako da bude bliže instanci koja mu se dodeljuje.
   Ako je veličina susedstva veća od 0, ažuriraju se i centri svakog susednog klastera tako da i njihovi centri budu bliže instanci koja se obrađuje.

Algoritam

Koraci

4. Na kraju svake iteracije ažurira se parametar učenja η.
5. Postupak se ponavlja dok se ne zadovolji neki uslov za zaustavljanje.

Algoritam

Model se trenira u 2 faze i za njih važi:

- u I fazi se zadaje veći broj suseda i veće η;
- u II fazi se zadaje mali broj suseda i malo η za fino podešavanje suseda.

Susedstvo

Za određvanje susednih klastera određenom klasteru koristi se Čebišljevo rastojanje:

$$d_c(x, y) = max_i|x_i - y_i|$$

gde je

- x_i lokacija čvora x na osi i na izlaznoj mreži ;
- y_i lokacija čvora y na osi i na izlaznoj mreži ;

Izlazni čvor o_j je sused izlaznom čvoru o_i ako je $d_c(o_i, o_j) < n$, gde je n veličina susedstva.

Ažuriranje težina

Pri obradi jedne instance, težine pobednika i njegovih suseda se ažuriraju dodavanjem dela razlike između instance i trenutnih težina. Intenzitet promene zavisi od parametra za učenje η. Promena težina se računa po formuli:

$$\Delta W = \eta * (W - I)$$

gde je

- W težina izlaznog čvora koji se ažurira;
- I instanca koja se obrađuje ;

Smanjivanje parametra za učenje η

Na kraju svake iteracije parametar η se smanjuje. Korisnik bira da li se smanjuje linearno ili eksponencijalno.

Parametri u SPSS modeleru

- Nastavak treniranja nad postojećim modelom
- Prikaz grafa tokom formiranja modela
- Maksimalno vreme zadato u minutima za treniranje mreže
- Seme za slučajno generisane brojeve
- Maksimalna širina i dužina mreže;
- Za svaku fazu
  - Veličina susedstva
  - Inicijalno η
  - Maksimalan broj iteracija