19. Grupiranje

Strojno učenje 1, UNIZG FER, ak. god. 2023./2024.

Jan Šnajder, vježbe, v2.2

1 Zadatci za učenje

1. [***Svrha: Razumjeti rad algoritma K-sredina u smislu minimizicije kriterija pogreške. Razumjeti kako rad algoritma ovisi o broju grupa K i odabiru početnih središta.***]

   Algoritam K-sredina minimizira kriterij pogreške $J(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K|\mathcal{D})$. Vrijednost tog kriterija ovisi o broju grupa K, koji je unaprijed postavljen, te o položajima središta, koja se mijenjaju kroz iteracije.

   (a) Nacrtajte skicu vrijednosti kriterija pogreške J kao funkcije broja grupa K. Koja je minimalna vrijednost funkcije J i zašto?

   (b) Izaberite na skici iz zadatka (a) tri vrijednosti za K i skicirajte na jednom grafikonu vrijednost kriterija pogreške J kao funkcije broja iteracija (tri krivulje).

   (c) Izaberite na skici iz zadatka (a) jednu vrijednost za K. Skicirajte na jednom grafikonu vrijednosti kriterija pogreške J kao funkcije broja iteracija, ali ovaj put uzevši u obzir stohastičnost uslijed slučajnog odabira početnih središta (nacrtajte nekoliko mogućih krivulja na istom grafikonu). Koje od tih krivulja su izglednije za algoritam K-means++?

2. [***Svrha: Isprobati rad algoritma K-sredina i K-medoida na konkretnom primjeru. Shvatiti da je složenost ovog drugog puno nepovoljnija.***] Raspolažemo skupom neoznačenih primjera:

   $$\mathcal{D} = \big\{a = (5,2), b = (7,1), c = (1,4), d = (6,2), e = (2,8), f = (3,6), g = (0,4)\big\}.$$

   (a) Izvedite jedan korak algoritma K-sredina uz $K = 3$. Za početna središta odaberite $\boldsymbol{\mu}_1 = b$, $\boldsymbol{\mu}_2 = c$ i $\boldsymbol{\mu}_3 = e$.

   (b) Izvedite jedan korak algoritma K-medoida uz $K = 3$. Za početna središta odaberite primjere b, c i e.

   (c) Usporedite računalnu složenost algoritma K-sredina i K-medoida.

   (d) Što su prednosti, a što nedostatci algoritma K-medoida?

3. [***Svrha: Isprobati izračun Randovog indeksa na konkretnom primjeru. Razumjeti primjenjivost Randovog indeksa.***] Nedostatak svih algoritama grupiranja koje smo razmotrili jest što se broj grupa K mora zadati unaprijed. Osim u rijetkim slučajevima kada nam je taj broj unaprijed poznat, to predstavlja problem.

   (a) Kada su primjeri ili podskup primjera označeni, kvaliteta grupiranja (uključivo i broj grupa K) može se procijeniti Randovim indeksom. Randov indeks zapravo izračunava točnost s kojom ćemo par jednako označenih primjera smjestiti u istu grupu, odnosno par različito označenih primjera u različite grupe. Izračunajte Randov indeks za sljedeću particiju označenih primjera (podskupovi su grupe dobivene grupiranje, a brojke su oznake klasa primjera):

   $$\big\{\{0,0,1,2\}, \{1,1\}, \{2,2,2,1,0\}\big\}.$$

   (b) Skicirajte vrijednost Randovog indeksa kao funkcije broja grupa K.

   (c) Randov indeks možemo koristiti samo ako su podatci označeni ili je podskup podataka označen. Međutim, čini se da to onda ujedno podrazumijeva da je unaprijed poznat broj grupa K. Imamo li koristi od Randovog indeksa čak i onda kada unaprijed znamo broj grupa? Možemo li ikako upotrijebiti Randov indeks, a da nam broj grupa nije unaprijed poznat?

2 Zadatci s ispita

1. (N) Raspolažemo sljedećim neoznačenim skupom primjera:

$$\mathcal{D} = \{(\mathbf{x}^{(i)})\}_i = \{(1,1), (1,2), (2,2), (2,3), (3,3)\}$$

Primjere grupiramo algoritmom K-sredina sa $K = 2$ grupe. Za početna središta odabrali smo primjere $\mathbf{x}^{(2)} = (1,2)$ i $\mathbf{x}^{(5)} = (3,3)$. Provedite prvu iteraciju algoritma K-sredina. **Koliko iznosi vrijednost kriterijske funkcije J nakon ažuriranja centroida?**

A 2.962 B 1.833 C 1.667 D 2.414

2. (P) Skup neoznačenih primjera u dvodimenzijskome ulaznom prostoru neka je sljedeći:

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{5} = \{(0,0), (0,4), (2,0), (2,4), (4,2)\}$$

Primjere grupiramo algoritmom K-sredina sa $K = 3$ grupe. Za početna središta grupa odaberemo nasumično primjere iz $\mathcal{D}$, pri čemu, naravno, pazimo da odaberemo različita središta. Ishod grupiranja i konačan iznos kriterijske funkcije J ovisit će o odabiru početnih središta. Neka je J^* vrijednost kriterijske funkcije u točki globalnog minimuma, dakle vrijednost koja odgovara najboljem grupiranju. Neka je J^+ vrijednost kriterijske funkcije u točki lokalnog minimuma, i to onoj točki lokalnog minimuma s najvećom vrijednošću funkcije J. **Koliko iznosi razlika $J^+ - J^*$?**

A 4 B 6 C 8 D 12

3. (P) Skup neoznačenih primjera u dvodimenzijskome ulaznom prostoru neka je sljedeći:

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{4} = \{(0,0), (0,4), (2,4), (4,2)\}$$

Primjere grupiramo algoritmom K-sredina u $K = 2$ grupe. Grupiranje možemo shvatiti kao pretraživanje prostora stanja, gdje svako pojedino stanje odgovara jednom pridjeljivanju primjera grupama. Pritom dva stanja smatramo identičnima ako su grupiranja identična, neovisno o identitetu grupa (npr., grupiranje kod kojega prva grupa sadrži samo primjer $\mathbf{x}^{(1)}$ je identično kao i grupiranje kod kojega drugo grupa sadrži samo primjer $\mathbf{x}^{(1)}$). Neka je A_1 algoritam K-sredina s potpuno slučajno inicijaliziranim središtima, a A_2 algoritam K-sredina gdje su središta inicijalizirana algoritmom K-means++. Neka je $S(A_1)$ skup stanja koje pretražuje algoritam A_1, a $S(A_2)$ skup stanja koje pretražuje algoritam A_2. Izračunajte veličine ovih skupova, uzevši u obzir mogućnost da pojedina grupa bude prazna te da dođe do izjednačenja udaljenosti primjera do centroida, što se razrješava slučajnim mehanizmom. **Koliko algoritam A_2 pretražuje manje stanja od algoritma A_1, tj. koliko iznosi $|S(A_1)| - |S(A_2)|$?**

A 4 B 6 C 12 D 10

4. (N) Algoritmom K-medoida (PAM) grupiramo $N = 5$ primjera. Za grupiranje koristimo mjeru različitosti, koja je za naših pet primjera definirana sljedećom matricom (matrica je simetrična, pa je donji trokut izostavljen):

$$\begin{array}{c} \\ \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \mathbf{x}^{(3)} \\ \mathbf{x}^{(4)} \\ \mathbf{x}^{(5)} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \mathbf{x}^{(3)} & \mathbf{x}^{(4)} & \mathbf{x}^{(5)} \end{array} \\ \begin{pmatrix} 0 & 0.2 & 0.9 & 0.7 & 0.5 \\ & 0 & 0.9 & 0.1 & 0.6 \\ & & 0 & 0.7 & 0.3 \\ & & & 0 & 0.8 \\ & & & & 0 \end{pmatrix} \end{array}$$

Grupiramo u $K = 2$ grupe, s primjerima $\mathbf{x}^{(1)}$ i $\mathbf{x}^{(5)}$ kao početnim medoidima. Provedite prvu iteraciju algoritma K-medoida (PAM). **Koje medoide dobivamo nakon prve iteracije?**

A $\mathbf{x}^{(1)}$ i $\mathbf{x}^{(3)}$ B $\mathbf{x}^{(3)}$ i $\mathbf{x}^{(4)}$ C $\mathbf{x}^{(1)}$ i $\mathbf{x}^{(2)}$ D $\mathbf{x}^{(2)}$ i $\mathbf{x}^{(5)}$

5. (N) Particijskim algoritmom grupiranja grupiramo $N = 1000$ primjera. Na temelju znanja o problemu zaključili smo da bi primjeri trebali formirati $K = 3$ grupe, pa smo s tim brojem grupa proveli grupiranje. Kako bismo evaluirali točnost grupiranja, slučajnim odabirom smo iz skupa

primjera uzorkovali 10 primjera, ručno smo označili primjere iz tog uzorka, i zatim na tom uzorku računamo Randov indeks. Označavanje smo proveli tako da smo svakom primjeru iz uzorka dodijelili oznaku točne grupe. Oznake grupe dobivene algoritmom grupiranja y_{pred} i oznake točnih grupa y_{true} za svih deset primjera u uzorku su sljedeće:

i	1	2	3	4	5	6	7	8	9	10
$y_{pred}^{(i)}$	0	1	2	2	1	0	0	2	1	2
$y_{true}^{(i)}$	1	1	0	2	0	0	1	1	1	2

Koliko iznosi Randov indeks grupiranja izračunat na ovom uzorku?

[A] 0.27 [B] 0.56 [C] 0.64 [D] 0.70

6. (N) Želimo grupirati $N = 1000$ primjera, ali nemamo nikakvih saznanja o optimalnom broju grupa. Kako bismo odredili optimalan broj grupa, odlučili smo označiti uzorak primjera i na tom uzorku izračunati Randov indeks $\mathrm{RI}(K)$ za grupiranja dobivena s različitim brojem grupa K. Naposlijetku ćemo onda kao optimalan broj grupa odabrati onaj K koji maksimizira Randov indeks, $K^* = \mathrm{argmax}_K \mathrm{RI}(K)$. Budući da ne znamo koji je točan broj grupa, umjesto označavanja pojedinačnih primjera označavamo parove primjera. U tu svrhu smo iz skupa primjera uzorkovali 16 različitih primjera, uparili ih u 8 različitih parova primjera, te smo za svaki par primjera ručno označili trebaju li dotični primjeri pripadati istoj grupi ili ne. Rezultat označavanja je takav da tri para primjera trebaju pripadati istoj grupi (indeksi parova 1–3), a pet različitim grupama (indeksi parova 4–8). Nakon toga proveli smo grupiranje za $K \in \{3, 4, 5\}$ grupa. Za uzorak označenih primjera dobili smo ovakve grupe:

$$
\begin{aligned}
K = 3 &: \{1,1,2,4,8\}\ \{2,3,7\}\ \{4,5,3,5,6,6,7,8\}\\
K = 4 &: \{1,1,2\}\ \{4,8,4\}\ \{2,3,7,5,7\}\ \{3,5,6,6,8\}\\
K = 5 &: \{1,1\}\ \{3,4,8\}\ \{2,2,4\}\ \{7,5,7,3,5,6\}\ \{6,8\}
\end{aligned}
$$

Brojke označavaju indeks para primjera. Na primjer, u grupiranju sa $K = 3$ grupe par primjera s indeksom 1 našao se u istoj grupi, a par primjera s indeksom 2 u različitim grupama. Izračunajte Randov indeks $\mathrm{RI}(K)$ te optimalan broj grupa K^* prema Randovom indeksu, za $K \in \{3, 4, 5\}$. **Koliko iznosi Randov indeks za optimalan broj grupa,** $\mathrm{RI}(K^*)$**?**

[A] 0.375 [B] 0.625 [C] 0.750 [D] 0.875

7. (N) Algoritmom K-sredina grupiramo $N = 1000$ primjera. U tom skupu nalazi se i uzorak od 11 primjera označenih oznakama $\mathcal{Y} = \{1, 2, 3, 4\}$. Međutim, nismo sigurni hoće li grupiranje u četiri grupe doista dati optimalne rezultate, pa isprobavamo grupiranje sa $K = 3$ i $K = 4$ grupe. Rezultati su sljedeći:

$$
\begin{aligned}
K = 3 &: \ \big\{\{1,2,4,4\},\{2,3,3\},\{1,1,3,4\}\big\}\\
K = 4 &: \ \big\{\{1,2,2,4,4\},\{3,3\},\{1,1\},\{3,4\}\big\}
\end{aligned}
$$

gdje podskupovi odgovaraju grupama, a brojke oznakama primjera. Izračunajte Randov indeks za oba ova grupiranja. **Koliko je Randov indeks za** $K = 4$ **veći od Randovog indeksa za** $K = 3$**?**

[A] 0.0164 [B] 0.0945 [C] 0.0727 [D] 0.0682