Skupina 22: *k*-means algoritem

Opis problema in načrt dela

Projekt v povezavi s predmetom Operacijske raziskave

Avtorja:
Matevž Raspet, Eva Šraj

Ljubljana, november 2018

1 OPIS PROBLEMA in NAČRT DELA

Pri projektu iz predmeta Operacijskih raziskav bova preučevala delovanje k-means algoritma.

K-means algoritem je metoda vektorske kvantizacije, namenjena za analizo grupiranja podatkov, predvsem pri rudarjenju s podatki. Ta algoritem naredi particijo n meritev v k različnih grup, v katerih vsaka meritev pripada grupi z najbližjo povprečno vrednostjo. Rezultat particije je razdelitev prostora v Voronojeve celice.
Bolj formalno: naj bodo $(x_1, x_2, \ldots, x_n)$ dane meritve, kjer je vsaka meritev x_i d-dimenzionalen realni vektor. K-means algoritem napravi particijo n meritev v $k(\leq n)$ množic $S = \{S_1, S_2, \ldots, S_k\}$, na podlagi najmanjše vsote kvadratov razdalj med meritvami posamezne množice in njihovim povprečjem μ_i.

$$\arg\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = \arg\min_S \sum_{i=1}^{k} |S_i| \cdot Var(S_i),$$

S pomočjo dveh metod bova zgenerirala in predstavila naključne podatke v prostorih $\mathbb{R}$, $\mathbb{R}^2$ in $\mathbb{R}^3$. Ti dve metodi sta: **metoda Manhattan** in **metoda najmanjših kvadratov** (krajše MNK). Uporabila ju bova na praktičnih življenjskih primerih, kot npr. *problem najbolj obiskane trgovine*.

Nato bova k-means algoritem uporabila na množicah, z 10% najbolj odstopajočimi meritami ter primerjala rezultata dobljena z Manhattan metodo in MNK.

Na koncu bova za množice, ki so generirane v prostoru $\mathbb{R}^2$, poiskala povezavo med k-means algoritmom in VORONOJEVIM DIAGRAMOM ter rešitev prikazala grafično.

1.1 NEZNANI POJMI

Kot sva omenila že zgoraj, bova pri projektu uporabila 2 metodi za računanje razdalj:

- Manhattan metodo in
- metodo najmanjših kvadratov.

Definicija: Naj bosta X in Y n-razsežna vektorja; $X = (x_1, x_2, \ldots, x_n)$ in $Y = (y_1, y_2, \ldots, y_n)$.
Razdalja po **metodi Manhattan** je definirana kot:

$$d(X, Y) = \sum_{i=1}^{n} |x_i - y_i|$$

Po **metodi najmanjših kvadratov** razdaljo definiramo kot:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Voronojev diagram
V matematiki Voronojev diagram prikazuje razdelitev ploskve na območja, osnovana na razdalji točk v posameznih podmnožicah območij.

Definicija: Naj bo X metrični prostor z razdaljo d, K indeksna množica in P_k ($k \in K$) nabor nepraznih podmnožic v (metričnem) prostoru X. Voronoijeva celica ali območje R_k povezano s P_k, je množica točk v X, katerih razdalja do P_k ni večja od razdalj do drugih P_j, kjer je j poljuben indeks različen od k ($j \neq k$).

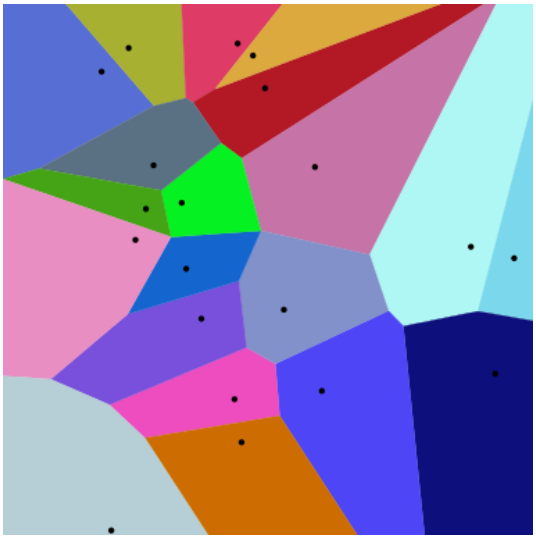
Z drugimi besedami: če je $d(x, A) = \inf\{d(x, a) | a \in A\}$ razdalja med X in A, je nato Voronojev diagram nabor celic $(R_k)_{k \in K}$, kjer je

$$R_k = \{x \in X | d(x, P_k) \leq d(x, P_j) zavsej \neq k\}$$

1.2 POTEK DELA S PRAKTIČNIM PRIMEROM

Kot enostaven primer bova preučevala *problem najbolj obiskane trgovine*.
Za množico bova vzela trgovine v nekem mestu. Predpostavila bova, da imajo v vseh trgovinah produkte z enako ceno in kvaliteto. Razumljivo je, da je kupčeva izbira odvisna le od njegove razdalje do trgovin (kupci bodo kupovali v njim najbližji trgovini). Midva pa želiva oceniti število kupcev v posamezni trgovini.

V tem primeru je Voronoijeva celica/območje R_k dane trgovine P_k približna ocena števila kupcev, ki bodo kupovali v tej trgovini P_k (trgovine so v našem modelu predstavljene kot točke v mestu).
Za večino mest bova razdaljo med točkami izmerila s pomočjo metode Manhattan in MNK. Pripadajoča grafa izgledata približno takole:



Slika 1: Voronojev diagram z MNK

Slika 2: Voronojev diagram z Manhattan metodo