KLASTERIZACIJA

Dat je set podataka genskih ekspresija pacijenata obolelih od raka koštane srži:

- *Yeoh* skup ima 248 uzorka i ukupan broj gena 2526 (dimenzionalnost), date su i labele koje ukazuju na 6 podtipova raka

Podaci su dobijeni sa čipom *Affymetrix* i imaju dosta neujednačene vrednosti ekspresija.

Prilikom klasterovanja uopšte ne koristite labele samo skup podataka. Da bismo olakšali proces možete da koristite informaciju o broju klasa prilikom klasterovanja na sledeći način: kod *k-means* algoritma za broj klastera postavićete broj klasa, a kod hijerarhijskog klasterovanja podesićete da se dendrogram preseca tako da rezultuje brojem klastera jednakom broju klasa (ako je takav presek moguć što ćete lako proveriti na osnovu dendrograma).

ZADATAK:

1. Nad datim skupom podataka uraditi normalizaciju i prikaz svake dobijene matrice podataka X kao slike (u *Matlabu*: imagesc(X); colormap(jet); colorbar):
    a. Skup koji sadrži originalne nenormalizovane podatke.
    b. Skup koji sadrži z-normalizovane podatke.
    c. Skup koji sadrži podatke normalizovane na opseg [0,1].

    Kao rezultat u ovom koraku imate 3 matrice podataka i 3 slike.

2. Nad svim napravljenim skupovima podataka (korak 1.) izvršiti *k-means* algoritam (*kmeans*) bar 15 puta uzimajući za inicijalne centroide slučajno odabrane uzorake (proverite parametre funkcije, može da se podesi). Neophodno je čuvati svih 15 dobijenih particija, najbolje u *mat* formatu (particija je data kao jedan vektor koji za svaki uzorak čuva oznaku labele klastera).
    a. Koristiti *sqeuclidean* rastojanje.
    b. Koristiti *cosine* rastojanje.

    Kao rezultat za svaku od 3 matrice podataka treba da dobijete po 15 particija za svako od rastojanja. Ukupno 45 particija. Maksimalan broj iteracija *k-means* algoritma je proizvoljan (Matlab će prijaviti grešku ako algoritam ne može da iskonvergira).

3. Nad svim napravljenim skupovima podataka (korak 1.) izvršiti algoritam hijerarhijske klasterizacije (*linkage i cluster*) koristeći obe mere rastojanja (*euclidean* i *cosine*) i iscrtati dendograme. Sačuvati dobijene particije.
    a. Koristiti *average linkage*.
    b. Koristiti *single linkage*

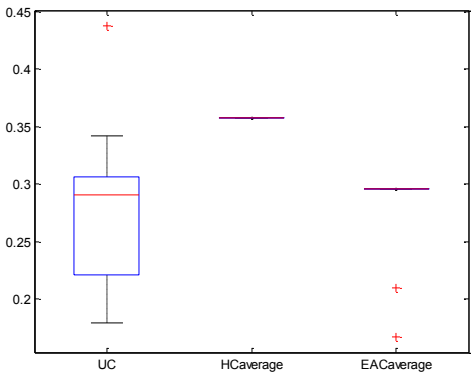
    Kao rezultat za svaku od 3 matrice podataka dobićete po 4 particije (kombinovanjem dve mere rastojanja (*euclidean* i *cosine*) i dve mere sličnosti klastera (*average* i *single*)). Ukupno 12 particija. Možete da isprobate i funkciju *clustergram* za vizuelni prikaz hijerarhijskog klasterovanja.

4. Za svaku dobijenu particiju izračunati interne mere validacije (*evalclusters*): *CalinskiHarabasz, DaviesBouldin i Silhouette* (prosečnu širinu). Dati i sliku *silhouette* koeficijenata.

a. *CalinskiHarabasz*: koristi pretpostavku da dobra klasterizacija ima karakteristiku male sume kvadrata rastojanja u okviru klastera, odnosno velike između različitih klastera; veće vrednosti su bolje.
b. *DaviesBouldin:* koristi pretpostavku da klasteri treba da budu kompaktni i da su međusobno što bolje razdvojeni; manje vrednosti su bolje.
c. *Silhouette*: koristi pretpostavku da uzorci međusobno treba da budu bliskiji sa uzorcima iz svog kalstera u poređenju sa njihovom blizinom do uzoraka iz najbližeg susednog klastera. Računa se za svaki uzorak i kreće se u opsegu [-1,1]. Za celu particiju vrednost indeksa se dobija kao srednja vrednost indeksa za sve uzorke; veće vrednosti su bolje.

5. Za eksternu evaluaciju koristiti priloženi kod *rand_index* koji poredi originalne labele uzoraka sa labelama dobijenim klasterizacijom. Koristiti *Adjusted rand index* koji je korigovan rand index kako bi se izbegla slučajna podudaranja. Ovo je eksterna mera validacije što znači da prilikom njenog izračunavanja morate da prosledite i labele uzoraka. Ovaj indeks ukazuje na to koliko se dobro klasteri poklapaju sa stvarnim klasama (ne garantuje se da su date klase odgovarajuće, te je moguće dobiti i bolje mere sa nekom novom raspodelom po klasama).
   ari=rand_index(p1,p2,'adjusted') % p1=labele dobijene klasterovanjem, p2=stvarne labele %p1 i p2 moraju biti vektori istih dimenzija

U izveštaju komentarisati dobijene rezultate i na osnovu dobijenih internih i eksternih mera validacije uporediti *k-means* i hijerarhijsku klasterizaciju. Za svaki skup podataka (dobijen u koraku 1), svaki indeks validacije (3 interna i 1 eksterni) i svaku meru rastojanja ((*sq)euclidean*, *cosine*) nacrtati po jedan *box plot* jednog indeksa za sva tri korišćena algoritma: opseg vrednosti indeksa za 15 particija k-means, vrednost indeksa za hijerarhijsku klasterizaciju sa *average linkage* i vrednost indeksa za hijerarhijsku klasterizaciju za *single linkage* (kod hijerarhijskog klasterovanja imate samo jednu particiju - samo jedna crta na *box plotu* jer imamo jednu vrednost indeksa, a ne opseg vrednosti). Jedan sličan primer je dat na slici 1.



Slika 1. Jedan primer box plot slike. Na y osi je vrednost indeksa, na x osi tri algoritma, a u ***nazivu slike*** treba da stoji ***skup podataka*** – ***vrsta normalizacije*** – ***vrsta rastojanja*** – ***naziv indeksa***

Važne napomene:

Skup podataka ima jako veliku dimenzionalnost i može da se desi da neki koraci potraju. Ako se to desi, a vi hoćete samo da testirate ili debagujete program napravite matricu podataka sa malom dimenzijom, na primer ekspresije prvih deset gena. Kada utvrdite da sve radi kako očekujete, počnite da radite sa stvarnim podacima. Najlakše je da napravite program koji za unet set podataka (matrica podataka, vektor stvarnih labela i string koji ukazuje na tip normalizacije) radi sve što se zahteva u domaćem zadatku počev od tačke 2. Na primer:

Domaci3(X, p2, 'nenormalizovan')

% X- matrica podataka

% p2- vektor stvarnih labela

% s – string koji sadrži način normalizacije.

Ovaj kod treba da prikaže set podataka kao sliku, izvrši k-means i hijerarhijsku klasterizaciju, izračuna indekse i generiše slike. String koji je jedan od ulaznih argumenata koristite kao prefiks nazivima slika ili prilikom snimanja mat fajlova.

Za svaki problem ili nedoumicu možete nas kontaktirati mejlom.