Istraživanje podataka 1 - vežbe 12, 2021.

1 Zadaci

1. Dat je skup *dogs* koji ima atribute:

   - *breed* - rasa psa
   - *height* - visina psa
   - *weight* - težina psa

   Primeniti hijerarhijsko sakupljajuće klasterovanje na skup korišćenjem biblioteke *scikit-learn* programskog jezika Python. Pri klasterovanju koristiti atribute visina i težina psa.

   Izvršiti klasterovanje korišćenjem različitih veza pri određivanju bliskosti dva klastera i izdvojiti 3 klastera.

   Rezultat svakog izvršenog klasterovanja prikazati pomoću grafika sa šemom sa raspršenim elementima. Svakom klasteru u rezultatu dodeliti jedinstvenu boju i njegove instance označiti tom bojom. U naslovu grafika ispisati ime korišćene veze i vrednost silueta koeficijenta.

   Uporediti dobijene modele.

2. Dat je skup *dogs* koji ima atribute:

   - *breed* - rasa psa
   - *height* - visina psa
   - *weight* - težina psa

   Izvršiti hijerarhijsko sakupljajuće klasterovanje na osnovu visine i težine pasa korišćenjem euklidskog rastojanja i *average* veze pri određivanju bliskosti dva klastera. Za klasterovanje koristiti biblioteku *scipy* programskog jezika Python.

   Rezultat izvršenog hijerarhijskog klasterovanja prikazati pomoću dendograma.

   Ako je prag za spajanje klastera 0,3, koliko klastera se izdvaja?

   Rezultat klasterovanja sa pragom za spajanje 0,3 prikazati pomoću šeme sa raspršenim elementima. Kao naslov gafika sa šemom sa raspršenim elementima ispisati vrednost silueta koeficijenta.

3. Dat je skup *dogs* koji ima atribute:

   - *breed* - rasa psa
   - *height* - visina psa

- *weight* - težina psa

Primeniti klasterovanje algoritmom DBSCAN na osnovu visine i težine pasa korišćenjem programskog jezika Python. Izvršiti klasterovanja korišćenjem različitih vrednosti za rastojanje među susedima (0.1, 0.2, 0.25, 0.27, 0.28, 0.3), a vrednost parametra za broj suseda koje mora da ima instanca u jezgru postaviti na 2.

Rezultat svakog izvršenog klasterovanja prikazati pomoću grafika sa šemom sa raspršenim elementima. Svakom klasteru u rezultatu dodeliti jedinstvenu boju i njegove instance označiti tom bojom. U naslovu grafika ispisati vrednost za parametar *eps* i vrednost silueta koeficijenta.

Uporediti dobijene modele.

4. U programskom jeziku Python izvršiti klasterovanje nad skupom *unbalance.csv* primenom algoritama K-sredina, DBSCAN i hijerarhijskog sakupljajućeg klasterovanja za različit broj klastera. Skup *unbalance.csv* sadrži numeričke atribute X i Y. Rezultate klasterovanja prikazati pomoću šeme sa raspršenim elementima. Za svaki algoritam napraviti i grafik koji prikazuje silueta koeficijent za različit broj izdvojenih klastera.

5. U programskom jeziku Python izvršiti hijerarhijsko sakupljajuće klasterovanje nad skupom u datoteci *studenti.csv* za različit broj klastera u intervalu od 2 do 34 i primenom različitih veza za određivanje bliskosti dva klastera. Skup u datoteci *studenti.csv* sadrži podatke o studentima koji su diplomirali. Za svakog studenta su izdvojeni sledeći podaci: indeks, naziv smera koji je diplomirao, dužina studiranja u godinama, broj položenih ispita i prosečna ocena. Za svaku primenjenu vezu

   - rezultat kalsterovanja prikazati pomoću grafika sa silueta koeficijentom i brojem izdvojenih klastera
   - za klasterovanje sa najvećim senka koeficijentom izdvojiti deskriptivne statistike za svaki klaster da bi se uočilo šta ga karakteriše