DN2: Razvrščanje v skupine

1. Opis podatkov. Ime raziskave oziroma podatkov. Leto in kraj/država, če je to za podatke smiselno. Enota analize. Velikost vzorca. Uporabljene spremenljivke (ime v datoteki in vsebinski pomen). Ostale posebnosti, za katere menite, da so vredne omembe (obravnava manjkajočih vrednosti, rekodirane spremenljivke ...). Zapišite, katere spremenljivke boste uporabili za razvrščanje in ali jih boste za ta namen standardizirali.

2. Hierarhično razvrščanje. Primerjajte tri različne metode (pri vsaki zapišite ime metode in uporabljeno razdaljo). Uporabite Wardovo metodo s kvadrirano Evklidsko razdaljo in še dve drugi metodi. Izberite najbolj primerno razvrstitev (metoda in število skupin). Vašo izbiro utemeljite.

3. Razvrščanje K-means. Izberite/določite najbolj primerno število skupin. Pomagajte si z rezultati hierarhičnega razvrščanja in dodatno za metodo K-means narišite grafe: koleno, pseudo F, gap statistika. Vašo izbiro utemeljite. Primerjajte vrednost kriterijske funkcije pri različnem številu skupin za Wardovo metodo in K-means. Katera metoda je boljša? Pri izbranem številu skupin primerjajte obe razvrstitvi. Izpišite kontingenčno tabelo in izračunajte popravljen Randov indeks.

4. Razvrščanje na podlagi modelov. Ali je porazdelitev spremenljivk primerna za uporabo te metode? Ne glede na primernost uporabite metodo za razvrščanje na vaših podatkih. Izberite najbolj primeren model in število skupin. Izbiro utemeljite.

5. Predstavitev skupin. Izberite najboljšo razvrstitev po vašem mnenju in predstavite skupine.
   a) Izračunajte povprečja standardiziranih spremenljivk po skupinah. Rezultat grafično prikažite. Skupine opišite po najbolj izstopajočih lastnostih.
   b) Skupine analizirajte bolj podrobno po nekaterih izbranih spremenljivkah. Preverite, ali obstaja povezanost med skupino in temi spremenljivkami. Povezanost analizirajte na podlagi grafičnih prikazov in z ustreznimi izračuni. Interpretirajte moč in smer povezanosti. Preverite, ali je povezanost statistično značilna pri 5 % stopnji značilnosti. Pred analizo navedite, kateri test ste uporabili in zakaj ste izbrali ta test.

---

cluster: clusGap
mclust: Mclust, adjustedRandIndex