Razvrščanje v skupine

24. april 2018

1 Uvod

Cilj naloge je čim bolje razvrstiti podatke o genomih (61637 primerov opisanih s 25133 atributi) v skupine. Pravilna razvrstitev je bila strežniku znana in kvaliteto naših oddaj je ocenjeval s popravljenim indeksom po Randu (ARI).

2 Metoda

Najboljši sprotni rezultat (0.40113) sem dosegla z metodo k-voditeljev (razred `KMeansClustering`). Podatke sem prebrala z metodo `io.mmread` iz knjižnice `scipy`, ki je med drugim namenjena tudi branju datotek MTX. Obdržala sem 550 značilk, ki imajo najbolj različne vrednosti in podatke binarizirala. Za k pri metodi k-voditeljev sem izbrala $k = 43$.

3 Konsenzno razvrščanje

V razredu `ConsensusClustering` sem implementirala metodo konsenznega razvrščanja. Metoda ob inicializaciji sprejme tabelo števil gruč za algoritem K-voditeljev, število iteracij prevzorčenja i in število primerov, ki jih obdržimo pri prevzorčenju. Metoda za vsak podani k i-krat izvede prevzorčenje in razvrščanje v skupine z metodo K-voditeljev. Zaradi varčevanja s prostorom si povezovalnih (angl. *connectivity*) in indikatorskih matrik nisem shranjevala, vendar jih sproti seštevala, tako da sem na koncu konsenzno matriko izračunala v enem koraku (le deljenje). Sledil je izračun empirične kumulativne porazdelitve (CDF) in območja pod krivuljo (AUC) iz posamezne konsenzne matrike ter na koncu iskanje najboljšega k. Konsenzno matriko, ki je pripadala najboljšemu k, sem podala kot vhod algoritmu spektralnega gručenja (angl. *Spectral Clustering*) in tako dobila končno razvrstitev primerov v razrede.

Metodo sem v procesu izdelave (in tudi na koncu) testirala s podatkovno zbirko `iris`, ki ni tako obsežna kot podana v nalogi. Ravno tako je pravilen k za zbirko `iris` že znan in to je 3, ki sem ga dosegla tudi v svoji implementaciji.

Metodo konsenznega razvrščanja sem zagnala nad podatki s strežnika in dosegla rezultat 0.12926. Zaradi dolgega trajanja izvajanja algoritma sem se v nadaljevanju posvetila le izboljševanju metode k-voditeljev.

4 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelala sama.