

## Slop 6: Rojenje s K-Means

### Uvod

Rojenje (ang. clustering), včasih tudi gručenje, je metoda nenadzorovanega strojnega učenja, pri kateri želimo vzorce podatkov razdeliti v skupine oziroma roje glede na njihovo medsebojno podobnost. Za razliko od klasifikacije pri rojenju vnaprej ne poznamo razredov, v katere spadajo posamezni vzorci.

Eden izmed najpogostejše uporabljenih algoritmov za rojenje je K-means. Algoritem podatke razdeli v vnaprej določeno število  $k$  gruč. Vsak roj je predstavljen s svojim centroidom. Algoritem iterativno: določi začetne centroide,

vsak vzorec dodeli najbližjemu centroidu,

ponovno izračuna centroide na podlagi dodeljenih vzorcev,

postopek ponavlja, dokler se razporeditev vzorcev bistveno ne spreminja.

Podobnost vzorcev se običajno določa z evklidsko razdaljo. Zato lahko različne merske lestvice značilk močno vplivajo na rezultat in je pogosto potrebno podatke pred rojenjem normalizirati.

### Predstavitev problema

Podatkovni nabor Iris vsebuje meritve cvetov treh vrst perunika:

Iris setosa, Iris versicolor, Iris virginica.

Za vsak cvet so podane štiri numerične značilke:

sepal length, sepal width, petal length, petal width.

Pri prejšnji vaji smo poznali vrsto cveta in na podlagi učnih podatkov zgradili klasifikator Naive Bayes. Tokrat bomo predpostavili, da vrste cvetov ne poznamo. Z metodo K-means bomo poskusili samo na podlagi štirih meritev ugotoviti, ali lahko podatke razdelimo v naravne skupine.

Ker dejansko vrsto cvetov poznamo, bomo lahko na koncu rezultat rojenja primerjali tudi s pravo razvrstitvijo.

### Naloga

Z uporabo operatorja **Read Excel** naložite podatke Iris.

Oglejte si podatke in njihove značilke.

Z uporabo operatorja **Set Role** značilki species nastavite posebno vlogo label.

Algoritem K-means te značilke ne sme uporabljati pri določanju rojev.

Narišite 3d raztresni graf podatkov:

$x$  = petal length

$y$  = petal width

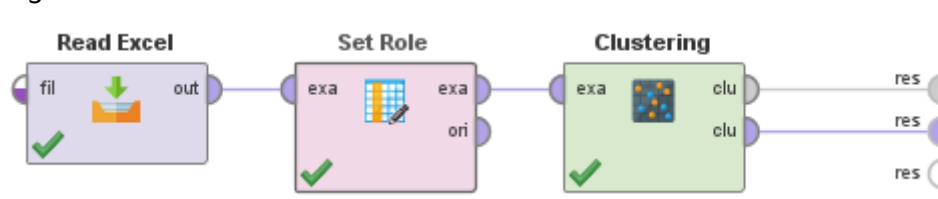
$z$  = sepal length

Color: species

Opazujte, ali lahko že iz grafa prepoznate naravne skupine podatkov.

Uporabite operator **K-Means** in nastavite:

$k = 3$



Oglejte si dobljeni model. Zapišite vrednosti centroidov posameznih gruč.

|           | sepal length | sepal width | petal length | petal width |
|-----------|--------------|-------------|--------------|-------------|
| cluster 0 |              |             |              |             |
| cluster 1 |              |             |              |             |
| cluster 2 |              |             |              |             |

Tabela: Centroidi rojev

Oglejte si podatke po izvedenem rojenju. Vsakemu vzorcu je algoritem pripisal enega izmed treh rojev.

Ponovno narišite raztresni graf:

x: petal length

y: petal width

z: sepal length

Color: cluster

Primerjajte ga z grafom, kjer je bila barva določena glede na species.

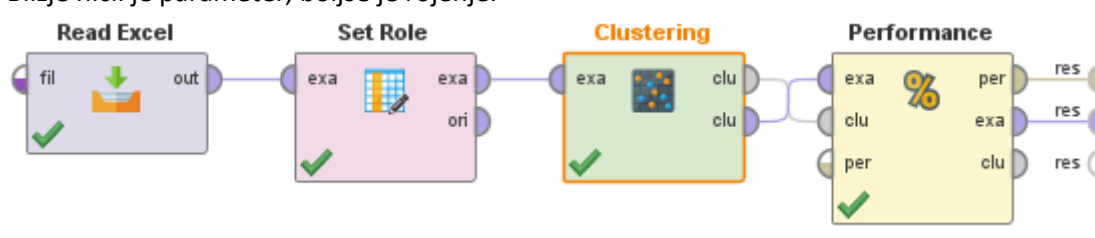
Odgovorite:

Ali trije roji, ki jih je našel K-means, ustrezajo trem dejanskim vrstam perunik?

Katero vrsto perunik algoritem najlažje loči od ostalih?

Pri katerih dveh vrstah prihaja do največjega prekrivanja?

V proces dodajte operator Cluster Distance Performance in za kriterij izberite Davies Bouldin (DBI). Bližje ničli je parameter, boljše je rojenje.



Spremenite število rojev in postopek ponovite za:

k=2, k=3, k=4, k=5

Opazujte rezultat rojenja in vrednost kriterija kakovosti rojenja. Če bi narisali vrednost DBI v odvisnosti od števila rojev bi opazili koleno – pri kolenu je izbira števila rojev najsmiselnejša.

## Vir

-Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining. Practical Machine Learning. Tools and Techniques. Third Edition

- <http://rapid-i.com/>

- <https://sites.google.com/site/dataminingforthemasses/>

Vaja pripravljena v okviru KCUI delavnice na UL-FE