Pēteris Račinskis
pr20015

Mājas darbs 3: Lēmumu koki, PCA

Uzdevums 1

a. Lēmumu koka parametri - koka izmēru samazināšana, nezaudējot precizitāti

Rezultāti, izmantojot noklusējuma parametrus:

```
Number of Leaves  :     34

Size of the tree :      67


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1436               95.7333 %
Incorrectly Classified Instances        64                4.2667 %
Kappa statistic                          0.9502
Mean absolute error                      0.0138
Root mean squared error                  0.1057
Relative absolute error                  5.6471 %
Root relative squared error             30.2115 %
Total Number of Instances             1500
```

Pirmais parametrs, ko mainot var samazināt koka izmērus, ir *minNumObj* - elementu skaits, pēc kura sasniegšanas koks netiek sīkāk dalīts un par klasi lapā tiek pieņemta biežākā (pēc noklusējuma = 2). Mainot tikai šo parametru, koka izmērus var samazināt aptuveni par 50%, saglabājot precizitāti virs 95% (gadījumā ar minimālo skaitu 8):

```
Number of Leaves  :     18

Size of the tree :      35


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1426               95.0667 %
Incorrectly Classified Instances        74                4.9333 %
Kappa statistic                          0.9424
Mean absolute error                      0.0199
Root mean squared error                  0.1107
Relative absolute error                  8.119  %
Root relative squared error             31.6474 %
Total Number of Instances             1500
```

Mainot citus parametrus, acīmredzamas sakarības grūti atrast. Dažos gadījumos *subtreeRaising* atslēgšana nedaudz pasliktina precizitāti. *confidenceFactor* palielināšana var palielināt koka izmērus, jo ierobežo *pruning*, taču precizitātes uzlabojumus, šķiet, nesniedz. Citiem parametriem robustas sakarības nav izdevies novērot. Izdrukāto rezultātu - precizitātes procentu - bieži vien var palielināt, palielinot krosvalidācijas atkārtojumu skaitu, taču šī atšķirība ir iluzora, jo atšķiras kļūdas novērtejuma metode, nevis rezultējošais modelis (kas vienmēr tiek ģenerēts no pilnās datu kopas, krosvalidācijā izmantotie modeļi tiek izmesti).

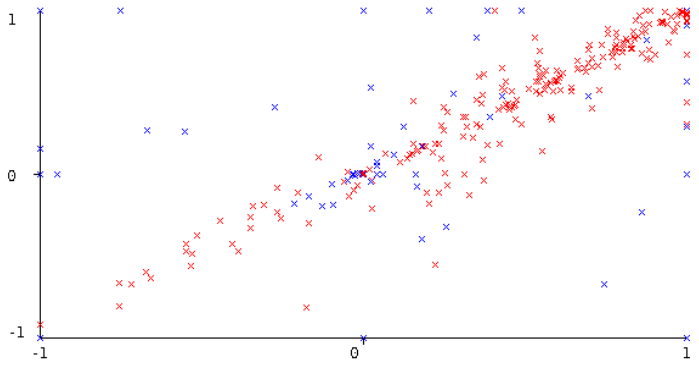
Pēteris Račinskis
pr20015

Uzdevums 2

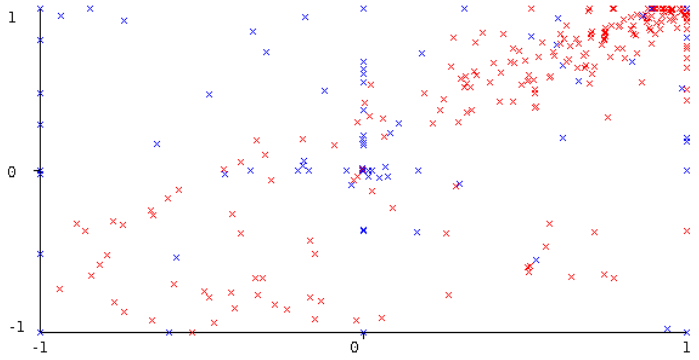
a. Projekciju salīdzinājums dotajās koordinātu asīs un PCA iegūtajās

Vienkārši projicējot datus plaknēs pa divām asīm var pamanīt sekojošās sakarības:

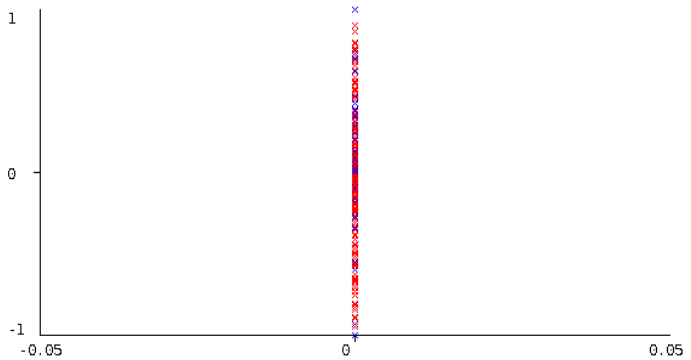
1. Daudzos gadījumos pastāv stipra korelācija starp argumentiem:



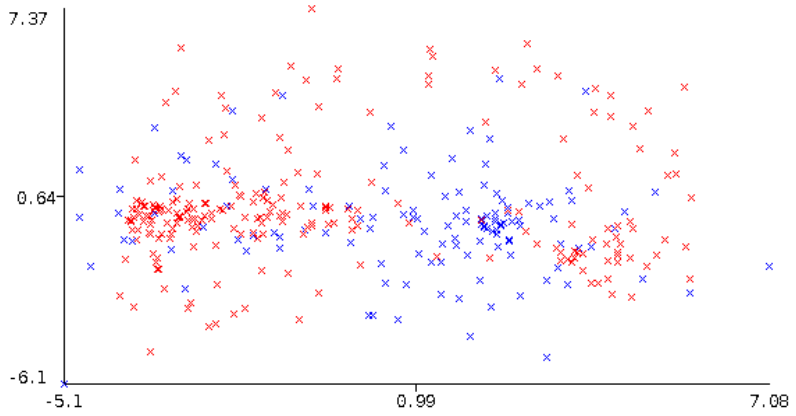
2. Izplatīti ir gadījumi, kad sagaidāmā vērtība vienai klase ir nošķirta no otras, taču lielu, pārklājošos dispersiju dēļ klasifikācija tāpat būtu sarežģīta (turklāt dati izskatās mākslīgi “nogriezti”):



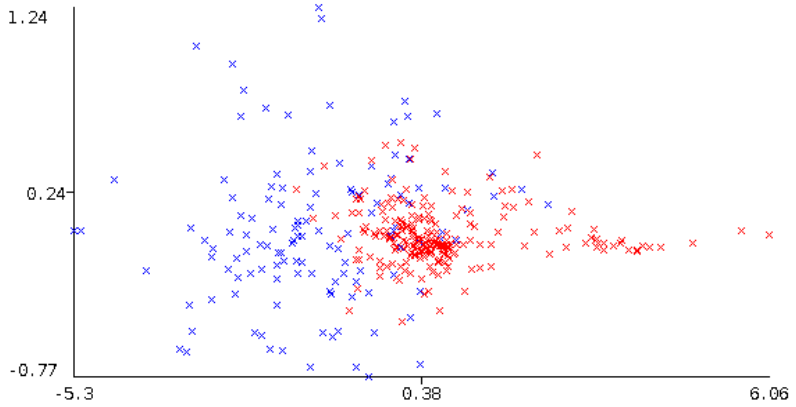
3. Pastāv asis, kurās atšķirīgu vērtību nav vispār:



Pēc PCA veikšanas iegūtajās projekcijās uzreiz redzams, ka ar trešo problēmu jāsaskarās nebūs. Ņemot divas labākās PCA asis, dati ir samērā vienmērīgi izkliedēti, bez izteiktām korelācijām, un klašu vidējās vērtības ir skaidri nošķirtas, taču dispersijas šajā plaknē tāpat ir lielas un klasifikācija, visticamāk, nebūtu īpaši precīza:



Pārskatot citas pieejamās kombinācijas, redzams, ka var atrast tādas kombinācijas, kurās, iespējams, klasifikācijas precizitāte būtu augstāka:



bet ļoti lielā daļā projekciju absolūti dominē klases “b” dispersija un par klašu vidējām vērtībām ir grūti izdarīt secinājumus: