Test from DataRobot

By Yiqian Jin
yj2342@columbia.edu

Apr.26

Topic: Breast Cancer Prediction

Data: 683 observations with 10 features:
"Cl.thickness" "Cell.size" "Cell.shape" "Marg.adhesion" "Epith.c.size"
"Bare.nuclei" "Bl.cromatin" "Normal.nucleoli" "Mitoses" "Class"

	row.names	Result	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1	1	benign	1000025	5	1	1	1	2	1	3	1	1
2	2	benign	1002945	5	4	4	5	7	10	3	2	1
3	3	benign	1015425	3	1	1	1	2	2	3	1	1
4	4	benign	1016277	6	8	8	1	3	4	3	7	1
5	5	benign	1017023	4	1	1	3	2	1	3	1	1
6	6	malignant	1017122	8	10	10	8	7	10	9	7	1
7	7	benign	1018099	1	1	1	1	2	10	3	1	1
8	8	benign	1018561	2	1	2	1	2	1	3	1	1
9	9	benign	1033078	2	1	1	1	2	1	1	1	5
10	10	benign	1033078	4	2	1	1	2	1	2	1	1

two classification results: {“benign”, “malignant”}

Process:
Randomly separate dataset into 90% training set(615 obs.) and 10% testing set(68 obs.). Build the model in API using training data and then predict and evaluate on testing data.

Result:

```
> table(pred_list, testset$Result)

pred_list   benign malignant
  benign        39         2
  malignant      4        23
```

The above is the table of prediction results(pred_list) and the true result(testset$Result).

The correct rate is 92.65%, showing that the prediction model performs quite well. Inspecting the data, the benign cells differ a lot from malignant cells in those 10 features. Then the model could have this good learning performance.
I also tried to use SVM classifier with the functionality of SVMLIB, associated with Cross-Validation, which also comes out a good prediction model.

Thanks!