Homework 4: Logistic Regression and Decision Trees

Theresa Wakonig
Data Mining I

November 23, 2020

1 Exercise 1

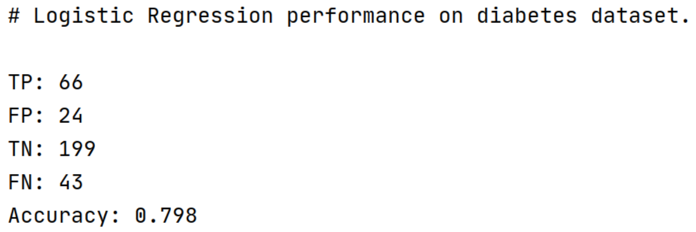
Logistic Regression on ’diabetes’ dataset.

1.a

See *ex1.py*.

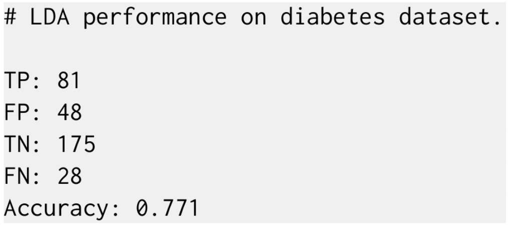
1.b

Output, performance metrics:



```
# Logistic Regression performance on diabetes dataset.

TP: 66
FP: 24
TN: 199
FN: 43
Accuracy: 0.798
```



```
# LDA performance on diabetes dataset.

TP: 81
FP: 48
TN: 175
FN: 28
Accuracy: 0.771
```

I believe the choice of classifier for the diabetes dataset also depends on what one wants to achieve with this classification process. The aim of reducing the FP and minimizing the FP rate in the predictions is best achieved by logistic regression. The TP rate can however be maximized with the LDA. I would perhaps choose to perform logistic regression because it yields higher accuracy and precision on the given data.

1.c

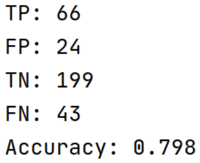
For any other dataset I would choose logisitc regression. Unlike for LDA, no prior assumptions on the data must hold when performing logistic regression. Achieving robust results with LDA would require close to normally distributed data as well as the same covariance among groups.

1.d

The two attributes which appear to contribute the most to the prediction are glu (plasma glucose concentration) and ped (diabetes pedigree function) with weights of about 0.97 and 0.53 respectively.

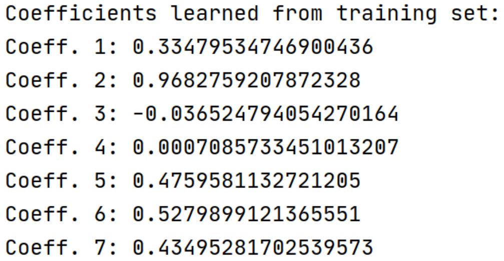
The coefficient for age is 0.43. Calculating the exponential function results in 1.54, which amounts to an increase in diabetes risk of 54% per additional year.

Performance on the reduced dataset:



```
TP: 66
FP: 24
TN: 199
FN: 43
Accuracy: 0.798
```

By comparing the performance and the coefficients obtained on the reduced dataset with the ones on the model including all the attributes, I observe that there is no difference in the performance metrics and hardly any change in the magnitude of the coefficients.
My explanation is that the coefficient of the attribute 'skin' is several orders of magnitude smaller than the other attributes' coefficients (see coeff. 4 in the list below; coeff. of unmodified dataset). The contribution of the 'skin' attribute to the overall prediction is therefore very small and negligible.



```
Coefficients learned from training set:
Coeff. 1: 0.33479534746900436
Coeff. 2: 0.9682759207872328
Coeff. 3: -0.036524794054270164
Coeff. 4: 0.0007085733451013207
Coeff. 5: 0.4759581132721205
Coeff. 6: 0.5279899121365551
Coeff. 7: 0.43495281702539573
```

2 Exercise 2

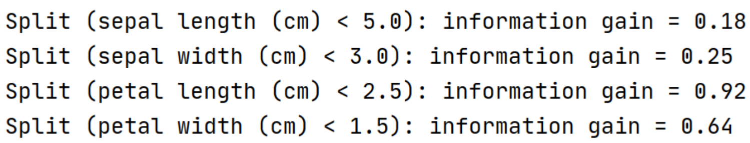
Construction of Decision Trees using 'Iris flower' dataset.

2.a

See *ex2.py*.

2.b

Output for the given splits:



```
Split (sepal length (cm) < 5.0): information gain = 0.18
Split (sepal width (cm) < 3.0): information gain = 0.25
Split (petal length (cm) < 2.5): information gain = 0.92
Split (petal width (cm) < 1.5): information gain = 0.64
```

2.c

I would select (petal length (cm) < 2.5) to be the first split, because it shows the maximum information gain among the four splits.

2.d

The mean accuracy is 95.33%.

For the original data, the two most important features are:
petal length (index: 2),
petal width (index: 3)

For the reduced data, the most important feature is:
petal length (index: 2)

The importance score for the 'petal length' is 1 for every fold (reduced dataset).
As this is a normalized score, all other features have a feature importance of 0 for this dataset. This means that in the scenario where only iris flowers of species 0 and 1 are looked at, the single split according to the 'petal length' already yields *pure* classes and the flowers can correctly be classified with respect to this attribute.