Logistic Regression with Adaboost

1605004: Md. Redwanul Haque

How to run:

Changing variables:

- changing learning rate: on line 253, change value of `alpha`
- changing maximum iteration of gradient descent: on line 254, change value of `max_iterations`
- changing count of features to keep based on information gain: on line 255, change value of `features_to_use`. Set value to 0 to use all features.
- changing error cutoff of gradient descent: on line 256, change value of `error_cutoff`
- showing learning curve of logistic regression: on line 257, change value of `show_plot`
- changing how many hypotheses to use in Adaboost: on line 258, change value of `hypothesis_count`
- chaning how many negative samples are taken in dataset 3: on line 134 change value of `n`

Changing which dataset to run:

1. dataset 1 on: line 260
2. dataset 2 on: line 261
3. dataset 3 on: line 262

Changing dataset input files:

1. dataset 1 on: line 71
2. dataset 2 on: line 92, 93
3. dataset 3 on: line 129

lines 265-268: runs a single Logistic Regression model
lines 271-276: runs Adaboost with multiple LR models

Performance:

Best Logistic regression only learners:

Dataset 1: alpha = 0.1, max_iter = 5000

Performance measure	Training	Test
Accuracy	80.58	80.2
True positive rate (recall)	0.55	0.52
True negative rate (specificity)	0.899	0.903
Positive predictive value (precision)	0.664	0.653
False discovery rate	0.336	0.347
F1 score	0.6	0.58

Dataset 2: alpha = 0.1, max_iter = 5000

Performance measure	Training	Test
Accuracy	84.1	84.3
True positive rate (recall)	0.56	0.56
True negative rate (specificity)	0.93	0.93
Positive predictive value (precision)	0.717	0.713
False discovery rate	0.283	0.287
F1 score	0.63	0.628

Dataset 3: alpha = 0.1, max_iter = 5000

Performance measure	Training	Test
Accuracy	98.95	99.17
True positive rate (recall)	0.573	0.638
True negative rate (specificity)	0.99	0.99
Positive predictive value (precision)	0.99	0.99
False discovery rate	0.009	0.0
F1 score	0.726	0.779

Adaboost:

Dataset 1: alpha = 0.1, max_iter = 100, cutoff = 0.9, features_used = 5

Number of boosting rounds	Training	Test
5	78.54	77.43
10	78.54	77.43
15	78.54	77.36
20	78.2	77.3

Dataset 2: alpha = 0.1, max_iter = 100, cutoff = 0.9, features_used = 5

Number of boosting rounds	Training	Test
5	82.23	82.57
10	81.0	81.18
15	81.02	81.2
20	82.33	82.56

Dataset 3: alpha = 0.1, max_iter = 100, cutoff = 0.9, features_used = 5

Number of boosting rounds	Training	Test
5	97.57	97.71
10	99.35	99.71
15	99.35	99.71
20	99.35	99.71