PS4 Machine Learning: Project

Cai Glencross and Katie Li

February 14, 2018

1 Perceptron:

b)

$\theta^{(0)}$	θ^*	mistakes
$(0,0)^T$	$(3.202, 2.128)^T$	1
$(1,0)^T$	$(0.912, 0.79)^T$	4

No, they do not converge to the same solution because they have different initial starting conditions. While we are guaranteed to converge to a solution for a given linearly-separable dataset, we are not guaranteed to converge to the same solution. The perceptron algorithm checks for whether each data point is correctly separated with the given θ, and once it is, it stops.

Since both classifiers correctly separate each point in the training dataset, they have the same performance for the training data. However, they probably do not have the same performance on a test set because the coefficient values (θ) are different.

c) The mistake bound is 675.2, which is much greater than the number of total mistakes in either of the cases starting with the initial $\theta^{(0)}$, of 1 and 4 mistakes. However, the mistake bound does not represent average number of mistakes but the maximum mistakes.

It is possible that we could have θ with very small geometric margins (resulting in a small γ term), and would sharply increase the mistake bound. γ^* is calculated in the given method based on the smallest L2 norm of the θ. So, a small norm would mean that we would have a very large mistake bound.

Finally, there are only 30 data points, and the mistake bound does not depend on the number of features, only R and γ. It appears unlikely to make over 600 mistakes with only 30 data points. Since the mistake bound does not depend on the particular data points, and must hold for the given R and gamma, it is higher than the experimental values.

d) In order to make the greatest number of mistakes, we would want to change the coefficients θ many times. First, we can calculate the mean distance of

the points. Then, we choose the point furthest away from the origin, called point q to test on θ. We test the other points in the order of q.

Since the algorithm first would find mainly one data label, it would continue to shift after switching to the other ones on the cluster.

2 Perceptron vs Logistic Regression

(a) Yes, the data is linearly separable because the perceptron algorithm found coefficients which separated the dataset. This is not surprising because it is a high dimensional dataset, of around 5 data points per feature, and a large feature space. Given higher dimensions, it would be easier to separate the data, and even over-fit it to noise.

(b) see above

accuracies

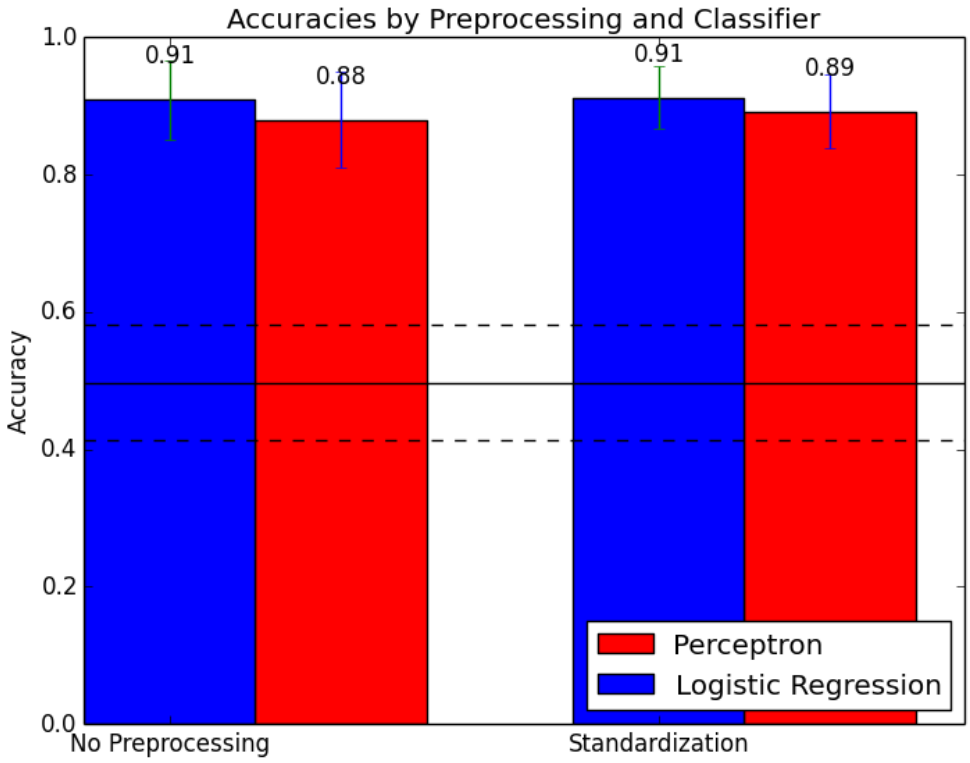
classifier	μ	σ
no preprocessing		
P	0.885	0.0701
L	0.907	0.0441
D	0.431	0.0547
with standardization		
P	0.892	0.0597
L	0.905	0.0463
D	0.431	0.0547

(c) see above

(d) see above

p-values

		no preprocessing			standardization		
		P	L	D	P	L	D
no preprocessing	P	-	.0024	1.9e-71	0.373	9e-4	1.3e-60
	L	-	-	1.9e-84	6.7e-3	0.32	1.4e-83
	D	-	-	-	1e-71	2e-84	nan
standardization	P	-	-	-	-	0.005	1.3e-80
	L	-	-	-	-	-	1.4e-83
	D	-	-	-	-	-	-


Accuracies by Preprocessing and Classifier
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0.91
0.88
0.91
0.89
Perceptron
Logistic Regression
No Preprocessing
Standardization

(e) see above