Registration Number: 2015/IS/038

Index Number: 15020381

IS4003 – Data Analytics

Assignment A on Machine Learning

- Dataset - Breast Cancer Wisconsin dataset from UCI data repository

http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

- Attributes

1.	Sample code number	id number
2.	Clump Thickness	1 – 10
3.	Uniformity of Cell Size	1 – 10
4.	Uniformity of Cell Shape	1 – 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class	2 for benign, 4 for malignant

- 16 missing attribute values were replaced with zero
- Replaced 11th attribute which is the class number (2 or 4) with (-1, +1)
- Removed the 1st attribute which is a unique identifier
- Results

PA

```
Algorithm:PA
Number of iterations:1
W: [-0.171468, 0.514711, 0.249339, 0.132195, -0.697327, 0.302809, -0.298703, 0.146631, -0.413485]
Training Accuracy:82.8326%
Test Accuracy:93.133%
Number of iterations:2
W: [-0.183447, 0.670755, 0.274053, 0.170809, -0.856113, 0.335139, -0.366811, 0.119813, -0.474911]
Training Accuracy:82.8326%
Test Accuracy:93.5622%
Number of iterations:10
W: [-0.190542, 0.705904, 0.282395, 0.180751, -0.895551, 0.345892, -0.383697, 0.118087, -0.492463]
Training Accuracy:82.618%
Test Accuracy:93.5622%
```

PA - I

```
Algorithm:PA-I
Number of iterations:1
W: [-0.171468, 0.514711, 0.249339, 0.132195, -0.697327, 0.302809, -0.298703, 0.146631, -0.413485]
Training Accuracy:82.8326%
Test Accuracy:93.133%
Number of iterations:2
W: [-0.183447, 0.670755, 0.274053, 0.170809, -0.856113, 0.335139, -0.366811, 0.119813, -0.474911]
Training Accuracy:82.8326%
Test Accuracy:93.5622%
Number of iterations:10
W: [-0.190542, 0.705904, 0.282395, 0.180751, -0.895551, 0.345892, -0.383697, 0.118087, -0.492463]
Training Accuracy:82.618%
Test Accuracy:93.5622%
```

PA - II

```
Algorithm:PA-II
Number of iterations:1
W: [-0.170935, 0.508179, 0.246931, 0.128491, -0.688027, 0.30294, -0.298107, 0.148707, -0.409149]
Training Accuracy:82.618%
Test Accuracy:93.5622%
Number of iterations:2
W: [-0.182174, 0.66311, 0.269889, 0.167382, -0.844297, 0.334431, -0.36449, 0.120674, -0.469359]
Training Accuracy:82.8326%
Test Accuracy:93.133%
Number of iterations:10
W: [-0.189146, 0.698081, 0.27796, 0.177175, -0.883046, 0.34504, -0.381256, 0.118516, -0.485697]
Training Accuracy:82.618%
Test Accuracy:93.5622%
```

It can be seen that the training accuracy is similar when the number of iterations equals 1 and 2, but the test accuracy is similar when the number of iterations equals 2 and 10 in PA and PA-I. But in PA-II training accuracies and testing accuracies are similar when the number of iterations is 1 and 10. When considering the three variants of the passive aggressive algorithm it can be seen that the test accuracy is greater than training accuracy in each number of iterations.

Since the three variants are having 93% accuracy for test data, we can conclude that the algorithms have classified the breast cancer dataset well.