Introduction to Data Mining: 21191441
Zhejiang University May 13, 2019
Professor Deng Cai Homework 2

Homework 2

Collaborators:

Name: Tian Zijun
Student ID: 3160104043

Problem 2-1. A Walk Through Linear Models

(a) Perceptron

Answer:

1. nTrain = 10, train error = 0, test error = 0.1106
   nTrain = 100, train error = 0, test error = 0.0136
2. nTrain = 10, average iteration = 5.748, nTrain = 100, average iteration = 40.983
3. the two classes cannot be divided by a line, this function will be a infinite loop

```
E_train is 0.0, E_test is 0.1105915
Average number of iterations is 5.41.
```

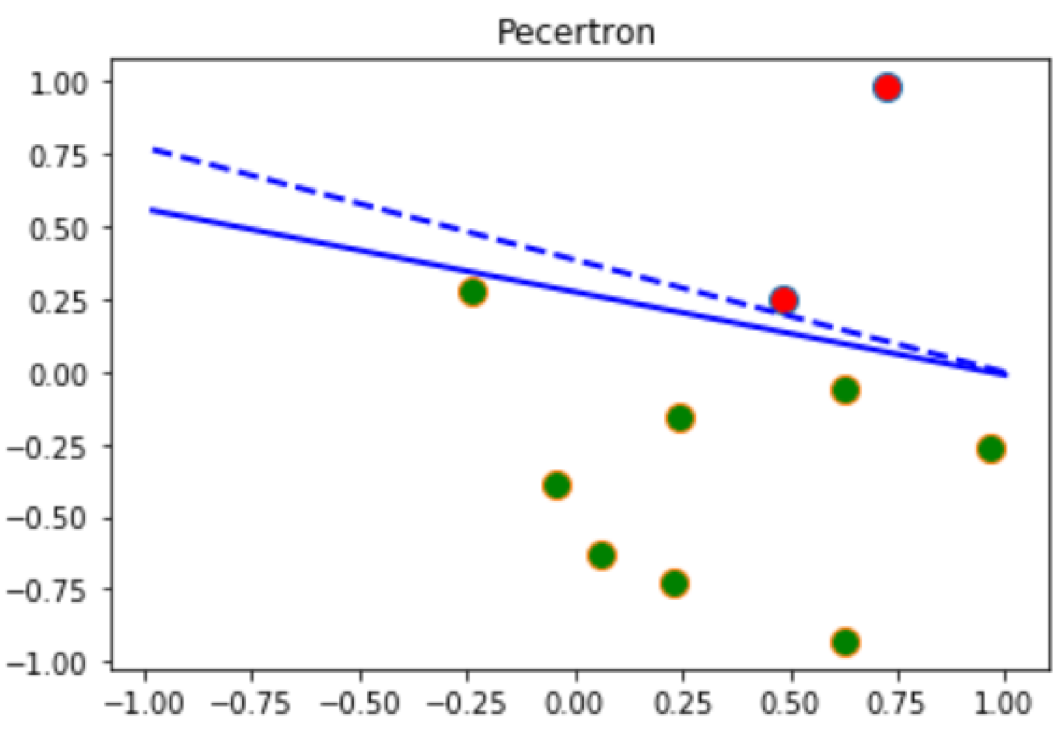

Figure 1: The plotting result for perceptron when nTrain = 10.

```
E_train is 0.0, E_test is 0.013635
Average number of iterations is 40.659.
```

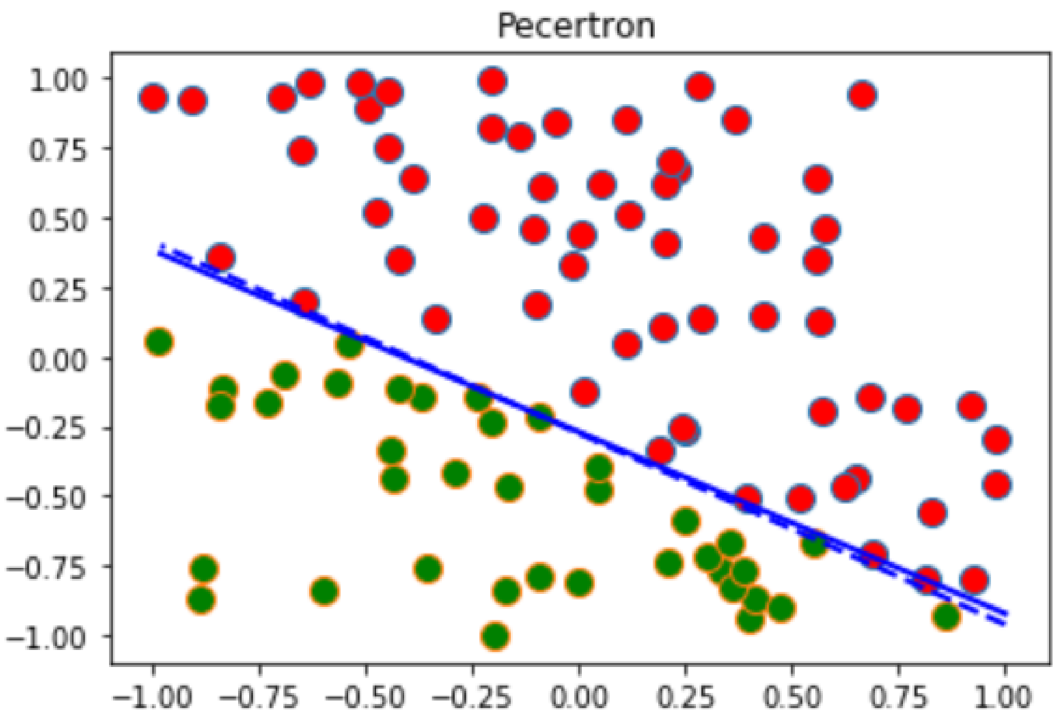

Figure 2: The plotting result for perceptron when nTrain = 100.

```
E_train is 0.26, E_test is 0.2202
Average number of iterations is 10001.0.
```

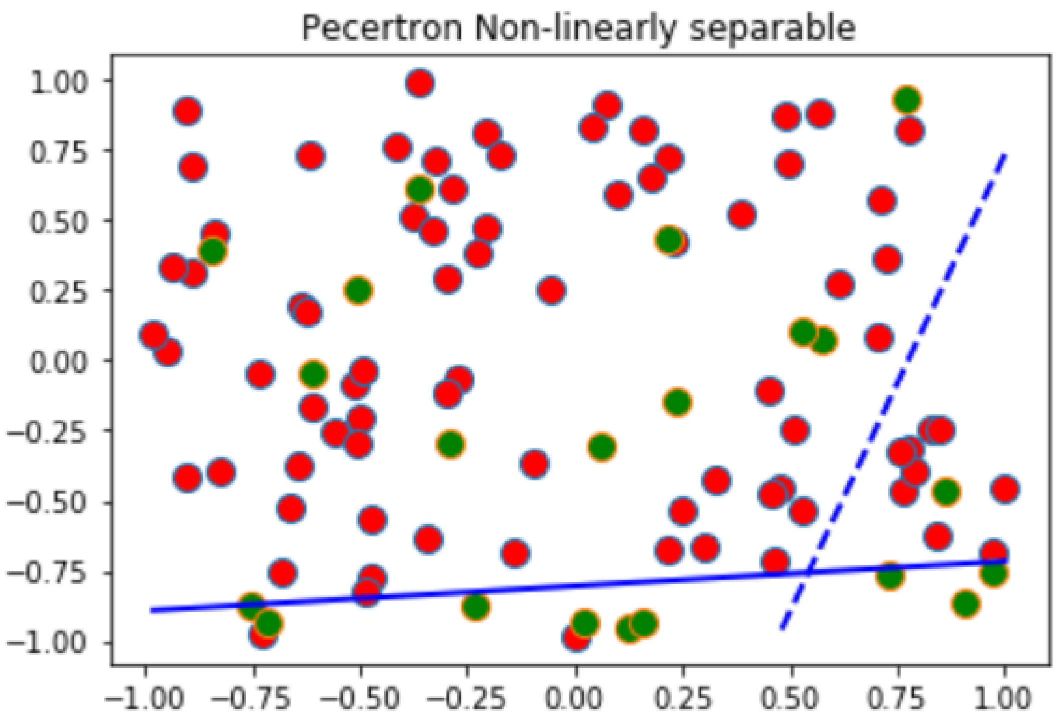

Figure 3: The plotting result for perceptron when training data is not linearly seperable.

(b) Linear Regression

Answer:

1. training error rate is 0.0408, testing error rate is 0.0506
2. training error rate is 0.13427, testing error rate = 0.14538
3. training error rate = 0.49, testing error rate = 0.5496
4. training error rate = 0.05, testing error rate = 0.066

```
E_train is 0.04013, E_test is 0.0486292
```

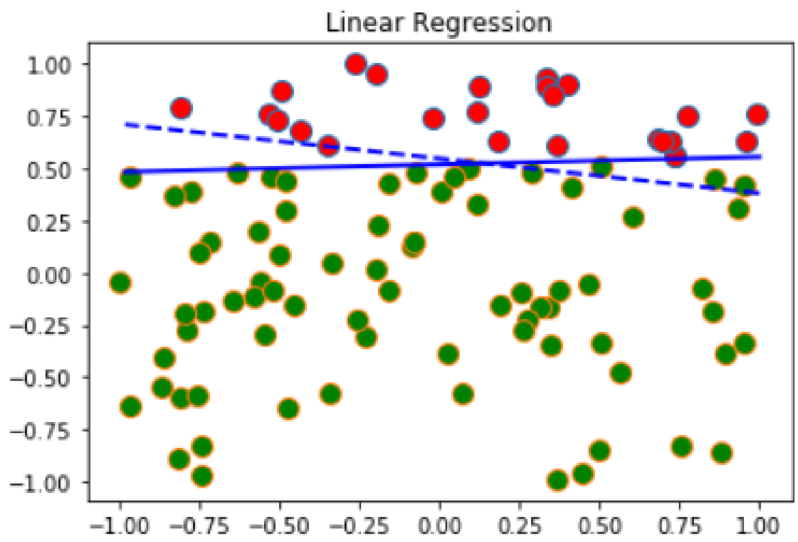

Figure 4: The plotting result for linear regression.

```
E_train is 0.13296, E_test is 0.14443910000000001
```

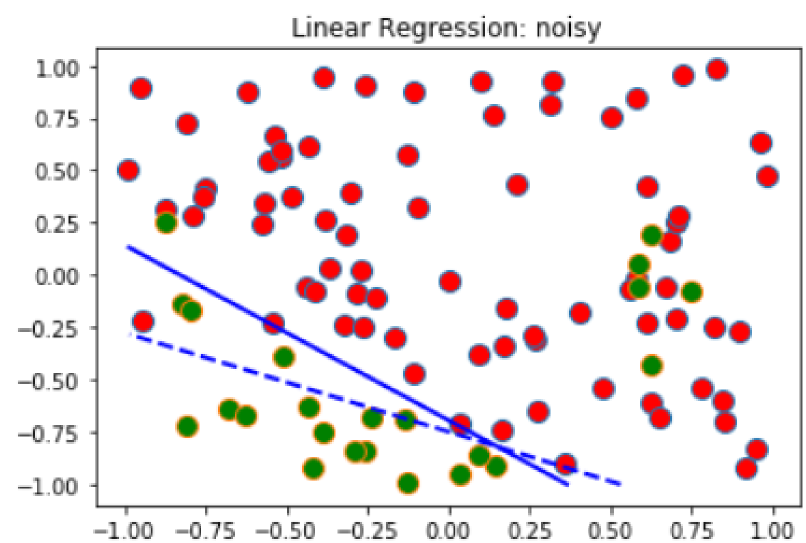

Figure 5: The plotting result for linear regression when training data is not linearly seperable.

(c) Logistic Regression

Answer:

1. training error = 0.0086, testing error = 0.017467

2. training error = 0.13, testing error = 0.1493

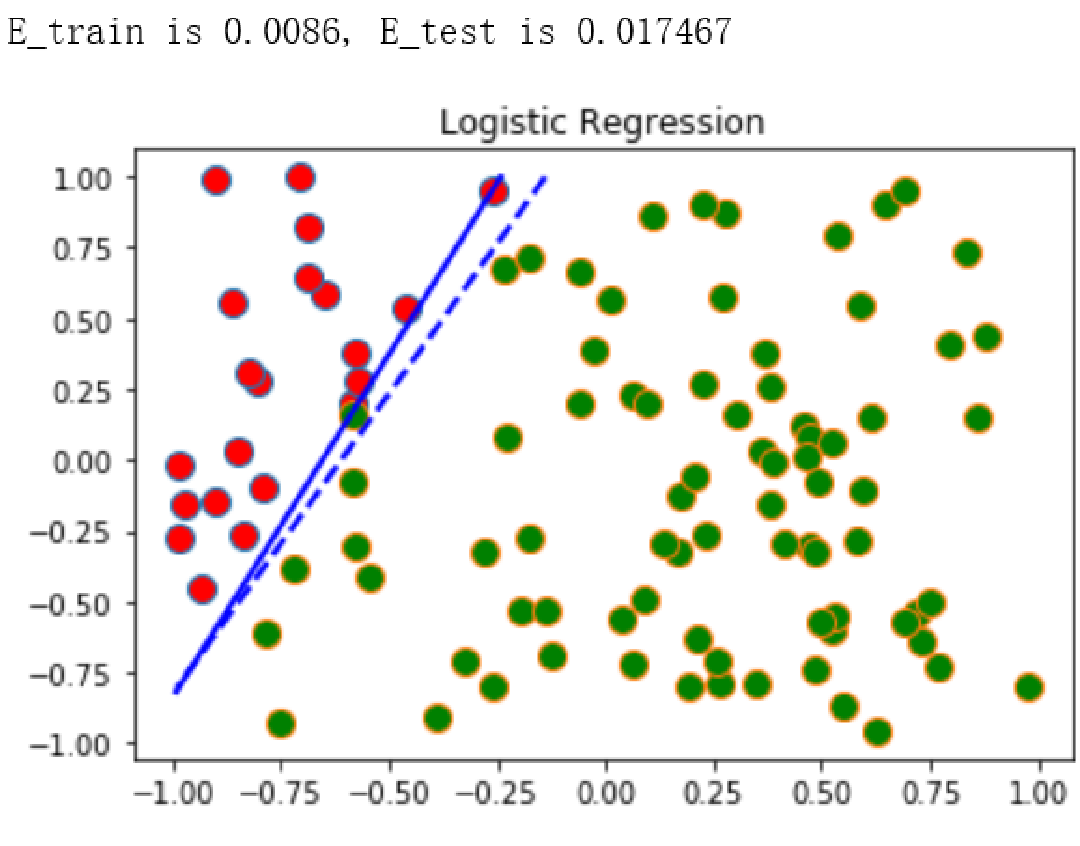

Figure 6: The plotting result for logistic regression.

E_train is 0.13, E_test is 0.1493

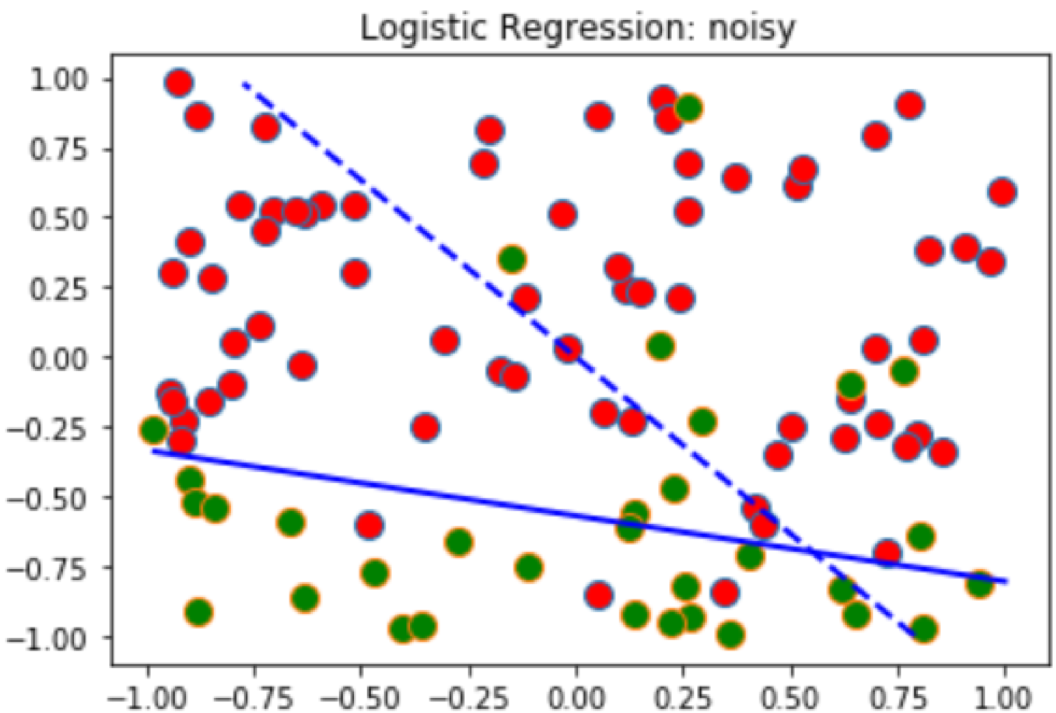

Figure 7: The plotting result for logistic regression when training data is not linearly seperable.

(d) Support Vector Machine

Answer:

1. training error = 0, testing error = 0.03546
2. training error = 0, testing error = 0.01010
3. support vector = 3.094

```
E_train is 0.0, E_test is 0.0101056
Average number of support vectors is 3.094.
```

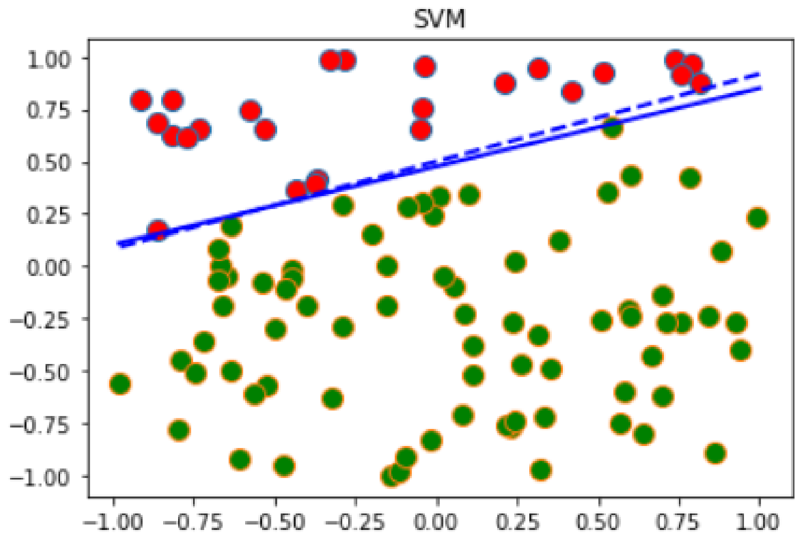

Figure 8: The plotting result for SVM when nTrain is 100.

Problem 2-2. Regularization and Cross-Validation

(a) Implement Ridge Regrssion, and use LOOCV to tune the regularization parameter λ.

Answer:

1. $\lambda = 100$
2. $\lambda = 0, \Sigma_i w^2 = 1.02, \lambda = 100, \Sigma_i w^2 = 0.133$
3. $\lambda = 0$, training error = 0, testing error = 0.126
   $\lambda = 100$, training error = 0, testing error = 0.0598

(b) Implement Logistic Regrssion, and use LOOCV to tune the regularization parameter λ.

Answer:

Problem 2-3. Bias Variance Trade-off

Let's review the bias-variance decomposition first. Now please answer the following questions:

(a) True of False

Answer:

1. False, adding more training examples will improve the training error
2. False, low variance is able to better fit the test set

3. True
4. False, regularization always results in equal or worse performance on training set
5. False, regularization parameter λ will hurt the performance