Introduction to Data Mining: Assignment #2

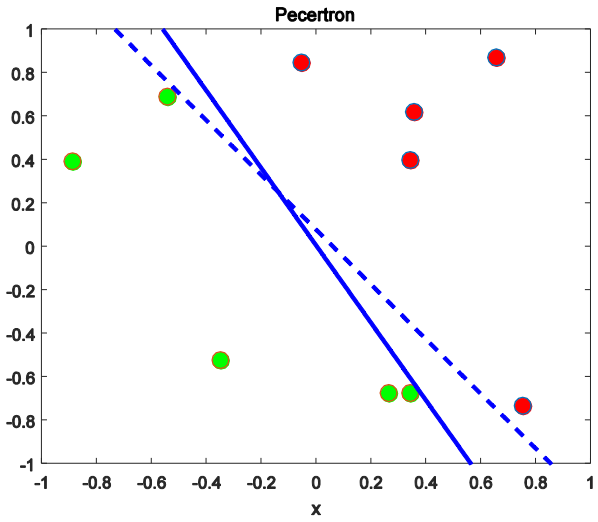
Junpei Zhou

3140102286

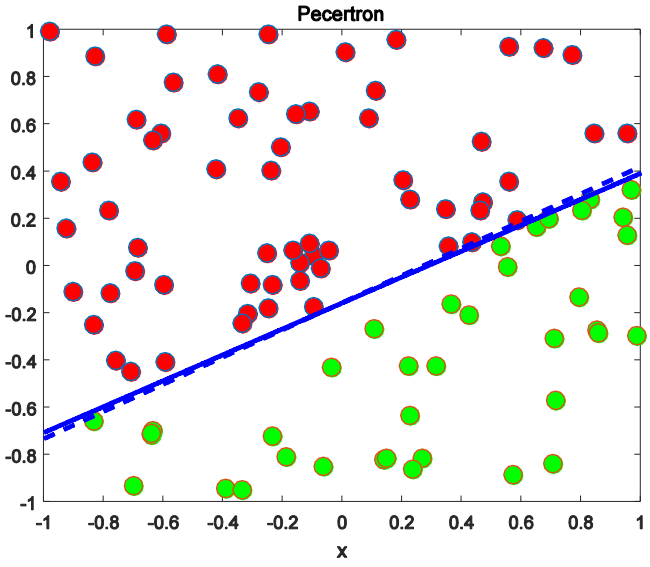
1. A Walk Through Linear Models

a) Perceptron

(i) When the size of training set is 10, the training error rate is 0.0, and the expected testing error rate is 0.136383. The plot is shown as below:



When the size of training set is 100, the training error rate is 0.0, and the testing error rate is 0.014594. The plot is shown as below:

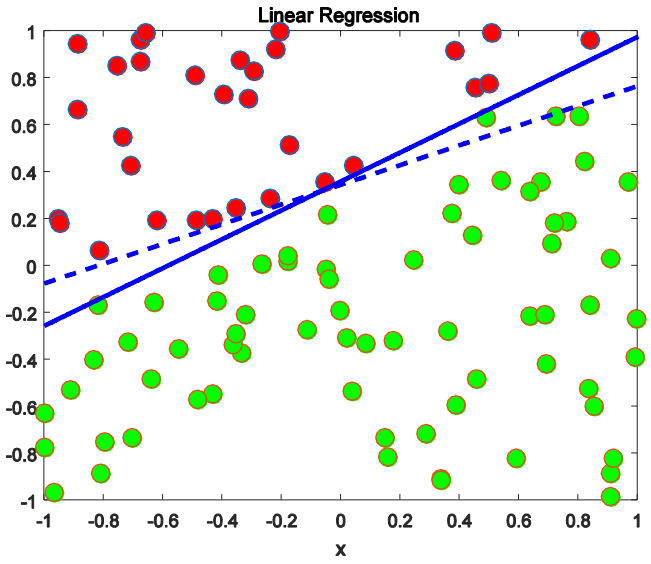


(ii) When the size of training set is 10, the average number of iterations is 3.3566. When the size of training set is 100, the average number of iterations is 3.4793.

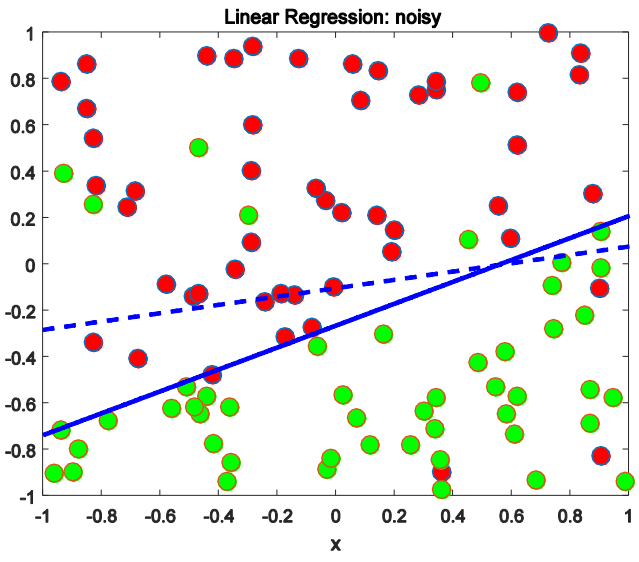
(iii) When the training data is not linearly separable, the algorithm will be in infinite loop.

b) Linear Regression

(i) When the size of training set is 100, the training error rate is 0.039700, and the testing error rate is 0.049288. The plot is shown as below:



(ii) When the size of training set is 100, and the training data is noisy and not linearly separable, the training error rate is 0.137930, and the testing error rate is 0.148424. The plot is shown as below:

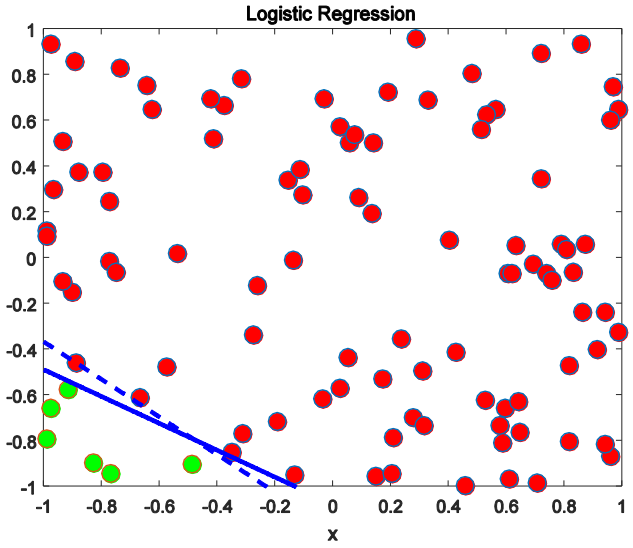


(iii) Run Linear Regression on dataset poly train.mat, the training error rate is 0.490000, and the testing error rate on poly test.mat is 0.549600.

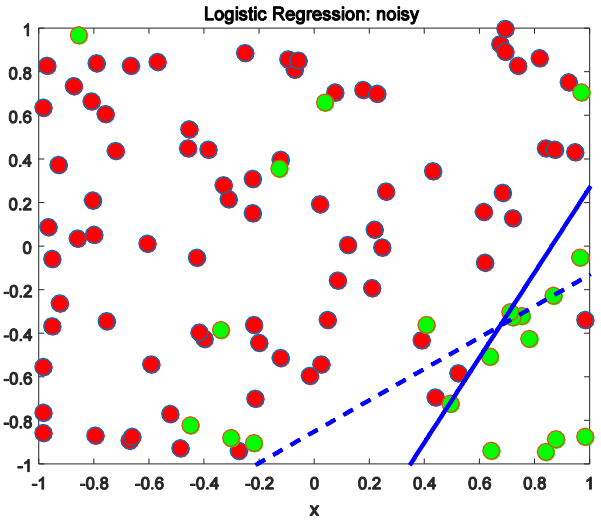
(iv) After this transformation, the training error rate is 0.050000, and the testing error rate is 0.066000.

c) Logistic Regression

(i) When the size of training set is 100, the training error rate is 0.010500, and the testing error rate is 0.018026. The plot is shown as below:

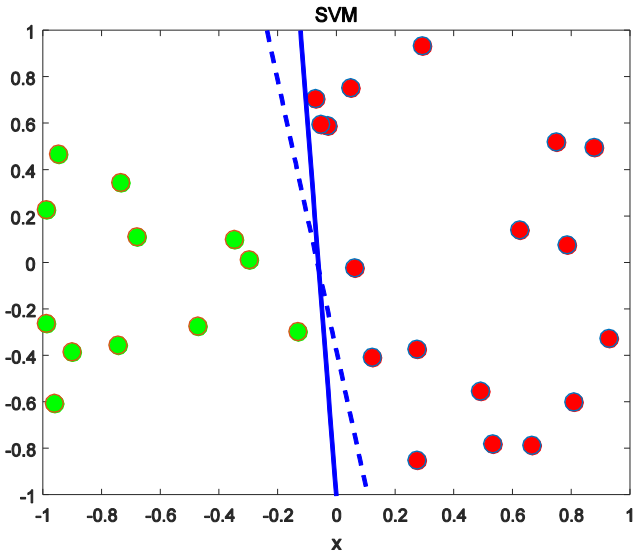


(ii) When the size of training set is 100, and the training data is noisy and not linearly separable, the training error rate is 0.139600, and the testing error rate is 0.148570. The plot is shown as below:

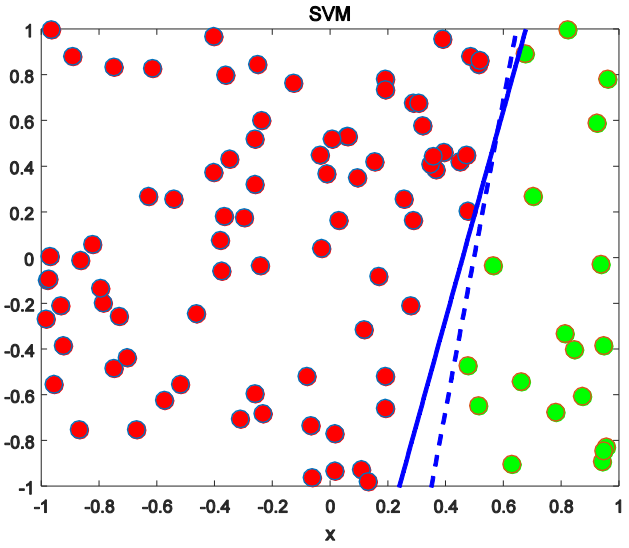


d) Support Vector Machine

(i) When the size of training set is 30, the training error rate is 0.0, and the testing error rate is 0.035179. The plot is shown as below, and we can see that the classifier got by SVM has max margin:



(ii) When the size of training set is 100, the training error rate is 0.0, and the testing error rate is 0.010512. The plot is shown as below:



(iii) For the case nTrain = 100, the average number of support vectors is 3.420000.

2. Regularization and Cross-Validation

a) Implement Ridge Regression

(i) LOOCV choose $\lambda = 100$.

(ii) With regularization, $\sum_{i=1}^{m} \omega_i^2 = 0.133222$

Without regularization, $\sum_{i=1}^{m} \omega_i^2 = 8634.450112$

(iii) With regularization, the training error rate is 0.0, the testing error rate is 0.065294.

Without regularization, the training error rate is 0.375000, the testing error rate is 0.500251.

b) Implement Logistic Regression with regularization

LOOCV choose $\lambda = 1$.

With regularization, $\sum_{i=1}^{m} \omega_i^2 = 23.659372$

Without regularization, $\sum_{i=1}^{m} \omega_i^2 = 171463.206106$

With regularization, the training error rate is 0.0, the testing error rate is 0.056755.

Without regularization, the training error rate is 0.000000, the testing error rate is 0.065796.

3. Bias Variance Trade-off

a) True of False

(i) False

(ii) False

(iii) True

(iv) False

(v) False