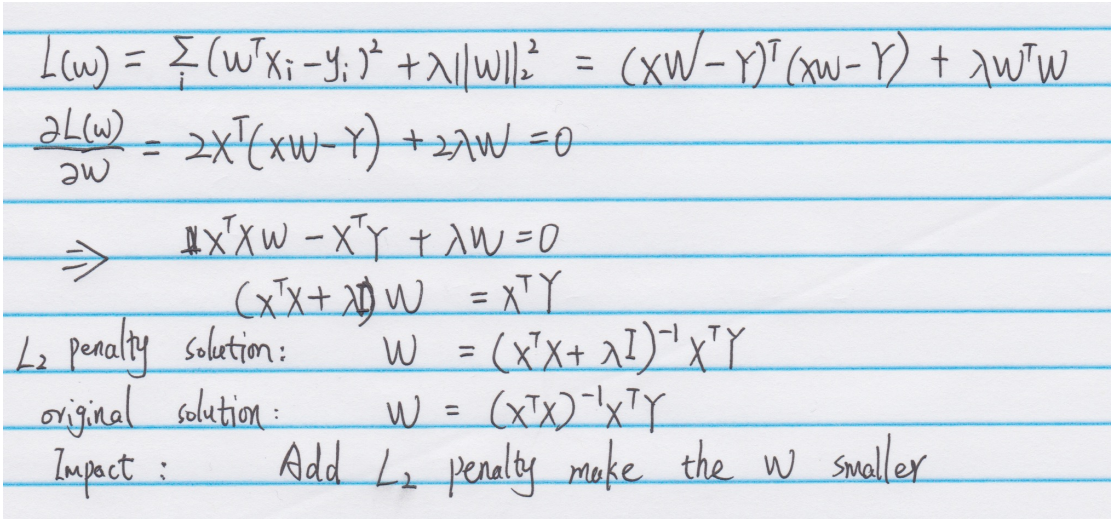
Homework 2
Tianyang Chen

Part A

1. Linear Regression and Regularization



2. Density Estimation

If k=n, which means we use all the data to estimate the classification of a given point. In this case, all the points will be determined in the same classification. The training error will be the sum of n/2 points' error, which is very large.

If k=1, the error will have more chance to happen in the overlapped range, if the overlapped range is large, then the training error may still be large.

In another word, there must be some $k = k_0, k_0 \in [1, n]$, that minimizes the training error.

3. Feature Selection & Preprocessing

The feature selection method he used may be computationally expensive, however, if he tried large number of subsets, he would get the most relevant features.

During the validation procedure, if he splits the data by 80% / 20%, then 50 times repeats will be too many. In this way, he has a high probability of getting at least 2 repeats which has similar training & testing sets. Thurs, I recommend using 5-fold cross validation.

Part B

1. Logistic Regression on Synthetic and Real-World Data

Using MATLAB, I write a function to compute the parameters W using stochastic gradient descent of cross-entropy. It will display the vector W in command window.

==========

function logisticR (Inputfile, Iteration_times)

==========

Inputfile – the name of input file

Iteration_times – iteration times during stochastic gradient descent

==========

Example:

logisticR ('balloons.csv', 500);

---

2. Dimensionality Reduction

I use Discriminant Analysis Classifier for this problem. In MATLAB, I write a function which will display the 5-fold cross validation error in command window for both with and without PCA.

==========

function PCA_compare (InputFile)

==========

InputFile - the name of input file

==========

Example:

PCA_compare ('contrac.csv');

Table 1 5-fold cross validation error – PCA

	5-fold cross validation error	
filename	without PCA	with PAC
abalone.csv	0.39095	0.405794
adult_test.csv	0.187335	0.189423
balance-scale.csv	0.0832	0.0864
balloons.csv	0.3125	0.25
bank.csv	0.147755	0.128733
breast-cancer.csv	0.297203	0.307692
breast-cancer-wisc.csv	0.04721	0.04721
breast-tissue.csv	0.330189	0.311321
connect-4.csv	0.295913	0.297467
contrac.csv	0.490835	0.507807

In this table, I find that After implementing PCA, sometimes it will significantly smaller the error, even if the error gets bigger, it will still be close to the error without PCA.

3. Density Estimation

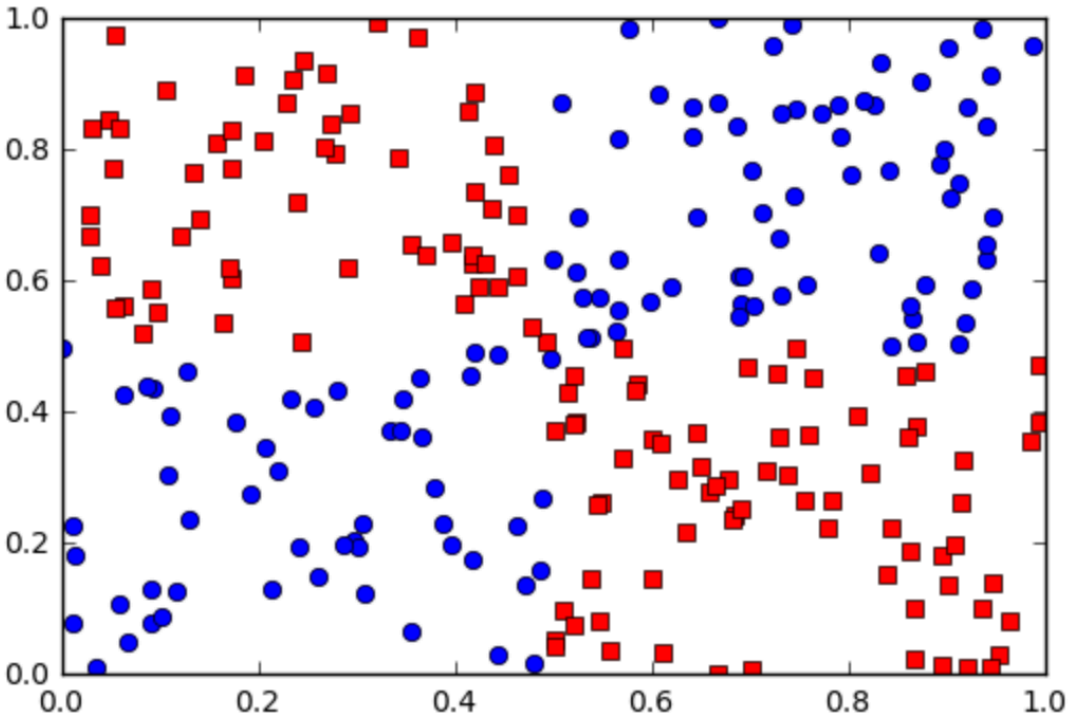


Fig.1 Traing Data 1

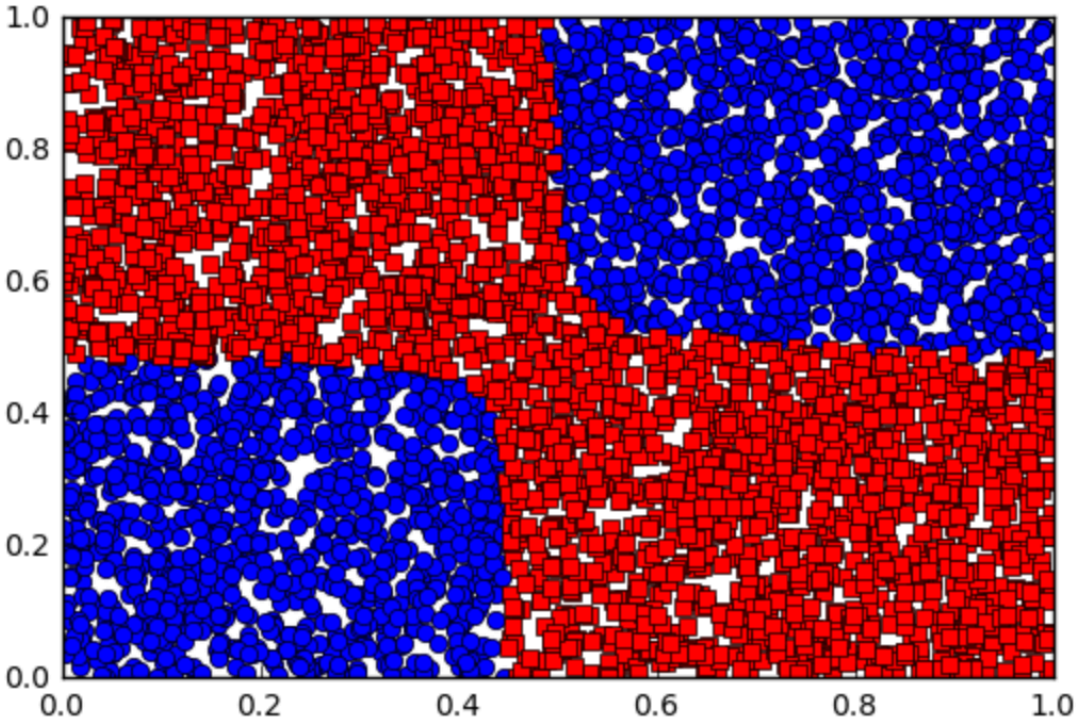


Fig.2 classified new data

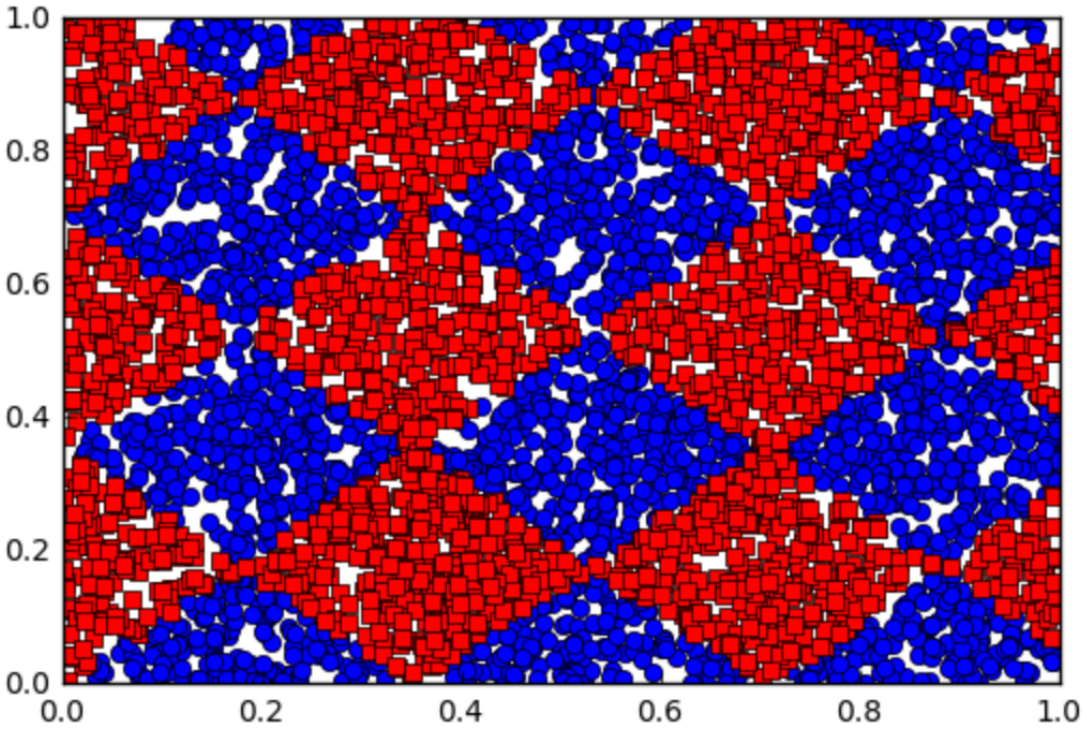


Fig.3 Training Data

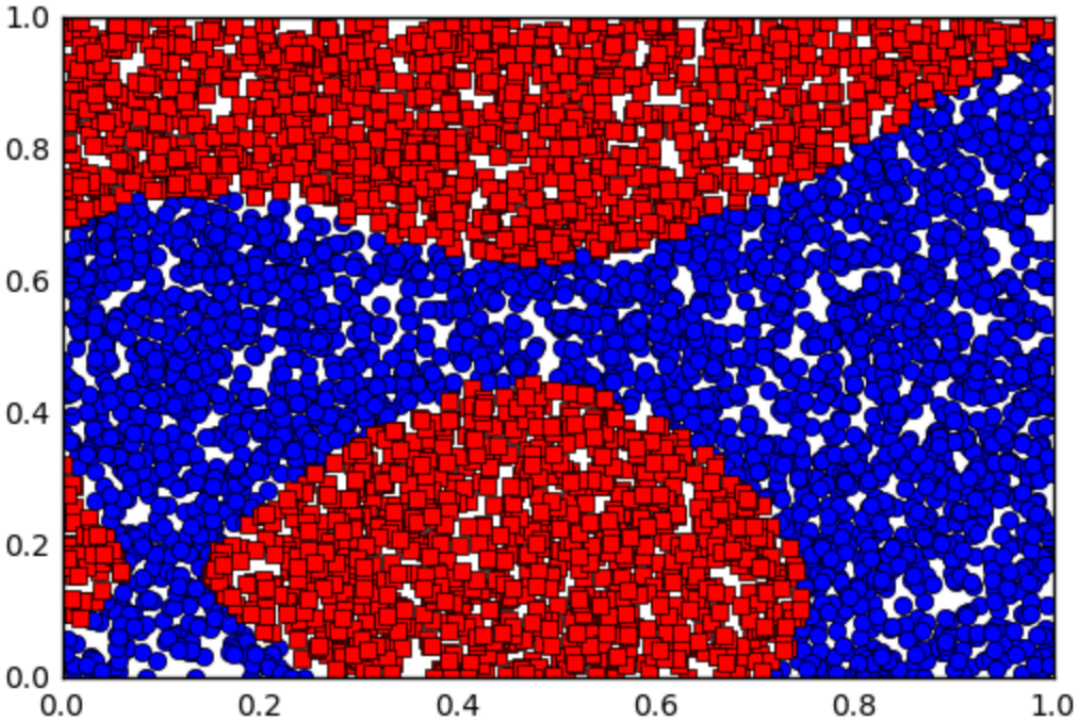


Fig.4 classified new data