Korea Advanced Institute of Science and Technology

School of Electrical Engineering

EE488 Introduction to Machine Learning Spring 2018

Name: Dinh Vu

Student ID: 20184187

**Assignment 1 (Problem 1-4)**

1. **Multilayer Perceptron**



is activation unit

is matrix of weight controlling function mapping from layer i to i + 1.

and are bias parameters, equal to 1.

So,

Assuming that we have a point is a test point. D is in the shade area or in triangle ABC if and only if , otherwise D is outside the shade area. We can choose

Final result: ,

1. **Robust Linear Regression**
2. Gaussian Distribution

Let for i = 1, 2, … , N

The loss function is the ordinary least squares, so:

Assume that is invertible.

1. Laplace Distribution

In matrix notation: where

So: with is invertible.

1. Robust fit data

|  |  |
| --- | --- |
| **Gaussian model** | **Laplace model** |
|  |  |

**Table 2.1.** Optimization function of Gaussian and Laplace model

Assuming that at the iteration number *t*, the acceptable hypothesis space is found with and at the next iteration step, there is a training data sample with great noise , the new hypothesis space with is estimated.

In Gaussian model, the value of increase rapidly because of square. Then the new hypothesis space at iteration *t* *+* 1 will move far away to the hypothesis space at iteration *t*. Hence, for the test data in Figure 2.1, it could be predicted with heavy error.

While in Laplace model the value of still increase but not so much compare to Gaussian model. Therefore, the new hypothesis space at iteration *t* + 1 will not change greatly and the test data is still predicted with good error.

To sum up, the Laplace model is more robust fit data than Gaussian noise.



**Figure 2.1.** The hypothesis space at iteration *t* and *t+*1 of Gaussian (a) and Laplace model (b)

1. **SVM Error Analysis**
2. Show that the expectation error rate is equal to the expectation of leave-one-out cross validation (LOOCV) error for *n* + 1 data points.

* Case 1:
* Step 1: Generating randomly n + 1 data point, then using SVM classification for n training data point to find a hypothesis space. To test the hypothesis space, is chose to be test data point. The expected error rate *err* is found.
* Step 2: Repeat step 1 to *N* times.

The expected error rate of case 1 is calculated by:

Let because all data points is drawn randomly with the same distribution

If , then

* Case 2:

Generating randomly n + 1 data points. Using LOOCV, the expectation of error for the i-th model is:

The expectation of LOOCV error in entire n + 1 different model is:

Finally, the expectation error rate is equal to the expectation of leave-one-out cross validation (LOOCV) error for *n* + 1 data points.

For the i-th LOOCV model of n + 1 data points, is the test point and the rest is the training data. After using SVM classification for the training data set , the hypothesis space with two margin is found, such as Figure 3.1. There are two case for test point .

First, is not support vector or in other word is classified correctly with the hypothesis space of the i-th model presented in Figure 3.1a

Second, is a support vector that means there is other hypothesis space better than the i-th model and the distant from test point to margin of the i-th model is less than 1 or classified incorrectly. That is shown in Figure 3.1b.

Hence the error rate of the i-th model is represented as below:



**Figure 3.1.** SVM classifier with LOOCV

1. **Gradient Descent**

According to the first order Taylor expansion:

Hence:

Let

Because is the minimum value of , we have:

. This is the basic formula of Gradient Descent.