CSCI567 2014 Homework Assignment 1

Student Name Yi Duan

USC ID 9966845562

I have collaborated with Meng Sun

Mingxuan Yue

1. Naive Bayes
   1. Parametric form of Naive Bayes with Gaussian Assumption

According to Bayes theorem,

*∴,*

*, for j = 1,…,D*

* 1. Parameter Estimation for Naive Bayes with Gaussian Assumption

Given that D=,

Estimating {}:

Suppose the number of data points label as 0 is m and the number of data points label as 1 is n. Then,

s.t. , m+n=N

we use the Lagrangian multiplier

L()=

Taking derivatives respectively with

Then we find

That is

∴

Estimating {}:

we use the Lagrangian multiplier

Taking derivatives respectively with ,

∴,

1. Nearest Neighbor
2. First we discretize the weights, for example, by 0.1. We set w1=0,w2=0,w3=1 for the first time, apply the three weights to the training data, calculate the distance and classify the data with different labels measure the performance of the classifier with the test data. Then we set w1=0,w2=0.1,w3=0.9 and repeat this process. We enumerate all the possible values for the three weights. Finally we compare their results and get the best weights.
3. When the feature dimensionality is very high, exhaustic search would cost too much. For example, we use D=10000 and discretize the weights by 0.1, then we would classify the data for 1010000 times, which is unacceptable. Given this condition, we can use coordinate descent to solve this problem. In each iteration, we choose one feature and does line search along the corresponding coordinate direction. In this way, we can find a local optimal with less iterations.
4. Logistic Regression
5. Probability of a single training example

Compact expression

Log-likelihood of the whole training data D

The negative log-likelihood is

1. The derivatives of is

Then,

For any vector ***v***,

Where is the i.th number of ***v***

Thus, is convex.

1. According to (b), the derivatives of is

Suppose. Since () () is linearly separable, there exists a situation that

Since is finite, can go to infinity.

1. For each , we have
2. Suppose

Since , is convex,  is convex. Thus, has a unique solution.

1. Decision Tree
2. The entropy for weather is

The entropy for traffic is

Since the entropy for traffic is smaller than that for weather, we choose traffic to split in the first step to maximize the information gain.

1. Suppose the dataset is (**X**i,Yi) (i=1,…,n). When building the decision tree, we compare the information gain at each branch node. When is fixed, we only need to compare .

For T1,

For T2,

Since X→X’ is bijection, and for k=1,…,n. Thus, T1 and T2 is equivalent.

1. In order to prove

We should prove

Suppose

Let . Then

When

Thus,

Since , there exists and

When

Since

Thus, . That is, for each ,. Thus,

1. Programming

5.4 Performance Comparison

kNN:

|  |  |  |  |
| --- | --- | --- | --- |
| k | Training | Validation | Test |
| 1 | 0.8326 | 0.8201 | 0.8535 |
| 3 | 0.9021 | 0.8920 | 0.9023 |
| 5 | 0.9263 | 0.9229 | 0.9409 |
| 7 | 0.9316 | 0.9254 | 0.9306 |
| 9 | 0.9242 | 0.9229 | 0.9254 |
| 11 | 0.9189 | 0.9177 | 0.9203 |
| 13 | 0.9147 | 0.9075 | 0.9126 |
| 15 | 0.9147 | 0.8946 | 0.9177 |
| 17 | 0.9137 | 0.8972 | 0.9075 |
| 19 | 0.9074 | 0.8843 | 0.9049 |
| 21 | 0.8979 | 0.8740 | 0.9075 |
| 23 | 0.8926 | 0.8843 | 0.9049 |

Decision Tree:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Gini-index | | | Cross-entropy | | |
| K | Training | Validation | Test | Training | Validation | Test |
| 1 | 0.9432 | 0.9332 | 0.9434 | 0.9516 | 0.9332 | 0.9486 |
| 2 | 0.9411 | 0.9332 | 0.9434 | 0.9495 | 0.9332 | 0.9486 |
| 3 | 0.9411 | 0.9383 | 0.9434 | 0.9495 | 0.9383 | 0.9486 |
| 4 | 0.9411 | 0.9383 | 0.9434 | 0.9495 | 0.9383 | 0.9486 |
| 5 | 0.9400 | 0.9409 | 0.9460 | 0.9484 | 0.9409 | 0.9512 |
| 6 | 0.9463 | 0.9460 | 0.9434 | 0.9547 | 0.9460 | 0.9486 |
| 7 | 0.9453 | 0.9486 | 0.9460 | 0.9505 | 0.9486 | 0.9512 |
| 8 | 0.9411 | 0.9512 | 0.9512 | 0.9537 | 0.9512 | 0.9512 |
| 9 | 0.9347 | 0.9460 | 0.9409 | 0.9400 | 0.9460 | 0.9409 |
| 10 | 0.9200 | 0.9383 | 0.9409 | 0.9274 | 0.9383 | 0.9409 |

Naïve Bayes:

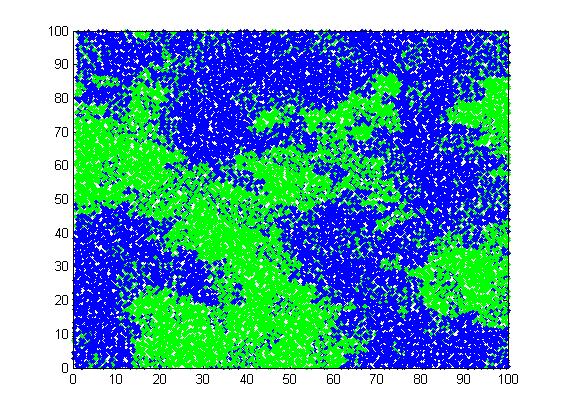
|  |  |  |
| --- | --- | --- |
| Training | Validation | Test |
| 0.8704 | 0.8380 | 0.8380 |

Logistic Regression:

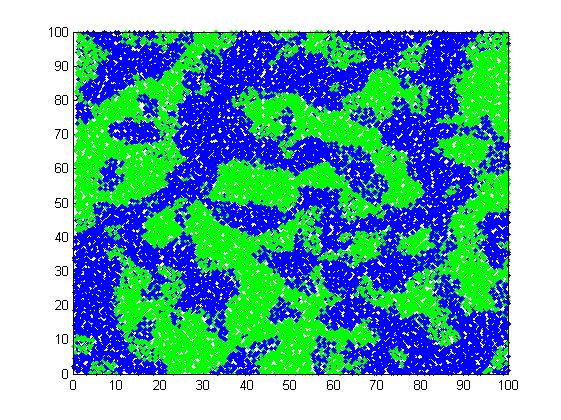
|  |  |  |
| --- | --- | --- |
| Training | Validation | Test |
| 0.8274 | 0.8149 | 0.8535 |

5.5 Decision Boundary

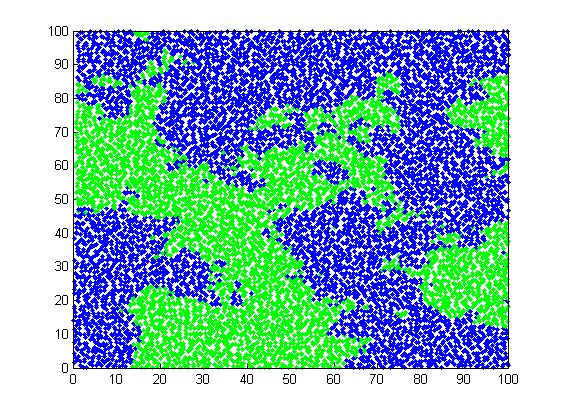
K=1:



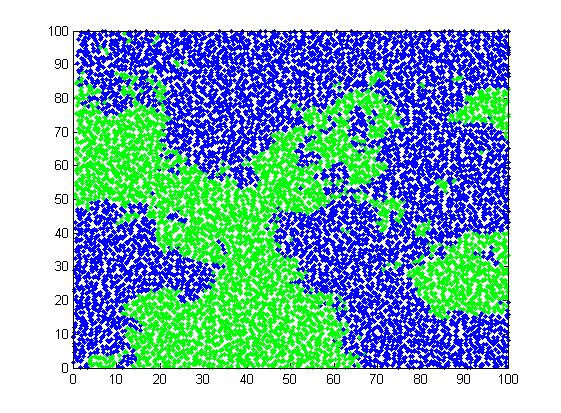
K=5:



K=15:



K=20:



According to the images, we can find that the decision boundary becomes smoother as k increases.