Introduction to artificial intelligence: Assignment #1

Num: 11521050 Name: Jiangrong Shen Date: 10/6/15

1. Machine Learning problems
2. 1) B & F

2) C

3) A & D

4) B & G

5) A & E

6) A & D

7) B & F

8) A & E

9) B & F

b) False.

On one hand, When the number of train data is too large may contribute overfitting, for maybe we would learn a few features not necessary.

On the other hand, we’d better use cross-validation to avoid overfitting.

1. Bayes Decision Rule
2. (i) P(B1=1) = 1/3

(ii) P(B2 =0| B1=1)=1

(iii) P(B2 =0) = P(B2 =0| B1=1)\*P(B2 =0| B1=0) = 1/3\*1 + 2/3\*(1/2) = 2/3

P(B1=1| B2 =0) = P(B2 =0| B1=1)/ P(B2 =0) = (1/3)/(2/3) = 1/2

(iv) I will not change my choice, for the probability is equal if I change my choice.

1. (i) The distribution of P(x|wi) is



The test error of maximum likelihood decision ruleis :

wrongnum : 64

test error rate: 21.33%

(ii) The distribution of P(wi | x) is



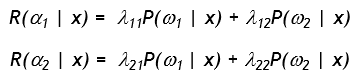
The test error of optimal bayes decision rule is:

Wrongnum: 47

test error rate: 0.1567

(iii) Get the minimal risk using optimal bayes decision rule and risk weights:

Refer to the Equs of :



We can get the conditional risk is:

Columns 1 through 5

0.0008 0.0025 0.0025 0.0075 0.0108

0 0 0 0 0

Columns 6 through 10

0.0375 0.0475 0.0592 0.0817 0.0817

0 0 0 0 0.0017

Columns 11 through 15

0.0825 0.0867 0.0642 0.0450 0.0250

0.0133 0.1033 0.2217 0.2133 0.1017

Columns 16 through 19

0.0142 0.0117 0.0033 0.0025

0.0100 0.0017 0 0

Wrongnum: 64

test error rate: 0.2133

1. Gaussian Discriminant analysis and MLE
2. Refer to the book 《pattern classification》 Page 23

Eqs.59:

Eqs.60:

Eqs.58:

setting to obtain the decision boundary:

1. …
2. Some kinds of decision boundary are as follows:



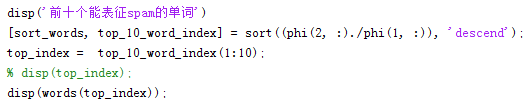


1. The Maximum likelihood estimate is:

That is,

Then we can get:

1. Text classification with Naïve Bayes
2. In the code file rum.m, attend the code :



then will get the list of the top 10 words:

'nbsp'

'viagra'

'pills'

'cialis'

'voip'

'php'

'meds'

'computron'

'sex'

'ooking'

1. accuracy: 84.52%
2. False.

Because although the accuracy is high, but it may not be general model, that is, maybe this model with high accuracy will not have a good performance on another dataset.

For example, when the ratio of spam and ham email is 1:99, then you will obtain a high accuracy 99%, however, we can’t just say that is a good model, for when many other cases will not have such high accuracy, it relates to the distribution of the test dataset.

1. precision: 99.79%

recall: 43.15%

why my recall is so low?

I think the Laplace smooth have a great influence, for it make some words that represent the spam email occurrence one time in the beginning.

1. I think recall is more important in this problem.

Because Compare the situation that we classify a ham email to spam email class, we prefer classifying the spam email to a ham email class, in order to avoid the lost of ham email. In most cases, we can tolerate a spam email to guarantee not lose the ham email.

However, if the classifier to identify drugs and bombs are at airport, I think the precision is more important in this case. Because we can’t to lose any drugs or bombs at airport for our safety, and we can permit there are some cases that regard other safe things to drugs and bombs to some extent.