Machine Learning:2112450

Zhejiang University
September 16 2020

Professor Deng Cai
Homework 1

HOMEWORK 1

Collaberators

Name: 陈志博

Student ID：3190100923

Problem 1-1. Machine Learning Problems

a) Choose proper word(s) from

Answer:

1. B.unsupervised learning   F.clustering
2. C.not learning
3. A.supervised learning   G. classification
4. B.unsupervised learning.   G.dimensionality reduction
5. C.not leaning
6. A.supervised leaning   D. classification
7. B. Unsupervised learning.
8. A.supervised leaning.   E.regression
9. C. Not learning

b) True or False: “To fully utilizing available data resource, we should use all the data we have to train our learning model and choose the parameters that maxi-mize performance on the whole dataset.” Justify your answer.

Answer: False

Because it only reduces the error rate of training samples. Overfitting may lead to high testing error.

Problem 1-2. Bayes Decision Rule

a) Suppose you are given a chance to win bonus grade points:

Answer:

1. P(B1=1)= $\frac{1}{3}$
2. P(B2=0 | B1=1)=1
3. P(B1=1 | B2=0)=$\frac{P(B2=0|B1=1)P(B1=1)}{P(B2=0)}=\frac{1}{3}$
4. P(B1=1 | B2=0)<P(B1=0 | B2=0)   So should change

b) Now let us use bayes decision theorem to make a two-class classifier. Please refer the codes in the bayes decision rule folder and main skeleton code is run.ipynb. There are two classes stored in data.mat. Each class has both training samples and testing samples of 1-dimensional feature x.

Answer:

1. The number of misclassified test samples is 64

   The test error is 0.213

   The distribution of P(x | ω i) is Figure 1:

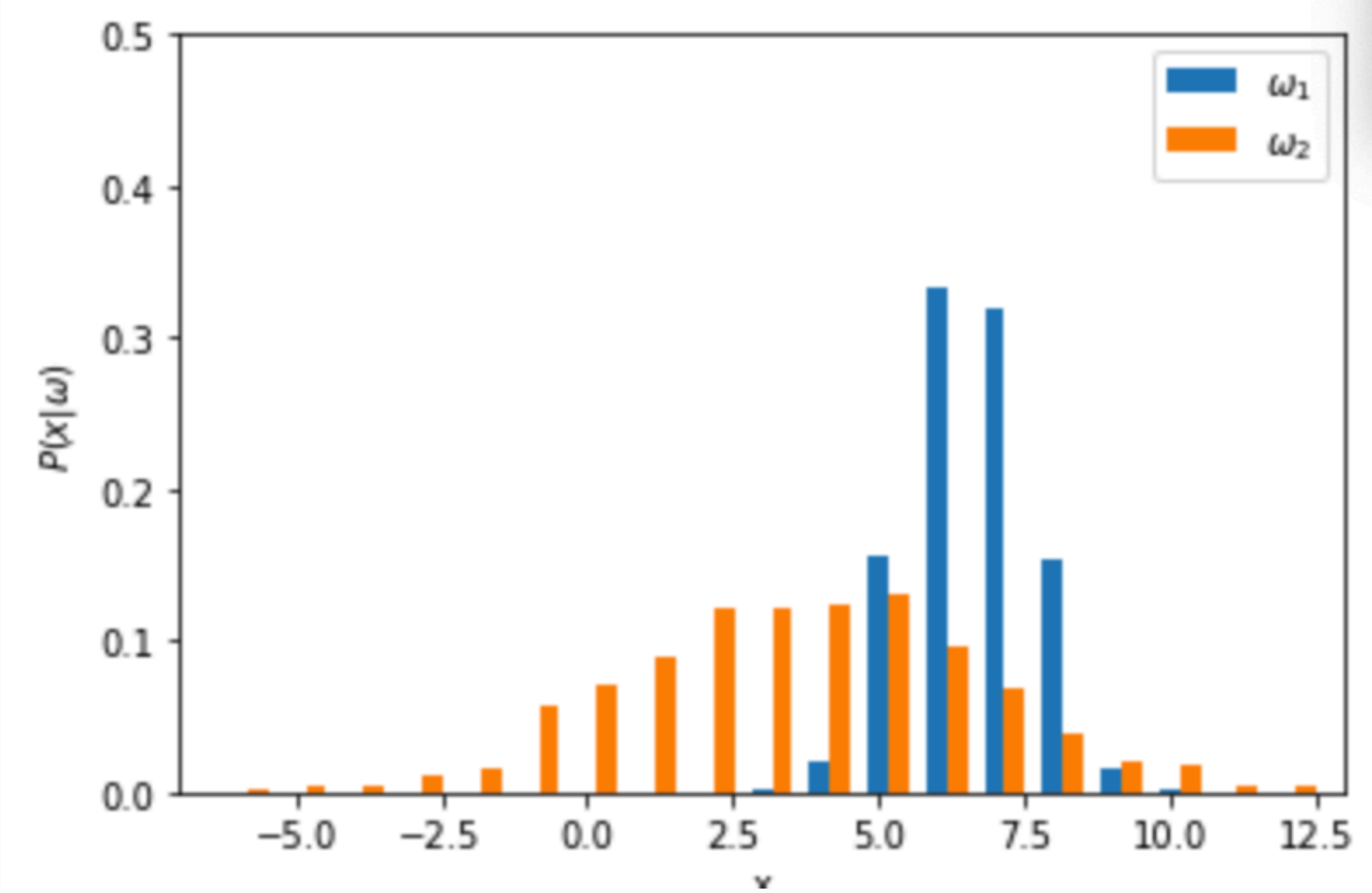


Figure 1 : Machine Learning

2. The number of misclassified test samples is 47

   The test error is 0.157

   The distribution of P(x | ω i) is Figure 2:

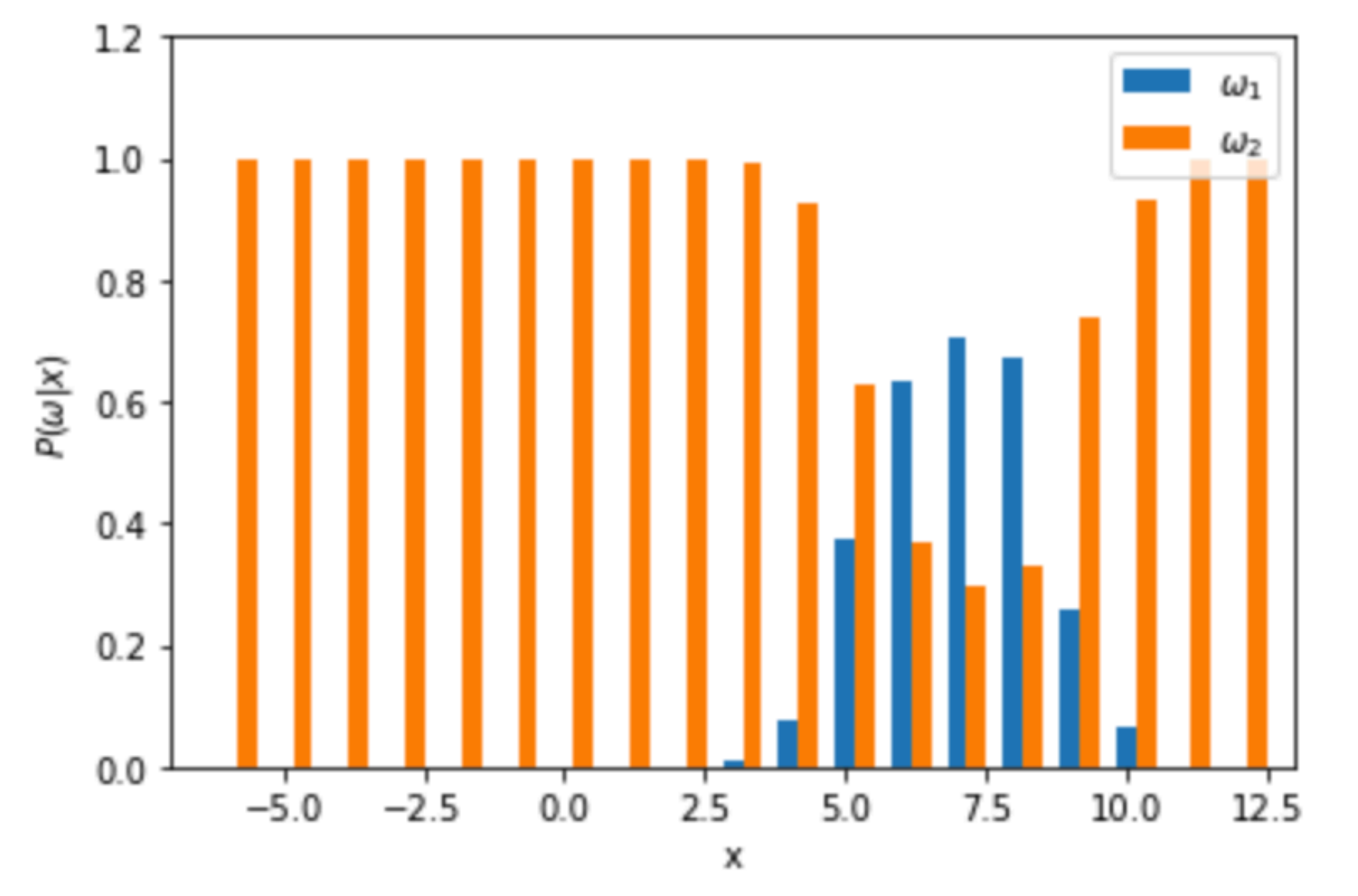


Figure 2 : Machine Learning 2

3. The minimal total risk we can get is 0.2475

Problem 1-3. Gaussian Discriminant Analysis and MLE

Given a dataset {(x (i), y (i)) | x ∈ R 2, y ∈ {0, 1}, i = 1, ..., m} consisting of m samples. We assume these samples are independently generated by one of two Gaussian distributions:

(a) What is the decision boundary?

Answer: $g_i(x)=\frac{(x-\mu_i)^T(x-\mu_i)}{2}+\ln(P(\omega_i))=\mu_i^T-\frac{\mu_i^T\mu}{2}+\ln(P(\omega))$

$g_0(x)-g_1(x)=x_1+x_2+\frac{1}{2}$

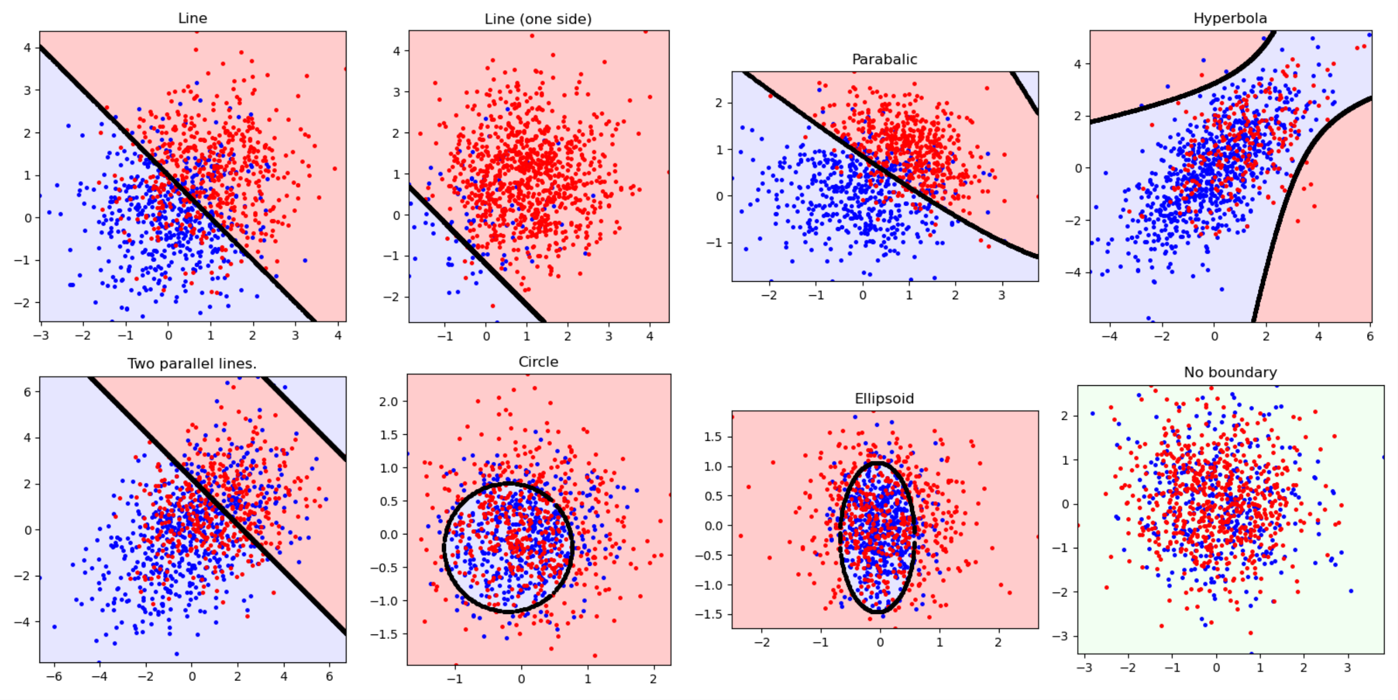
(b) An extension of the above model is to classify K classes by fitting a Gaussian distribution for each class

Answer:

```
import numpy as np
import math

def gaussian_pos_prob(X, Mu, Sigma, Phi):
    '''
    Inputs:
        'X'     - M-by-N numpy array, N data points of dimension M.
        'Mu'    - M-by-K numpy array, mean of K Gaussian distributions.
        'Sigma' - M-by-M-by-K  numpy array (yes, a 3D matrix), variance matrix
of
                  K Gaussian distributions.
        'Phi'   - 1-by-K  numpy array, prior of K Gaussian distributions.
    Outputs:
        'p'     - N-by-K  numpy array, posterior probability of N data points
                with in K Gaussian distribsubplots_adjustutions.
    '''
    N = X.shape[1]
    K = Phi.shape[0]
    p = np.zeros((N, K))

    for i in range(K):
        x=(X.T-Mu[:,i].T).T
        si=Sigma[:,:,i]
        s=np.linalg.inv(si)
        d=np.linalg.det(si)
        x1=np.dot(s,x)
        x2=np.sum(x*x1,axis=0)
        l=1/(2*math.pi*d**0.5)*np.exp(-0.5*x2)
        l1=(l*Phi[i]).reshape(N)
        p[:,i]=l1
    px=np.sum(p,axis=1)
    p=(p.T/px.T).T

    return p
```

(c) Now let us do some field work – playing with the above 2-class Gaussian dis-criminant model.

Answer:



(d) What is the maximum likelihood estimation of ϕ, μ_0 and μ_1

$$\mu_0 = \frac{1}{n} \sum_{y_k=0} x_k$$

$$\mu_1 = \frac{1}{n} \sum_{y_k=1} x_k$$

$$\phi = \frac{\sum_{y_k=1} 1}{\sum_{y_k=0} 1}$$

Problem 1-4. Text classification with Naive Bayes

(a) list the top 10 words

> Answer: nbsp viagra pills cialis voip php meds computron sex ooking

(b) What is the accuracy of your spam filter on the testing set?

> Answer: 0.9786476868327402

(c) True or False: a model with 99% accuracy is always a good model. Why?

> Answer: False
>
> When the ratio of spam and ham email is 1:99, and all spam e-mail are wrongly predicted. For ham e-mail the accuracy is 99% , but it's obviously not a good model

(d) Compute the precision and recall of your learnt model.

> Answer:
>
> ```
> precision= 0.9725906277630415
> recall= 0.9786476868327402
> ```

(e) For a spam filter, which one do you think is more important, precision or recall? What about a classifier to identify drugs and bombs at airport? Justify your answer.

> Answer:
>
> ```
> I deem recall to be more important.
> Because in a prediction process, we don't know exactly its label,  what
> matters more is to tell the accuracy of our prediction
> ```