Machine Learning: 21121450

Zhejiang University Sept 16, 2020

Professor Deng Cai Homework 1

Homework 1

Collaborators:

Name: Youchao Zhang

Student ID: 3170100125

Problem1-1.Machine Learning Problems

(a) Choose proper word(s) from

Answer

task	choice
1	B F
2	C
3	A D
4	G
5	A D
6	A D
7	C
8	A E
9	B F

(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 maximize performance on the whole dataset.” Justify your answer.

Answer :False

If we use all the data to train our model, the model may overfit. And the model is not robust enough. So the performance in the new data may be very poor. We should split the data into train set and test set, use test set to test the model.

For example if our model is a linear regression, if we use all the data to train our model, finally we may get a polynomial function which fit our data perfectly and the error is zero. But when we use the model to predict our new data the error will be very significant.

Problem1-2.Bayes Decision Rule

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

Answer

(i)

$$P(B_1 = 1) = \frac{1}{3}$$

(ii)

$$P(B_2 = 0|B_1 = 1) = \frac{P(B_1 = 1, B_2 = 0)}{P(B_1 = 1)} = \frac{\frac{1}{3}}{\frac{1}{3}} = 1$$

(iii)

$$P(B_1 = 1|B_2 = 0) = \frac{P(B_1 = 1, B_2 = 0)}{P(B_2 = 0)} = \frac{\frac{1}{3}}{1} = \frac{1}{3}$$

(iiii)

B_1	B_2	B_3
0	0	1
1	0	0
0	0	1

$$P(B_1 = 1|B_2 = 0) = \frac{P(B_2 = 0|B_1 = 1)P(B_1 = 1)}{P(B_2 = 0)} = \frac{1 * \frac{1}{3}}{1} = \frac{1}{3}$$

$$P(B_3 = 1|B_2 = 0) = 1 - P(B_1 = 1|B_2 = 0) = \frac{2}{3}$$

Since $P(B_3 = 1|B_2 = 0) > P(B_1 = 1|B_2 = 0)$,I should change my choice

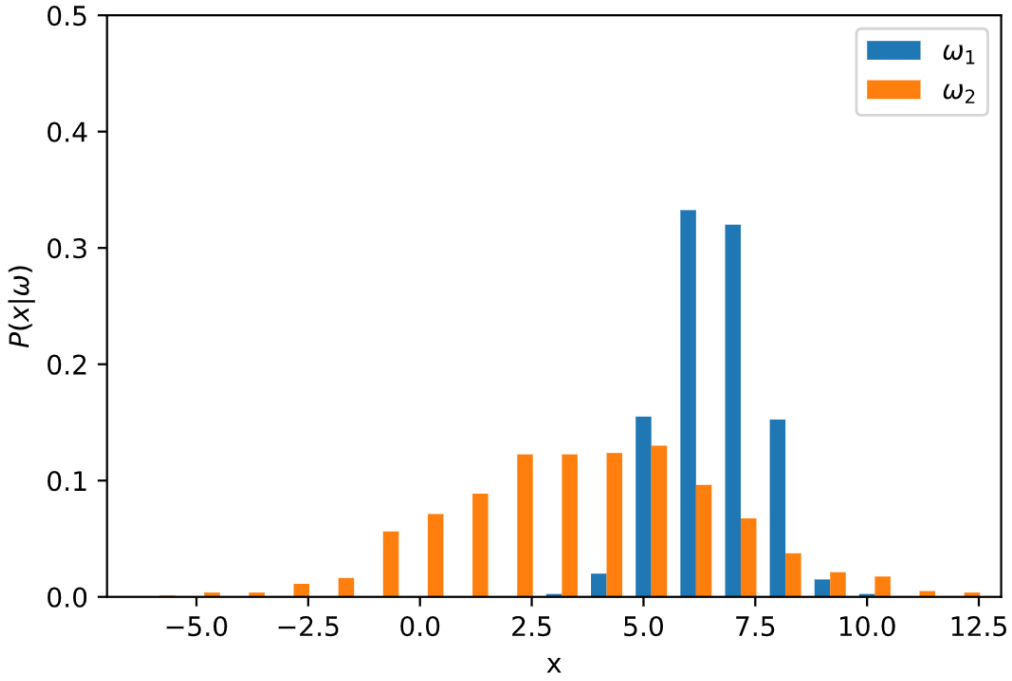
(b) Now let us use bayes decision theorem to make a two-class classifier…..

Answer :

(i)

The number of misclassified test samples is 64

The test error is 21.3%

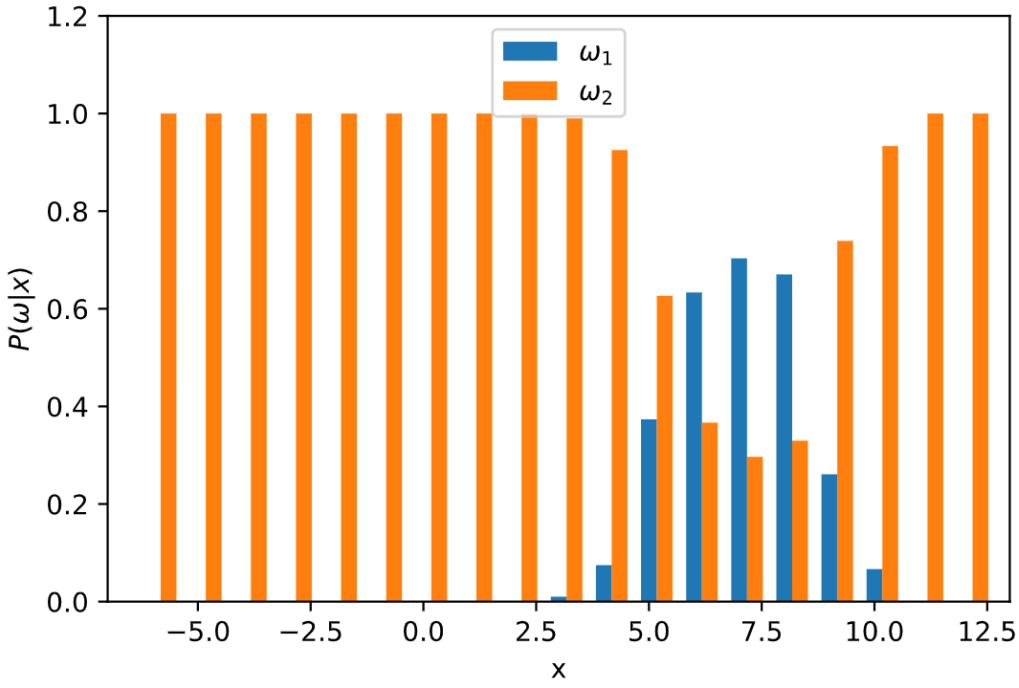


The distribution of $P(x|\omega_i)$

(ii)

The number of misclassified test samples is 47

The test error is 15.7%



The distribution of $P(\omega_i|x)$

(iii) The minimal total risk = 0.2529

Problem1-3. Gaussian Discriminant Analysis and MLE

(a) What is the decision boundary?

Answer

$$p(X|y=1) = N(\mu_1, \Sigma_1) = \frac{1}{2\pi} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)} = \frac{1}{2\pi} e^{-\frac{1}{2}[(x_1-1)^2+(x_2-1)^2]}$$

$$p(y=1) = \frac{1}{2}$$

$$p(X|y=0) = N(\mu_0, \Sigma_1) = \frac{1}{2\pi} e^{-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0)} = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)}$$

$$p(y=0) = \frac{1}{2}$$

$$\therefore p(y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\Sigma_1) = \frac{p(X|y=1)p(y=1)}{p(x)}$$

$$= \frac{\frac{1}{4\pi} e^{-\frac{1}{2}[(x_1-1)^2+(x_2-1)^2]}}{\frac{1}{4\pi} e^{-\frac{1}{2}[(x_1-1)^2+(x_2-1)^2]} + \frac{1}{4\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)}}$$

$$= \frac{e^{-\frac{1}{2}[(x_1-1)^2+(x_2-1)^2]}}{e^{-\frac{1}{2}[(x_1-1)^2+(x_2-1)^2]} + e^{-\frac{1}{2}(x_1^2+x_2^2)}} = \frac{1}{1+e^{1-x_1-x_2}}$$

The decision boundary is: $0 = x_1 + x_2 - 1$

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

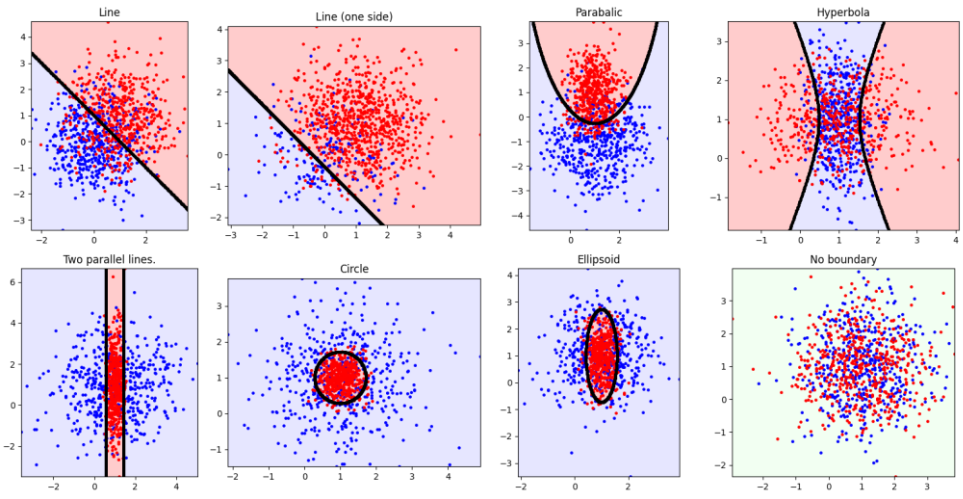
Answer

see the gaussian_pos_prob.py

(c) Now let us do some field work – playing with the above 2-class Gaussian discriminant

model.

Answer



(i) A linear line.
(ii) A linear line, while both means are on the same side of the line.
(iii) A parabolic curve.
(iv) A hyperbola curve.
(v) Two parallel lines.
(vi) A circle.
(vii) An ellipsoid.
(viii) No boundary, i.e. assigning all samples to only one label.

(d) What is the maximum likelihood estimation ofϕ, μ_0, μ_1
Answer

$$P(D|\phi) = \Pi_j P(x_j|\phi) = \Pi_j\{\phi^{y_j}(1-\phi)^{1-y_j}\} = \phi^{\Sigma y_j}(1-\phi)^{\Sigma(1-y_j)}$$

$$\ln\big(P(D|\phi)\big) = ln\phi * \Sigma y_j + ln(1-\phi) * \Sigma(1-y_j)$$

$$\frac{\partial \ln\big(P(D|\phi)\big)}{\partial \phi} = \frac{1}{\phi} * \Sigma y_j + \frac{1}{(1-\phi)} * \Sigma\big(1-y_j\big) = 0$$

$$\therefore \phi = \frac{\Sigma y_j}{\Sigma y_j + \Sigma\big(1-y_j\big)}$$

$$P(D|\mu) = \Pi_j P(x_j|\mu) = \Pi_j \left(\frac{1}{2\pi\sigma} e^{-\frac{(x_j-\mu)^2}{2}}\right) = \left(\frac{1}{2\pi\sigma}\right)^n e^{\Sigma_j[-\frac{(x_j-\mu)^2}{2}]}$$

$$\ln\big(P(D|\mu)\big) = \Sigma_j \frac{(x_j-\mu)^2}{2} * nln(2\pi\sigma)$$

$$\frac{\partial \ln\big(P(D|\mu)\big)}{\partial \mu} = \Sigma_j^n\big(x_j-\mu\big) = 0$$

$$\hat{\mu} = \frac{1}{\mathrm{n}}\Sigma_j^n x_j$$

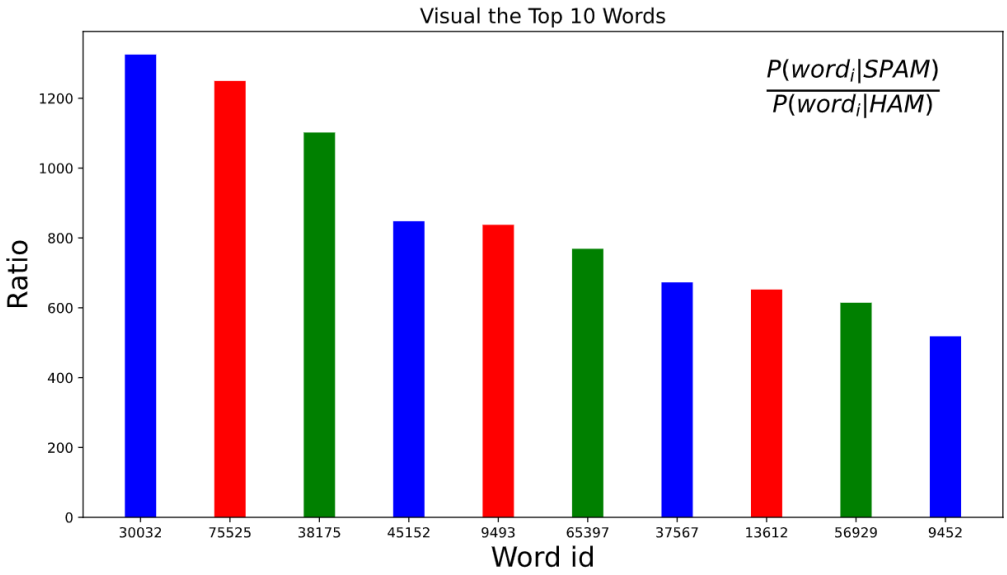
$$\therefore\ \mu_0 = \hat{\mu} = \frac{1}{\mathrm{n}}\Sigma_j^n x_j^{(0)}$$

$$\therefore\ \mu_1 = \hat{\mu} = \frac{1}{\mathrm{n}}\Sigma_j^n x_j^{(1)}$$

Problem1-4. Text Classification with Naive Bayes

(a) List the top 10 words.

Answer



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

Answer

Error_of_spamToham = 31

Error_of_hamTospam = 28

The total accuracy = 99.961879%

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

Answer

False.

Because if the ratio of spam and ham email is 1:99, we can easily separate the spam from ham and we can also get a model with 99% accuracy. But since the number of spam is far less than the ham, the model can't get enough features from spam. If we have another spam which is rather different from the train spam, the model may can not perform very well to identify it.

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

Answer

Table: confusion matrix

	Spam(label)	Ham(label)
Spam(predict)	77355	28
Ham(predict)	31	77358

$$Precision = \frac{TP}{TP + FP} = 99.963816\%$$

$$Recall = \frac{TP}{TP + FN} = 99.959941\%$$

(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

For the spam filter issue, I think Precision is more important, because higher precision means lower FP , which means less ham is wrongly identified to spam to be put into Spam box, so we won't miss the normal message about work or family.

For a classier to identify drugs and bombs at airport, I think Recall is far more important, because higher recall means lower FN, which means we can identify more drugs or bombs ,after that staff will do a double check. If the classifier makes a mistake, it not a big deal. But if FN is high, the airport will suffer danger.