DM-HW1

Zijun Tian

May 2019

1 Machine Learning Problems

(a)
(1) BF
(2) C
(3) C
(4) BG
(5) AE
(6) AD
(7) BF
(8) AE
(9) BF

(b)
False, the dataset should be divided into training set, cross validation set and test set. Overfitting, which means the performance on new data can be much worse than the dataset, may happen when training set accounts for the large part of the whole dataset or the parameters are chosen for training set instead of test set used as new data.

2 Bayes Decision Rule

(a)
(i)

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

(ii)

$$P(B_2 = 0|B_1 = 1) = 1$$

(iii)

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

(iv)

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

According to the previous calculation, I should change my choice.

(i)

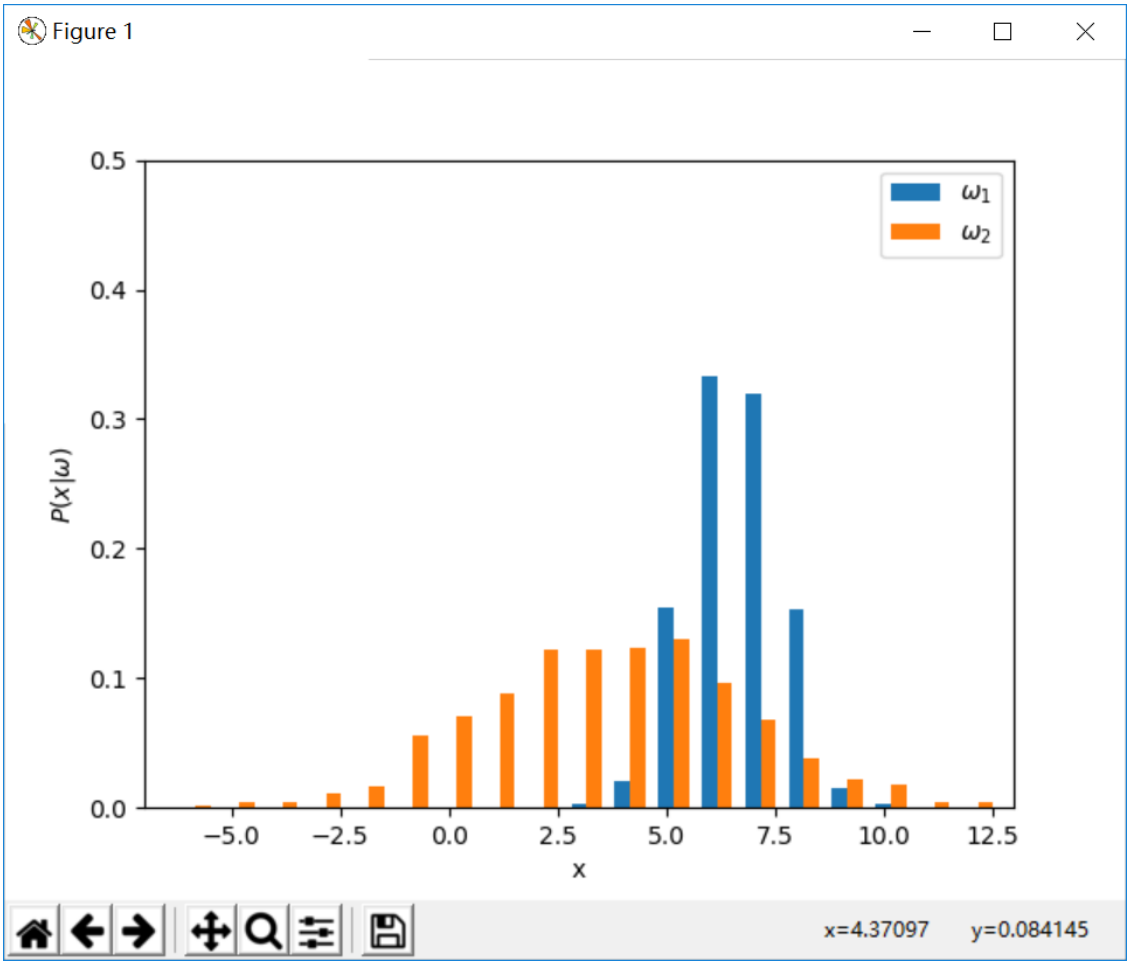

Figure 1: $P(x \mid \omega_i)$

test error = 64

(ii)

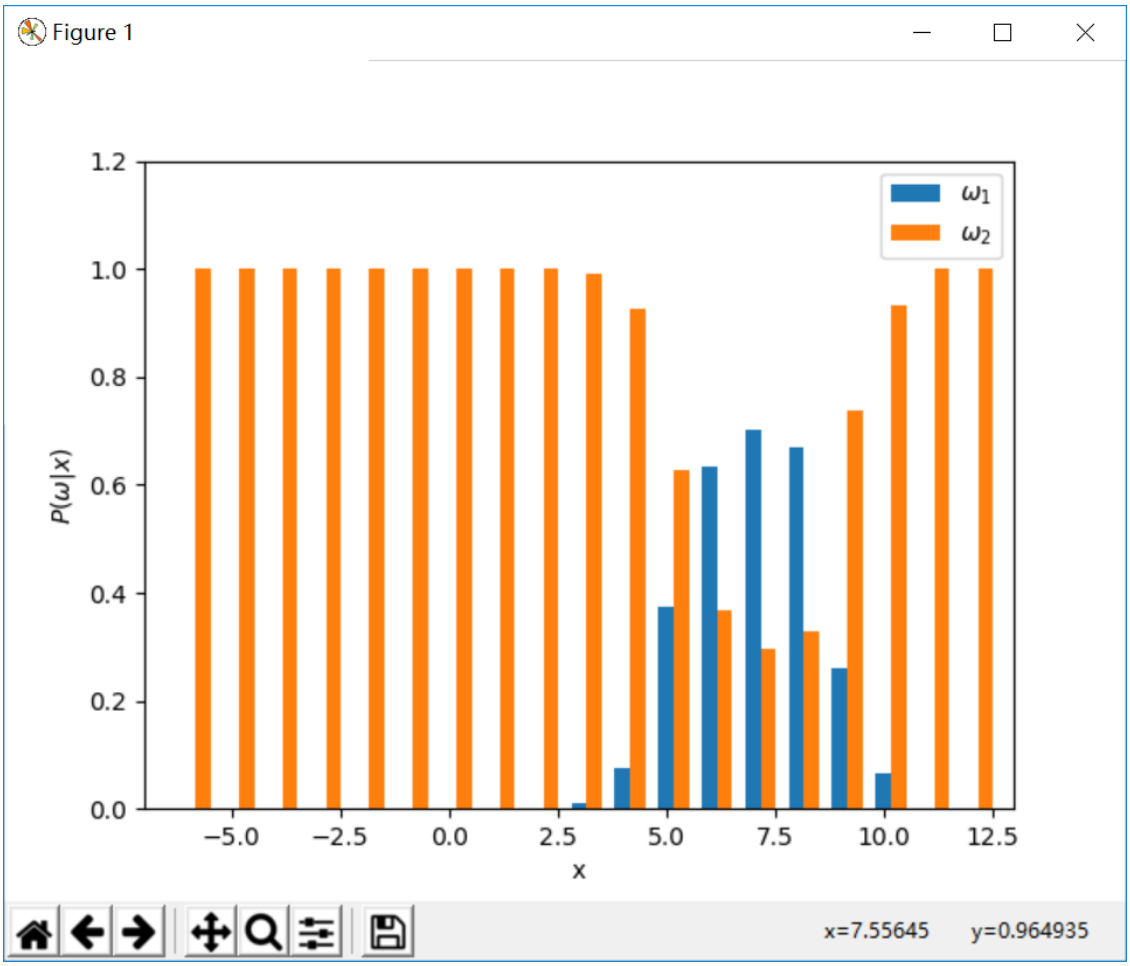

Figure 2: $P(\omega_i \mid x)$

test error = 47

(iii)

R = 0.243

3 Gaussian Discriminant Analysis and MLE

(a)

$$P(y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1) = \frac{P(x|y=1)*P(y=1)}{P(x)}$$
$$P(x) = P(x|y=0)*P(y=0) + P(x|y=1)*P(y=1)$$
$$\phi = \frac{1}{2}, \Sigma_0 = \Sigma_1 = 1, \mu_0 = (0,0)^T, \mu_1 = (1,1)^T$$
$$P(x|y=0) = \frac{1}{2\pi}e^{-\frac{1}{2}(x_1^2+x_2^2)}$$
$$P(x|y=1) = \frac{1}{2\pi}e^{-\frac{1}{2}(x_1-1)^2-\frac{1}{2}(x_2-1)^2}$$
$$P(y=0) = P(y=1) = \frac{1}{2}$$

With equations above, we can conclude that

$P(y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1) = \frac{\frac{1}{2\pi}e^{-\frac{1}{2}(x_1^2+x_2^2)}}{\frac{1}{2\pi}e^{-\frac{1}{2}(x_1-1)^2-\frac{1}{2}(x_2-1)^2}+\frac{1}{2\pi}e^{-\frac{1}{2}(x_1-1)^2-\frac{1}{2}(x_2-1)^2}} = \frac{e^{-\frac{1}{2}(x_1^2+x_2^2)}*e^{x_1+x_2-1}}{e^{-\frac{1}{2}(x_1^2+x_2^2)}+e^{-\frac{1}{2}(x_1^2+x_2^2)}*e^{x_1+x_2-1}} = \frac{e^{x_1+x_2-1}}{e^{x_1+x_2-1}+1}$

$P(y=0|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1) = 1 - P(y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1) = \frac{1}{e^{x_1+x_2-1}+1}$

To find the boundary, let $P(y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1) = P(y=0|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1)$

we can find that at this time $x_1+x_2=1$

when $x_1+x_2>1, P(y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1) > (y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1)$

when $x_1+x_2<1, P(y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1) < (y=1|x;\phi,\mu_0,\mu_1,\Sigma_0,\sigma_1)$

(c)

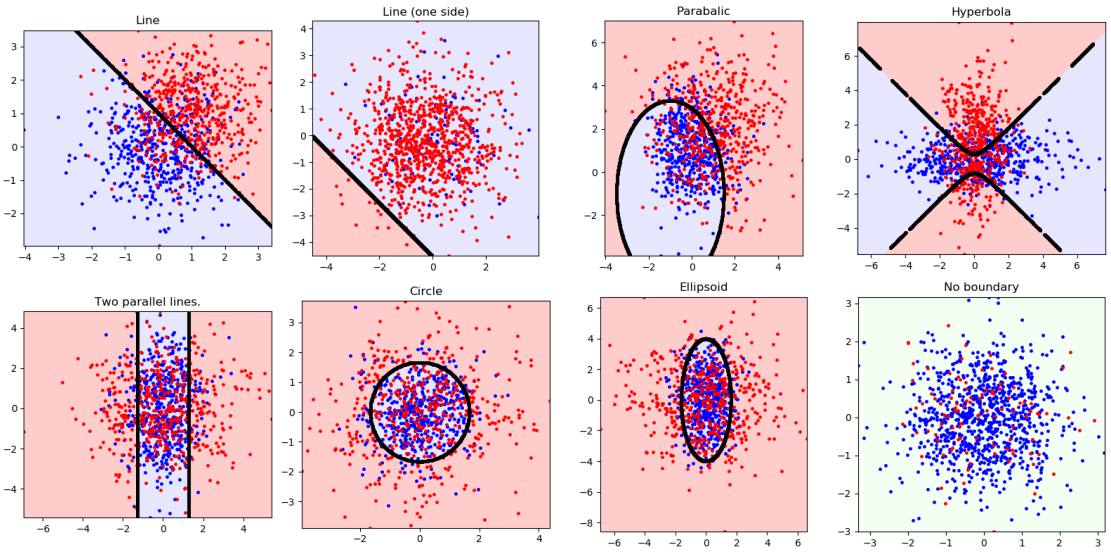

Figure 3: $P(\omega_i \mid x)$

(d)

Using MLE

$$\begin{aligned} l(x|y=0) &= \log L(x|y=0) \\ &= \log \prod_{i=1}^{N} p(x|y=0) \\ &= \Sigma_{i=0}^{N}\{\log(-2\pi^d\Sigma) - \frac{(x-\mu_0)^T * \Sigma_0^{-1} * (x-\mu_0)}{2}\} \end{aligned}$$

we can find that $\mu_0, \mu_1 and \Sigma_0, \Sigma_1 is the same, therefore, we use it generally for k classes$

$$\begin{gathered} \frac{\partial l}{\partial \mu} = \Sigma_{i=1}^{N}(x-\mu) = 0 \\ \frac{\partial l}{\partial \Sigma} = -\frac{N}{2}\Sigma^{-1} - \frac{1}{2}\Sigma^{-1}(\Sigma_{i=1}^{N}(x-\mu)(x-\mu)^T)\Sigma^{-1} = 0 \\ \mu = \frac{1}{K}\Sigma_{i=1}^{K}x_i \\ \Sigma = \frac{1}{K}\Sigma_{i=1}^{K}(x_i-\mu)(x_i-\mu)^T \\ \phi = \frac{|x_k|}{\Sigma_{i=1}^{K}} \end{gathered}$$

4 Text Classification with Naive Bayes

(a)

30032 75525 38175 45152 9493 65397 37567 13612 56929 9452

nbsp viagra pills cialis voip php meds computron sex ooking

(b)

acc = 98.57%

(c)

P(S)=0.01 denotes p of spam, P(H)=0.99 denotes p of ham. P(PS) denotes p of predicted as spam, P(PH) denotes p of predicted as ham.

$$P(S|PS) = \frac{P(PS|S) * P(S)}{P(PS)} P(PS) = P(PS|S) * P(S) + P(PS|H) * P(H) P(S|PS) = \frac{0.99 * 0.01}{0.99 * 0.01 + 0.99 * 0.01} = 0.5$$

We can conclude that the accuracy of prediction is only o.5, which is not a good model.

(d) precision = 97.5% recall = 97.24%

(e)

For email filter, precision is more important because the low recall brings some spams, but the low precision may block some important emails which can be very harmful. As for me, human beings' life is the important thing in the world. The low precision only costs more but low recall takes others' life. In this way, we should do everything to improve the recall.