Assignment #1

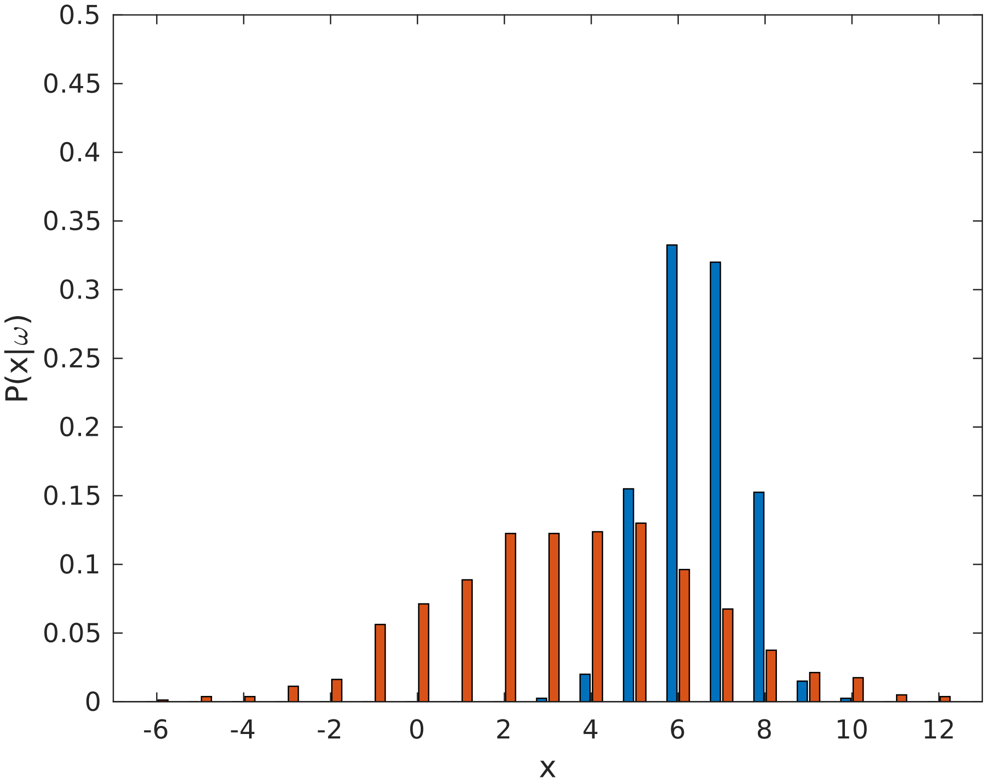
1. Machine Learning Problems

- (a)
    1. BF
    2. C
    3. C
    4. BG
    5. AE
    6. AD
    7. BF
    8. AE
    9. BF
- (b) No. Data set should be divided into training set and test data. And we should not maximize the performance of the training set, but that of the test set.

2. Bayes Decision Rule

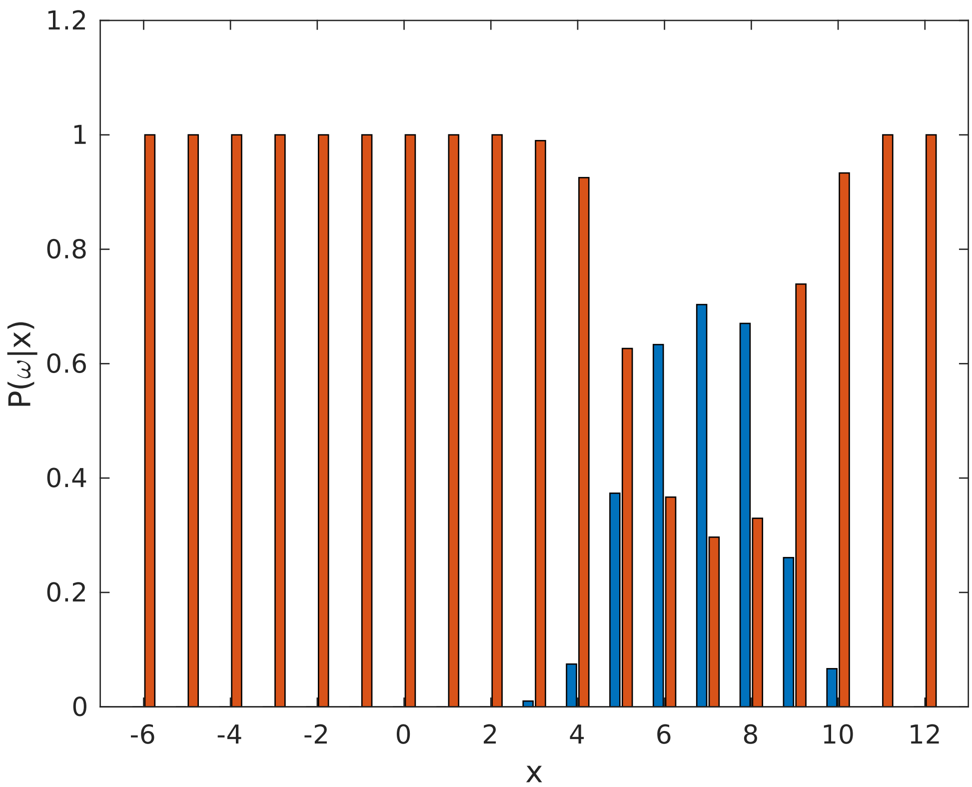
- (a)
    - (i)
    $$P(B_1 = 1) = \frac{1}{3}$$
    - (ii)
    $$P(B_2 = 0 \mid B_1 = 1) = 1$$
    - (iii)
    $$P(B_1 = 1 \mid B_2 = 0) = \frac{P(B_2 = 0 \mid B_1 = 1) + P(B_2 = 1 \mid B_1 = 1)}{P(B_1 = 1)} = \frac{1}{2}$$
    - (iv)
    $$P(B_3 = 1 \mid B_2 = 0) = P(B_1 = 1 \mid B_2 = 0) = \frac{1}{2}$$
    So, after knowing B~2~ contains nothing, the probability of B~1~ and B~2~ containing the bonus is the same. So changing my choice or not are both okay.
- (b)

- (i) The distribution of $P(x \mid \omega_i)$ is shown as follow:



  And the test error is 64.

- (ii) The distribution of $P(\omega_i \mid x)$ is shown as follow:



  And the test error is 47.

- (iii) The minimal total risk ($R = \sum_x \min_i R(\alpha_i \mid x)$) is 2.444354.

3. Gaussian Discriminant Analysis and MLE

- (a) When

$$\Sigma_0 = \Sigma_1 = \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I, \phi = \frac{1}{2}, \mu_0 = (0,0), \mu_1 = (1,1)^T$$

We have:

$$P(\mathtt{x} \mid y=0) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)}$$

$$P(\mathtt{x} \mid y=1) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1-1)^2 - \frac{1}{2}(x_2-1)^2} = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} e^{x_1+x_2-1}$$

So:

$$\begin{aligned} P(y=1 \mid \mathtt{x}; \phi, \mu_0, \mu_1, \Sigma_0, \Sigma_1) &= \frac{P(\mathtt{x} \mid y=1)P(y=1)}{P(\mathtt{x} \mid y=0)P(y=0) + P(\mathtt{x} \mid y=1)P(y=1)} \\ &= \frac{\frac{1}{4\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} e^{x_1+x_2-1}}{\frac{1}{4\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} + \frac{1}{4\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} e^{x_1+x_2-1}} \\ &= \frac{e^{x_1+x_2-1}}{1+e^{x_1+x_2-1}} \\ &= \frac{1}{1+e^{1-x_1-x_2}} \end{aligned}$$

And at the same time, we get:

$$\begin{aligned} P(y=0 \mid \mathtt{x}; \phi, \mu_0, \mu_1, \Sigma_0, \Sigma_1) &= 1 - P(y=1 \mid \mathtt{x}; \phi, \mu_0, \mu_1, \Sigma_0, \Sigma_1) \\ &= \frac{1}{1+e^{x_1+x_2-1}} \end{aligned}$$

Let $P(y=0) = P(y=1)$, we can get the solution of the discriminant plane:

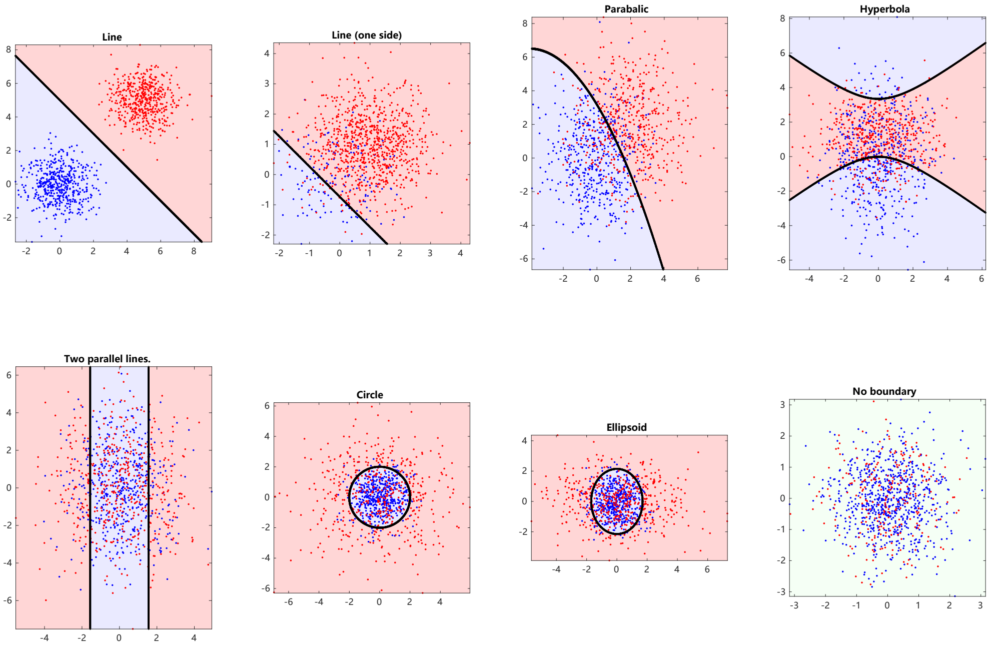
$$x_1 + x_2 = 1$$

When $x_1 + x_2 < 1$, $P(y=0) > P(y=1)$, and when $x_1 + x_2 > 1$, $P(y=0) < P(y=1)$. So the decision boundary is:

$$\int_{x_1+x_2<1} P(\mathtt{x} \mid y=0)P(y=0)\, d\mathtt{x} + \int_{x_1+x_2>1} P(\mathtt{x} \mid y=1)P(y=1)\, d\mathtt{x} = 0.76205$$

- (b) See the code.

- (c) The result plots are shown as follow:



- (d) I directly start from K-class gaussian model. Firstly divided the data set into K sub set.

  $$
  \mathtt{x}_k=\left\{ \mathtt{x}^{(i)} \mid y^{(i)}=k, i=1,\dots ,m\right\}\\
  N_k=\left|\mathtt{x}_k\right|
  $$

$$\mathtt{x}_k=\left\{ \mathtt{x}^{(i)} \mid y^{(i)}=k, i=1,\dots ,m\right\}$$
$$N_k=\left|\mathtt{x}_k\right|$$

Applying MLE, we have:

$$\max P\left(\mathtt{x}_k \mid \mu_k, \mathbf{\Sigma}_k\right)$$

, where

$$P\left(\mathtt{x}_k \mid \mu_k, \mathbf{\Sigma}_k\right)=\prod_{\mathtt{x}\in\mathtt{x}_k}\frac{1}{(2\pi)^{D/2}\left|\mathbf{\Sigma}_k\right|}\exp\left\{-\frac{1}{2}(\mathtt{x}-\mu_k)^T\mathbf{\Sigma}_k^{-1}\left(\mathtt{x}-\mu_k\right)\right\}$$
$$\ln P\left(\mathtt{x}_k \mid \mu_k, \mathbf{\Sigma}_k\right)=-\frac{N_kD}{2}\ln 2\pi-\frac{N_k}{2}\ln\left|\mathbf{\Sigma}_k\right|-\frac{1}{2}\sum_{\mathtt{x}\in\mathtt{x}_k}(\mathtt{x}-\mu_k)^T\mathbf{\Sigma}_k^{-1}\left(\mathtt{x}-\mu_k\right)$$

Let

$$\frac{\partial}{\partial\mu_k}\ln P\left(\mathtt{x}_k \mid \mu_k, \mathbf{\Sigma}_k\right)=\sum_{\mathtt{x}\in\mathtt{x}_k}\mathbf{\Sigma}_k^{-1}\left(\mathtt{x}-\mu_k\right)=\mathbf{0}$$
$$\Rightarrow \qquad \mu_{ML_k}=\frac{1}{\left|\mathtt{x}_k\right|}\sum_{\mathtt{x}\in\mathtt{x}_k}\mathtt{x}$$

The same for $\mathbf{\Sigma}_k$:

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln P\left(\mathbf{x}_k \mid \mu_k, \boldsymbol{\Sigma}_k\right) \\
&= -\frac{N_k}{2} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln \left|\boldsymbol{\Sigma}_k\right| - \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \sum_{\mathbf{x} \in \mathbf{x}_k} \left(\mathbf{x}-\mu_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}-\mu_k\right) \\
&= -\frac{N_k}{2}\left(\boldsymbol{\Sigma}_k^{-1}\right)^T - \frac{1}{2}\left(\boldsymbol{\Sigma}_k^{-1}\right)^T\left(\sum_{\mathbf{x} \in \mathbf{x}_k}\left(\mathbf{x}-\mu_k\right)\left(\mathbf{x}-\mu_k\right)^T\right)\left(\boldsymbol{\Sigma}_k^{-1}\right)^T \\
&= -\frac{N_k}{2} \boldsymbol{\Sigma}_k^{-1} - \frac{1}{2} \boldsymbol{\Sigma}_k^{-1}\left(\sum_{\mathbf{x} \in \mathbf{x}_k}\left(\mathbf{x}-\mu_k\right)\left(\mathbf{x}-\mu_k\right)^T\right) \boldsymbol{\Sigma}_k^{-1} = \mathbf{0} \\
\Rightarrow \qquad & N_k \boldsymbol{\Sigma}_k^{-1} = \boldsymbol{\Sigma}_k^{-1}\left(\sum_{\mathbf{x} \in \mathbf{x}_k}\left(\mathbf{x}-\mu_k\right)\left(\mathbf{x}-\mu_k\right)^T\right) \boldsymbol{\Sigma}_k^{-1} \\
\Rightarrow \qquad & \boldsymbol{\Sigma}_{ML_k} = \frac{1}{\left|\mathbf{x}_k\right|} \sum_{\mathbf{x} \in \mathbf{x}_k}\left(\mathbf{x}-\mu_{ML_k}\right)\left(\mathbf{x}-\mu_{ML_k}\right)^T
\end{aligned}
$$

For unbiased estimation:

$$
\begin{aligned}
\mathbb{E}\left(\mu_{ML_k}\right) &= \mathbb{E}\left(\frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k} \mathbf{x}\right) = \frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k} \mathbb{E}(\mathbf{x}) = \mu_k \\
\mathbb{E}\left(\boldsymbol{\Sigma}_{ML_k}\right) &= \mathbb{E}\left[\frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k}\left(\mathbf{x}-\mu_{ML_k}\right)\left(\mathbf{x}-\mu_{ML_k}\right)^T\right] \\
&= \frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k} \mathbb{E}\left[\left(\mathbf{x}-\mu_{ML_k}\right)\left(\mathbf{x}-\mu_{ML_k}\right)^T\right] \\
&= \frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k} \mathbb{E}\left(\mathbf{x}\mathbf{x}^T - 2\mu_{ML_k}\mathbf{x}^T + \mu_{ML_k}\mu_{ML_k}^T\right) \\
&= \frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k} \mathbb{E}\left(\mathbf{x}\mathbf{x}^T\right) - 2\mathbb{E}\left(\frac{\mu_{ML_k}}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k} \mathbf{x}^T\right) + \mathbb{E}\left(\mu_{ML_k}\mu_{ML_k}^T\right) \\
&= \frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k} \mathbb{E}\left(\mathbf{x}\mathbf{x}^T\right) - 2\mathbb{E}\left(\mu_{ML_k}\mu_{ML_k}^T\right) + \mathbb{E}\left(\mu_{ML_k}\mu_{ML_k}^T\right) \\
&= \frac{1}{N_k} \sum_{\mathbf{x} \in \mathbf{x}_k}\left(\mu_k\mu_k^T + \boldsymbol{\Sigma}_k\right) - 2\left(\mu_k\mu_k^T + \frac{\boldsymbol{\Sigma}_k}{N_k}\right) + \mu_k\mu_k^T + \frac{\boldsymbol{\Sigma}_k}{N_k} \\
&= \frac{N_k - 1}{N_k} \boldsymbol{\Sigma}_k
\end{aligned}
$$

So evidently:

$$
\begin{cases}
\mu_k = \dfrac{1}{|\mathbf{x}_k|} \sum_{\mathbf{x} \in \mathbf{x}_k} \mathbf{x} \\
\mathbf{\Sigma}_k = \dfrac{1}{|\mathbf{x}_k| - 1} \sum_{\mathbf{x} \in \mathbf{x}_k} (\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^T \\
\phi_k = \dfrac{|\mathbf{x}_k|}{\sum_{i=1}^{K} |\mathbf{x}_i|}
\end{cases}
$$

4. Text Classification with Naive Bayes

- (a)

  1. nbsp
  2. viagra
  3. pills
  4. cialis
  5. voip
  6. php
  7. meds
  8. computron
  9. sex
  10. ooking
- (b) 98.57%
- (c) False. Consider the ratio of spam and ham email is 1:99. Then according to Bayes theorem:

$$
\begin{aligned}
p(S) = 0.01, p(H) = 0.99 \\
p(S \mid P_s) = & \frac{p(P_s \mid S) p(S)}{p(P_s \mid S) p(S) + p(P_s \mid H) p(H)} \\
= & \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} = 0.5
\end{aligned}
$$

  , where P_s means predicted as spam, and P_h means predicted as ham, S means its a spam, and H means its a ham. From the result, we know that if our model says an email is spam, the probability of that it's really a spam is only 0.5.
- (d) For my classifier, $tp = 1093, fp = 28, fn = 31, tn = 2983$. So:

$$
\begin{aligned}
p = \frac{tp}{tp + fp} = 97.5\% \\
r = \frac{tp}{tp + fn} = 97.2\%
\end{aligned}
$$

- (e) **Precision** is more important. Because we don't what to mis-predict a ham as spam. And it's acceptable to let go some spams. But for identifying drugs and bombs, the **recall** is more important because we cannot let go any drugs and bombs.