1. Machine Learning Problems

(a)

1. BF
2. C
3. BD
4. BG
5. AE
6. AD
7. BF
8. AE
9. BF

(b)

False. You should have the test dataset and the train dataset. Because the training on the training set will cause overfit, if you do not distinguish between the test set and the training set, the final result will be false high.

2. Bayes Decision Rule

(a)

1. $P(B_1 = 1) = \frac{1}{3}$
2. $P(B_2 = 0|B_1 = 1) = 1$
3. $P(B_1 = 1|B_2 = 0) = \frac{1}{2}$
4. Either choises have the same probility

3. Gaussian Discriminant Analysis and MLE

(a)

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

Discisoin Boudry: $x_1^2 * x_2^2 = (x_1 - 1)^2(x_2 - 1)^2$

(d)

we define n as number of y=0, m as number of y = 1

$\phi = \frac{m}{m+n}$

$\mu_0 = \frac{\sum_{y=0} x_i}{n}$ the average x when y = 0

$\mu_1 = \frac{\sum_{y=1} x_i}{m}$ the average x when y = 1

$$\sum_0 = \frac{\sum_{y=0}(xi - \mu_0)(xi - \mu_0)^T}{n}$$

$$\sum_1 = \frac{\sum_{y=1}(xi - \mu_1)(xi - \mu_1)^T}{m}$$

$p(X|y = k) = N(\mu_k, \sum_k)$

4. Text Classification with Naive Bayes

(a)

52913, 27709, 2571, 23471, 23631, 8058, 22602, 33609, 35796,60120

(b)

0.9857315598548972

(c)

we define a classifier which defines all messages as ham messages. If the number of spam messages in all test sets is only 1%, then the final accuracy rate is 99%. In fact, this classifier can't find spam, it has no effect.

(d)

0.9750223015165032 0.9724199288256228

(e)

In spam filter, precision is more important because you should not miss any information that should be a normal mail.

In airport, recall is more important because any bomb or drug neglected by the classifier can cause irreparable consequences