Homework

Collaborators:
Name: Zhuo Chen
Student ID: 3170101214

Problem -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* D) *Classification*
4. B) *Unsupervised Learning* G) *Dimensionality Reduction*
5. A) *Supervised Learning* D) *Regression*
6. A) *Supervised Learning* D) *Classification*
7. B) *Unsupervised Learning* F) *Clustering*
8. A) *Supervised Learning* D) *Regression*
9. B) *Unsupervised Learning* G) *Dimensionality Reduction*

(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. We also need a test dataset to evaluate generalization ability of our trained model.

Problem -2. Bayes Decision Rule

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

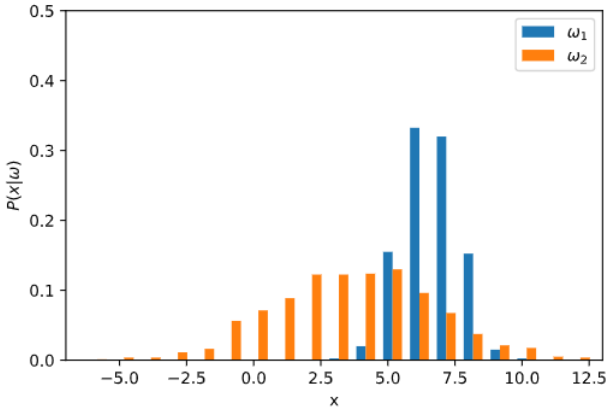
Answer:

1. $P(B_1) = \frac{1}{3}$
2. $P(B_2 = 0|B_1 = 1) = \frac{1}{2}$
3. $$\begin{aligned} P(B_1 = 1|B_2 = 0) &= \frac{P(B_2 = 0|B_1 = 1)P(B_1 = 1)}{P(B_2 = 0)} \\ &= \frac{P(B_2 = 0|B_1 = 1)P(B_1 = 1)}{\sum_{i=1}^{3} P(B_2 = 0|B_i = 1)P(B_i = 1)} \\ &= \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3}} \\ &= \frac{1}{3} \end{aligned}$$
4. Change.

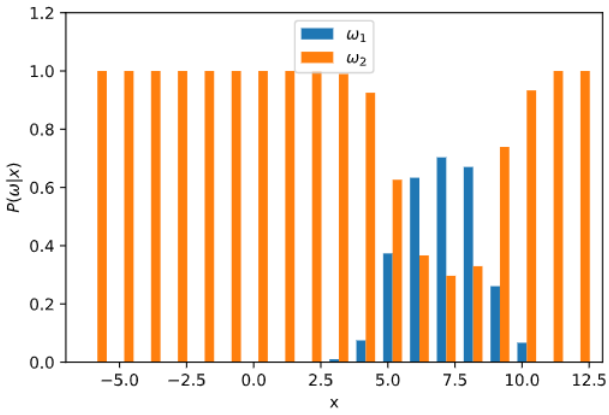
(b) Now let us use bayes decision theorem to make a two-class classifier $\cdots$.

Answer:

1. test error: 3%(3/100) for class 1, 30.5%(61/200) for class 2, 21.3%(64/300) for total.



2. test error: 15%(15/100) for class 1, 16%(32/200) for class 2, 15.7%(47/300) for total.



3. minimum risk: 70.9.

Problem -3. Gaussian Discriminant Analysis and MLE

Given a dataset consisting of m samples. We assume these samples are independently generated by one of two Gaussian distributions$\cdots$

(a) What is the decision boundary?

Answer: $x_1 + x_2 = 1$

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

Answer:

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

Answer:

1. $\mu_0 = (-3,-3)^T, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_1 = (3,3)^T, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \phi = 0.5$
2. $\mu_0 = (0,0)^T, \Sigma_0 = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}, \mu_1 = (-2,0)^T, \Sigma_1 = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \phi = 0.3$
3. $\mu_0 = (0,0)^T, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_1 = (-1,-1)^T, \Sigma_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \phi = 0.5$
4. $\mu_0 = (0,0)^T, \Sigma_0 = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}, \mu_1 = (0,0)^T, \Sigma_1 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}, \phi = 0.5$
5. $\mu_0 = (0,0)^T, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_1 = (0,0)^T, \Sigma_1 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}, \phi = 0.5$
6. $\mu_0 = (0,0)^T, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_1 = (0,0)^T, \Sigma_1 = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}, \phi = 0.5$
7. $\mu_0 = (0,0)^T, \Sigma_0 = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix}, \mu_1 = (0,0)^T, \Sigma_1 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}, \phi = 0.5$
8. $\mu_0 = (0,0)^T, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mu_1 = (0,0)^T, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \phi = 0.5$

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

Answer: WER END

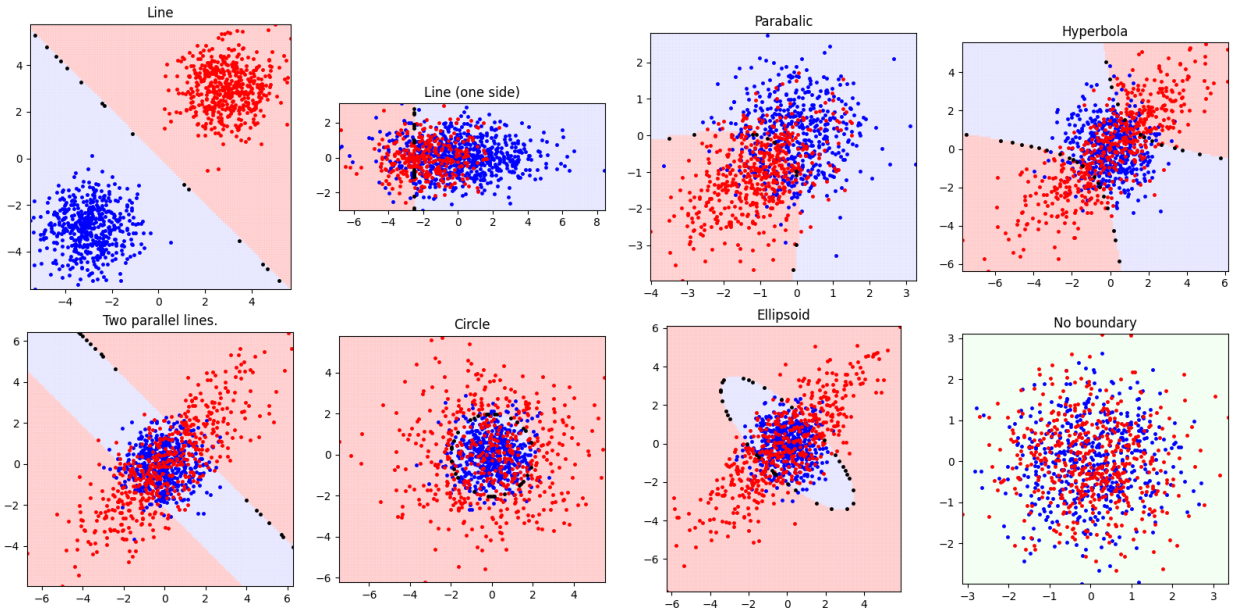


Figure 1: 3(c)

Problem -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: M 4(a) ANSWER END

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

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

(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.

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: 99.1% for ham emails, 97.2% for spam, 98.6% for total.

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

Answer: False. For example, when there are only 100 emails (99 hams and 1 spam), the model which always gives positive prediction can achieve a 99% accuracy.

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

Answer: We can get the following confusion matrix:

	Spam (label)	Ham (label)
Spam (predict)	1093	29
Ham (predict)	31	2983

So, $precision = 97.4\%$, $recall = 97.2\%$.

(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: Precision is more important in a spam filter, because it can be endured if a few spams are not detected. However, recall is dominant in airports security, because the danger resulting from not identifying drugs and bombs is more than the bother brought to passengers.