Homework 1

Collaborators:
Name: Jiayi Cai
Student ID: 3160101462

Problem 1-1. Machine Learning Problems

(a) Choose proper word(s) from

Answer:

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

(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. Validation data should be used to cross-validate the model in order to improve its generalization ability.

Problem 1-2. Bayes Decision Rule

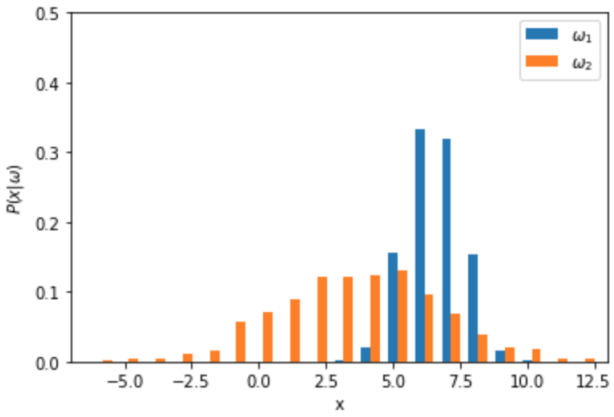
(a) Suppose you are given a chance to win bonus grade points:

Answer:

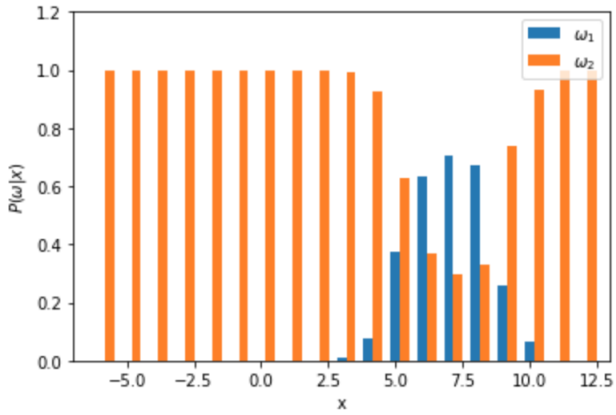
1. P(B_1=1)=$\frac{1}{3}$
2. P(B_2=0|B_1=1)=1
3. P(B_1=1|B_2=0)=$\frac{1}{3}$
4. You should change your choice.

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

Answer:

1. 

test error = 64

2. 

test error = 47

3. $R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x} = 0.24528976503338282$

Problem 1-3. 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:

P(y=1|x) = $\frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)+P(x|y=1)P(y=1)} = \frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{2}e^{-\frac{1}{2}(x1+x2-1)}}$

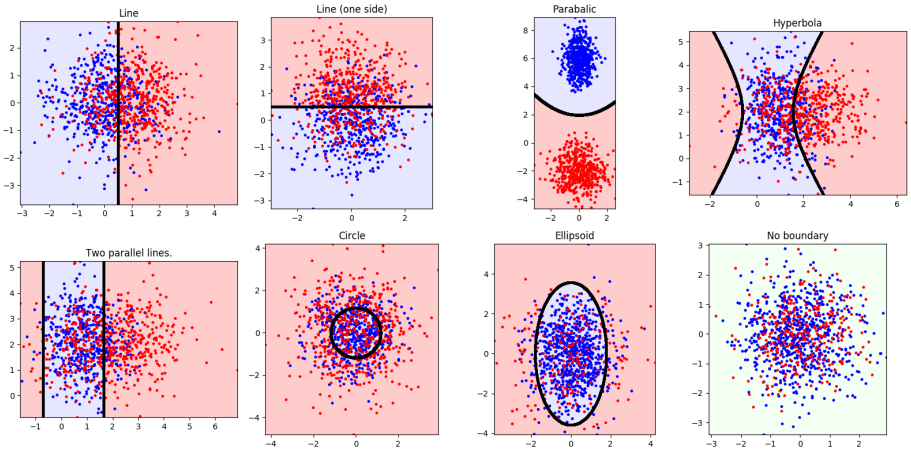
Decision boundary: $x_1 + x_2 = 1$

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

Answer: See code.

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

Answer:



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

Answer:

$\phi = \frac{1}{k}$,

$\hat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i$

$\widehat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} (\mathrm{x}_i - \hat{\boldsymbol{\mu}})(\mathrm{x}_i - \hat{\boldsymbol{\mu}})^t$

Problem 1-4. Text Classification with Naive Bayes

(a) List the top 10 words.

Answer:
30032 nbsp
75525 viagra
38175 pills
45152 cialis
9493 voip
65397 php
37567 meds
13612 computron
56929 sex
9452 ooking

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

Answer:
$accuracy = \frac{TP+FP}{P+Q} = 98.60\%$

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

Answer:
False. In the case when spam:ham=1:99, an always-negative classifier would achieve 99% accuracy, but it is not helpful at all. A good model should be able to distinguish the features between classes.

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

Answer:
TP = 1097, FP = 31, TN = 2980, FN = 27
$precision = \frac{tp}{tp+fp} = \frac{1097}{1097+31} = 97.25\%$
$recall = \frac{tp}{tp+fn} = \frac{1097}{1097+27} = 97.60\%$

(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 a spam filter, precision is more important because people don't want their important emails to be blocked by the filter (which is the case of a false positive). Recall is not so important because false negatives merely result in more trash mails in our inbox.

For a classifier to identify drugs and bombs, recall is more important because we can't afford the consequence of a false negative. Precision is not so important because the cost of a false positive is small (only more inspections on innocent people).