Homework 1

Collaborators:
Name: Shanbin Ke
Student ID: 3160104210

Problem 1-1. Machine Learning Problems

(a) Choose proper word(s) from

Answer:

1. B
2. C
3. A
4. G
5. E
6. D
7. F
8. 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. Using all the data we have to train our model could cause over fitting. Our model could underestimate the error and lose the ability to generalize, thus reducing the accuracy when dealing with data that does not appear in the training set.

A better approach would be to divide the train data to be train data and test data, then use train data to train model and test data to measure the accuracy of our model.

Problem 1-2. Bayes Decision Rule

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

Answer:

1. $\frac{1}{3}$
2. 1
3. $\frac{1}{2}$
4. assume opening box B_2 after choosing box B_1 to be event H and $B_i = 1$ to be box i containing the special prize, $B_i = 0$ to be the opposite.

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

$$P(B_1 = 1) = P(B_2 = 1) = P(B_3 = 1) = \frac{1}{3}$$

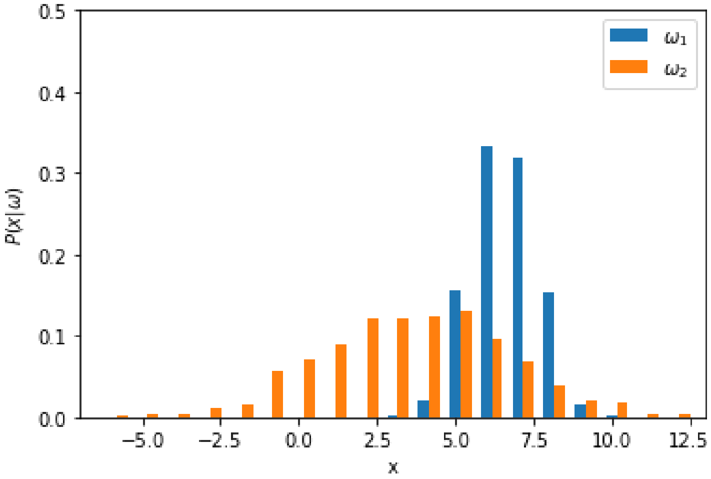
$$P(B_1 = 1|H) = \frac{P(H|B_1 = 1)P(B_1 = 1)}{P(H)} = \frac{\frac{1}{2} * \frac{1}{3}}{P(H)} = \frac{1}{6P(H)}$$

$$P(B_3 = 1|H) = \frac{P(H|B_3 = 1)P(B_3 = 1)}{P(H)} = \frac{1 * \frac{1}{3}}{P(H)} = \frac{1}{3P(H)}$$

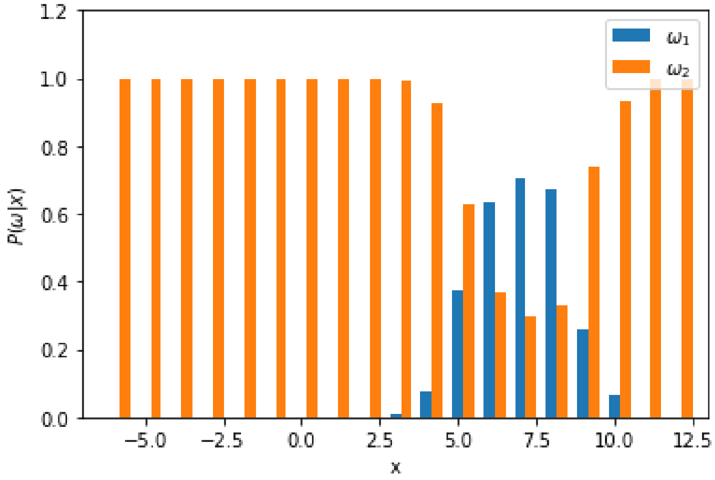
so $P(B_3 = 1|H) > P(B_1 = 1|H)$, the special prize is more likely to be in box B_3, i will change my choice.

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

Answer:



1. test error: 21.34%



2. test error: 15.67%
3. minimal total risk: 0.24

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:

$$p(y=1|X;\phi;\mu_0;\mu_1;\Sigma_0;\Sigma_1) = P(X|y=1)P(y=1) = \frac{1}{2\pi}e^{(-\frac{1}{2}(x_1^2+x_2^2))} * \frac{1}{2}$$

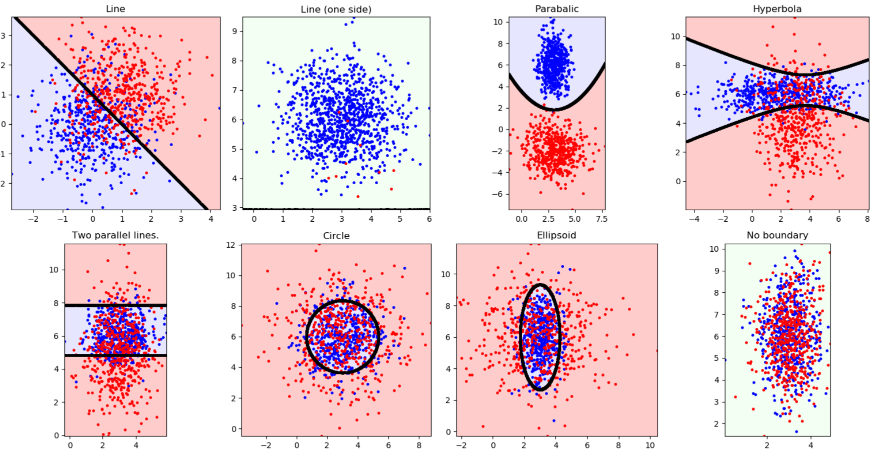
$$p(y=0|X;\phi;\mu_0;\mu_1;\Sigma_0;\Sigma_1) = P(X|y=1)P(y=0) = \frac{1}{2\pi}e^{(-\frac{1}{2}((x_1-1)^2+(x_2-1)^2))} * \frac{1}{2}$$

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: check file *gaussian_pos_prob.py* for implementation.

(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:

$$P(x_k|\mu_0,\phi) = \frac{1}{2\pi\sqrt{|\Sigma_0|}}e^{-\frac{1}{2}(x_k-\mu_0)^T\Sigma_0^{-1}(x_k-\mu_0)} * (1-\phi)$$

$$P(x_k|\mu_1,\phi) = \frac{1}{2\pi\sqrt{|\Sigma_1|}}e^{-\frac{1}{2}(x_k-\mu_1)^T\Sigma_1^{-1}(x_k-\mu_1)} * \phi$$

$$P(D|\mu_0,\mu_1,\phi) = \Pi_{y_k=0}P(x_k|\mu_0,\phi)\Pi_{y_k=1}P(x_k|\mu_1,\phi)$$

$$l(\mu_0, \mu_1, \phi)$$
$$= lnP(D|\mu_0, \mu_1, \phi)$$
$$= \Sigma_{y_k=0}(-\frac{1}{2}ln(\sqrt{2\pi}|\Sigma_0|(1-\phi)) - \frac{1}{2}(x_k-\mu_0)^T\Sigma_0^{-1}(x_k-\mu_0))$$
$$+\Sigma_{y_k=1}(-\frac{1}{2}ln(\sqrt{2\pi}|\Sigma_1|\phi) - \frac{1}{2}(x_k-\mu_1)^T\Sigma_1^{-1}(x_k-\mu_1))$$

$$\bigtriangledown_\phi = \Sigma_{y_k=0}(\frac{1}{2}ln(\sqrt{2\pi}|\Sigma_0|) + \Sigma_{y_k=1}(-\frac{1}{2}ln(\sqrt{2\pi}|\Sigma_1|)$$
$$\bigtriangledown_{\mu_0} = \Sigma_{y_k=0}(\Sigma_0^{-1}(x_k-\mu_0))$$
$$\bigtriangledown_{\mu_1} = \Sigma_{y_k=1}(\Sigma_1^{-1}(x_k-\mu_1))$$

Problem 1-4. Text Classification with Naive Bayes

(a) List the top 10 words.

Answer:

1. nbsp
2. viagra
3. pills
4. cialis
5. voip
6. php
7. meds
8. computron
9. sex
10. width

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

Answer: 98.45%

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

Answer: False.Accuracy is misleading when there is a heavy class imblance, for example, if 99% items belong to class A and 1% items belong to class B, then there are not enough samples for classifier to learn about class B. classifier might just learn to classify everything as class B (in bayesian classifier case, prior of class B will be very small, which leads to small posterior) to achieve 99% percentage accuracy, which is not what we want.

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

Answer: $precision = 96.01\%, recall = 98.40\%$

(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 represents the relevance of the result(may lose a lot of information), and recall represents the completeness of the result(may contain a lot of false information).

For a spam filter, i think precision is more important. If we put recall ahead of precision, then though more spam mails will be identified, more ham mails will go to trash with spam mails too, which is not we want.

For a classifier to identify drugs and bombs at the airport, i think recall is more important. If we put precision ahead of recall, then though it's very likely to be true every time the detector reports about drugs and bombs, it's also more likely to miss cases when drugs and bombs appear, which is very dangerous.