Homework 1

Collaborators:
Name: Yanwei Wang
Student ID: 11821049

Problem 1-1. Machine Learning Problems

(a) Choose proper word(s) from

Answer:

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

(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, Because we not only need the model to fit the training data, but also have ablity of generalization on similar data except for training data.

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. $P(B_1 = 1|B_2 = 0) = \frac{P(B_2=0|B_1=1)P(B_1=1)}{P(B_2=0)} = \frac{1}{3}$
4. $P(B_1 = 1) = \frac{1}{3}$
   $P(B_3 = 1) = 1 - P(B_1 = 0, B_2 = 0) = \frac{2}{3}$

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

Answer:

1. Error num is 64 using MLE
2. Error num is 47 using MAP
3. The minimal total risk is 0.2475

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· · ·

(a) What is the decision boundary?

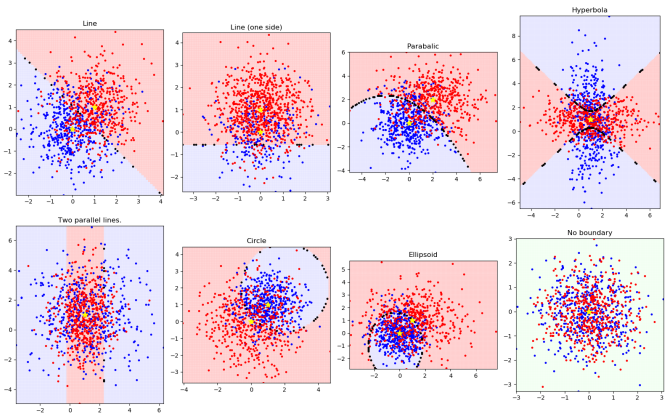
Answer: The boundary is a line, $x + y - 1 = 0$

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

Answer: Pass

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

Answer: For the speed of the code, I change the plot step to 0.1.



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

Answer: I think ϕ is the prior of $p(y)$, and μ,Σ should be calculated respectively,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

,

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})$$

Problem 1-4. Text Classification with Naive Bayes

(a) List the top 10 words.

Answer: aural bihamdihi leary cowpea wxb eep proffer invesco fetid tours

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

Answer: 0.7828

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

Answer: Fasle, because if the data is imbalanced, such as abnormal detection task, one kind data is much larger than the other,even if the model outputs only one value, the accuracy is high.

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

Answer:

	Spam(label)	Ham(label)
Spam(predict)	227	1
Ham(predict)	897	3010

(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: Precison is more important for this question, but recall is more important to identify drugs and bombs. For email, we only need to discard the true spam, for some we can not decide we should keep them in order to prevent losing importtant thing; For bomb detection, even if the possibility of bomb is very small, we should check the luggage.