Homework 1

Collaborators:
Name: Zhao Yi
Student ID: 21921266

Problem 1-1. Machine Learning Problems

(a) Choose proper word(s) from

Answer:

1. B,F
2. C
3. A
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.
The final goal of learning is to get a model that has the best performance on test data set(has the best generalization ability). Due to over-fitting or other issues, parameters working well on train data may not have the same performance on test data. The right way of learning is to leave a portion of data as validation set (no overlap with train set), then use the performance on the validation set as the criteron for model selection.

Problem 1-2. Bayes Decision Rule

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

Answer:

1. $P(B_1 = 1) = 1/3$
2. $P(B_2 = 0|B_1 = 1) = 1$
3. $P(B_1 = 1|B_2 = 0) = \frac{P(B_2=0|B_1=1)*P(B_1=1)}{P(B_2=0)} = \frac{1*1/3}{1} = 1/3$
4. I will change my choice, for $p(B_3 = 1|B_2 = 0) = 1 - P(B_1 = 1|B_2 = 0) = 2/3 > 1/3 = P(B_1 = 1|B_2 = 0)$.

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

Answer:

1. The distribution is shown in Fig 1.
   total err 64/300=21.33%. (x1_test: err 3/100, x2_test: err 61/200)

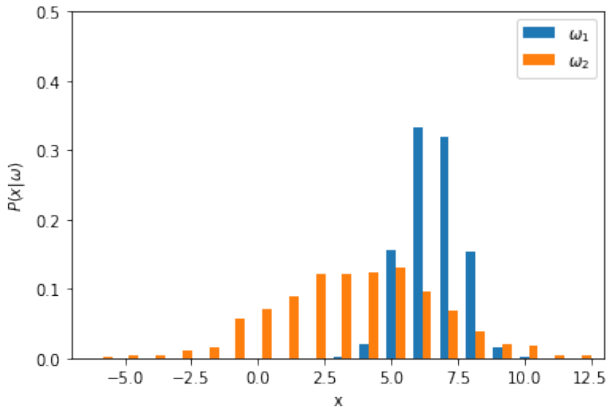

Figure 1: $P(x|w_i)$

2. The distribution is shown in Fig 2.
   total err 47/300=15.667%. (x1_test: err 15/100, x2_test: err 32/200)
3. min total risk is 41.663.

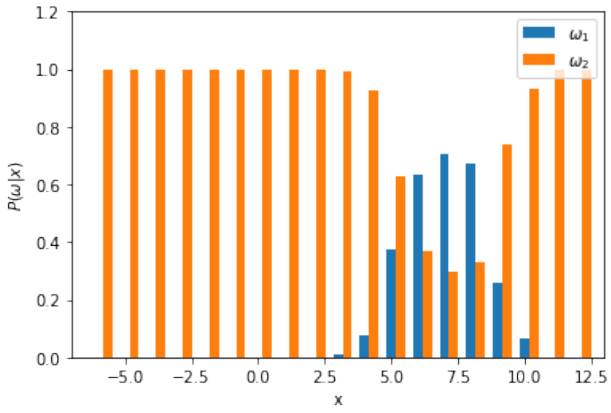

Figure 2: $P(w_i|x)$

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:

$$P(y=1|\mathbf{x}) = \frac{P(\mathbf{x}|y=1) * P(y=1)}{P(\mathbf{x})} = \frac{N(\mu_1, \Sigma_1) * \phi}{P(\mathbf{x})} = \frac{e^{-1/2(\mathbf{x}-\mu_0)^T(\mathbf{x}-\mu_0)}}{4\pi P(\mathbf{x})}$$

Similary,

$$P(y=0|\mathbf{x}) = \frac{e^{-1/2\mathbf{x}^T\mathbf{x}}}{4\pi P(\mathbf{x})}$$

So, the decision boundary is

$$\begin{aligned} P(y=0|\mathbf{x}) &= P(y=1|\mathbf{x}) \\ (\mathbf{x}-\mu_0)^T(\mathbf{x}-\mu_0) &= \mathbf{x}^T\mathbf{x} \\ (x_1-1)^2 + (x_2-1)^2 &= x_1^2 + x_2^2 \\ x_1 + x_2 &= 1 \end{aligned}$$

The decision boundary is line $x_1 + x_2 = 1$

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

Answer:
see hw_1/ml2019fall_hw1/gaussian_discriminant/gaussian_pos_prob.py

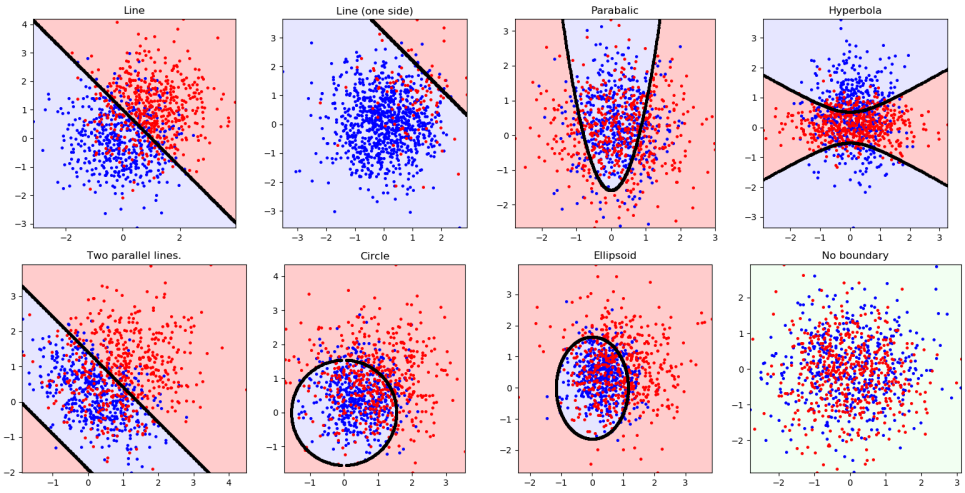

Figure 3: Boundaries of 2 Gaussian, in subfigure 'line (one side)', mean of 2 Gaussian is (1,1) (0,0) respectively

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

Answer: Shown in Fig 3

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

Answer:
Let's denote θ as all parameters. $\phi_k = p(y = k)$, $I[b] = 1(0)$ if b is ture (false). $c_k = \sum_{i=1}^{M} I[y^{(i)} = k]$, $R_k = \{\mathbf{x}_i | y^{(i)} = k\}$

$$
\begin{aligned}
L(\theta) &= -\ln(\prod_{i=1}^{M}(p(\mathbf{x}^{(i)}|y^{(i)},\theta) * p(y^{(i)}|\theta))) \\
&= -\sum_{i=1}^{M}\prod_{k=1}^{K}\ln(p(\mathbf{x}^{(i)}|y^{(i)},\theta)^{I[y^{(i)}=k]} - \sum_{i=1}^{M}\ln(p(y^{(i)}|\theta)) \\
&= \sum_{i=1}^{M}\prod_{k=1}^{K}[\ln(2\pi) - \frac{1}{2}|\Sigma_k| + \frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)]^{I[y^{(i)}=k]} - \sum_{i=1}^{M}\prod_{k=1}^{K}[\ln\phi_k]^{I[y^{(i)}=k]}
\end{aligned}
$$

First, we consider ϕ which is only related to $-\sum_{i=1}^{M}\prod_{k=1}^{K}[\ln\phi_k]^{I[y^{(i)}=k]}$, we want

to maximize this item which equal to

$$\begin{aligned}\phi &= argmin_{\phi} \sum_{i=1}^{M} \prod_{k=1}^{K} [\ln \phi_k]^{I[y^{(i)}=k]} \\ &= argmin_{\phi} \sum_{k=1}^{K} c_k \ln(\phi_k) \\ subject\ to.\ \sum_{k}^{K} \phi_k &= 1\end{aligned}$$

This can be solved by Lagrange minimum multiplier method, and we will get

$$\phi_k = c_k/M$$

Second, we consider μ which is related to former component, to maximize this item, we get

$$\begin{aligned}\mu &= argmax_{\mu} \sum_{k=1}^{K} \sum_{x}^{R_k} \ln(2\pi) - \frac{1}{2}|\Sigma_k| + \frac{1}{2}\left(\mathbf{x}-\mu_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}-\mu_k\right) \\ &= argmax_{\mu} \sum_{k=1}^{K} \sum_{x}^{R_k} \frac{1}{2}\left(\mathbf{x}-\mu_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}-\mu_k\right)\end{aligned}$$

For u_k, we have

$$\mu_k = argmax_{\mu_k} \sum_{x}^{R_k} \frac{1}{2}\left(\mathbf{x}-\mu_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}-\mu_k\right)$$

Let partial derivative w. r. t. μ_k: $-\sum_x^{R_k}(\mathbf{x}-\mu_k) = 0$, we get

$$\mu_k = \frac{\sum_x^{R_k} \mathbf{x}}{c_k}$$

Finally, we substitute μ_k into formula

$$\sum_{x}^{R_k} \frac{1}{2}\left(\mathbf{x}-\mu_k\right)^T \Sigma_k^{-1}\left(\mathbf{x}-\mu_k\right)$$

minimize this formula, we have

$$\Sigma_k = \sum_{x}^{R_k} (\mathbf{x}-\mu_k)^T (\mathbf{x}-\mu_k)$$

In summary, we have

$$\phi_k = c_k/M$$
$$\mu_k = \frac{\sum_x^{R_k} \mathbf{x}}{c_k}$$
$$\Sigma_k = \sum_x^{R_k} (\mathbf{x} - \mu_k)^T (\mathbf{x} - \mu_k)$$

Problem 1-4. Text Classification with Naive Bayes

(a) List the top 10 words.

Answer:

id	word	ratio	occurences in spam	occurences in ham
30032	nbsp	1325.1002358991152	385	0
75525	viagra	1249.5763882571969	363	0
38175	pills	1101.9615951389017	320	0
45152	cialis	847.9268348888121	246	0
9493	voip	837.6281283921868	243	0
65397	php	768.9700850813518	223	0
37567	meds	672.8488244461829	195	0
13612	computron	652.2514114529324	189	0
56929	sex	614.4894876319731	178	0
9452	ooking	518.3682269968041	150	0

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

Answer:

1. ham_test: err 28/all 3011, accuracy:99.07%
2. spam_test: err 31/all 1124, accuracy:97.24%
3. whole test set: err 59/ all 4135, accuracy:98.573%

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

Answer: False.
There exist many unbalanced set (for binary classification) in practice in which the positive (negative) samples dominate. Under this condition, model is influenced too much by prior probability. A model that always gives the same prediction (always positive or always nagetive) can get a pretty good accuracy, but obviously this model is totally useless.

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

Answer:

	spam(label)	ham(label)
spam(predict)	1093	28
ham(predict)	31	2983

- precision: 97.502%
- recall: 97.242%

(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 spam filter, I think the precsion of spam e-mail is important (actually I think the most important item is the recall of ham). Beacuse if a spam filter system causes me to read one more spam e-mail, it doesn't matter much. But if the system causes me to omit a ham e-mail, I may suffer huge loss.
Similarily, for drugs or bombs detection system, the recall is more important than precision. If a security system catches a normal person wrongly, it will be complained at most. However if the security system miss a terrorist, the consequences may be very serious, some people may even lose their lives.
Concretely, you can set different loss of FP and FN and select a model can get the minimal loss.