Introduction to Data Mining: 21191441
Zhejiang University — April 29, 2019
Professor Deng Cai — Homework 1

Homework 1

Collaborators:
Name: Junlin Yin
Student ID: 3160104340

Problem 1-1. Machine Learning Problems

(a) Choose proper word(s) from

Answer:

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

(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.

1. Sometimes the whole data set is so large that we need to spend much of time training the model if we utilize all the data.
2. Aside from training the model, we also need to test it. Therefore, data set should be divided for both training and testing.

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. $P(B_1 = 1|B_2 = 0) = \frac{1}{2}$
4. Let $X = \{$the first choice B_1 is actually right$\}$, $Y = \{$you choose B_1 and B_2 opens without prize$\}$, then

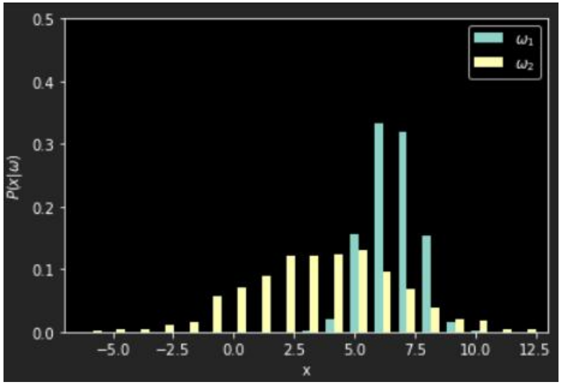
$$\begin{aligned} P(X|Y) &= \frac{P(Y|X) * P(X)}{P(Y|X) * P(X) + P(Y|\bar{X}) * P(\bar{X})} \\ &= \frac{\frac{1}{2} * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + \frac{1}{2} * \frac{2}{3}} \\ &= \frac{1}{3} \end{aligned}$$

That means if you stick to your first choice, you may have a probability of $\frac{1}{3}$ to win the prize, so I will change my choice.

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

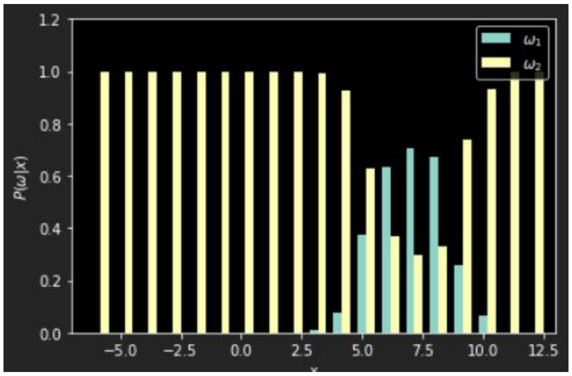
Answer:

1. The likelihood distribution is shown as follows:



There're totally 64 errors when predicting test data, and the error rate is 21.3%.

2. The posterior distribution is shown as follows:



There're totally 47 errors when predicting test data, and the error rate is 15.7%.

3. The minimal total risk is 0.243.

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:

$$\begin{aligned}
P(y=1|x) &= \frac{P(x|y=1)*P(y=1)}{P(x|y=0)*P(y=0)+P(x|y=1)*P(y=1)}\\
&= \frac{N(\mu_1,\Sigma_1)*\phi}{N(\mu_0,\Sigma_0)*(1-\phi)+N(\mu_1,\Sigma_1)*\phi}\\
&= \frac{1}{1+\exp\left[2x_1+2x_2-2\right]}
\end{aligned}$$

Decision boundary:

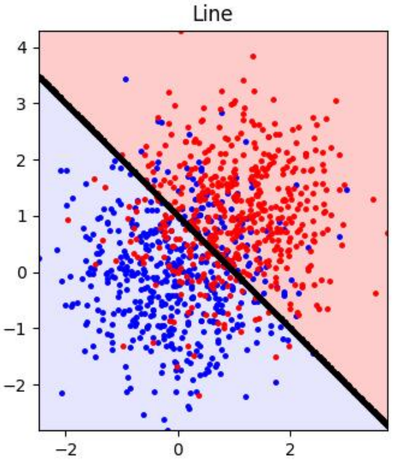
$$\begin{aligned}
P(y=1|x) &= P(y=0|x)\\
P(x|y=1)*P(y=1) &= P(x|y=0)*P(y=0)\\
x_1+x_2 &= 1+\frac{1}{2}*\left[\ln\phi-\ln(1-\phi)\right]\\
x_1+x_2 &= 1
\end{aligned}$$

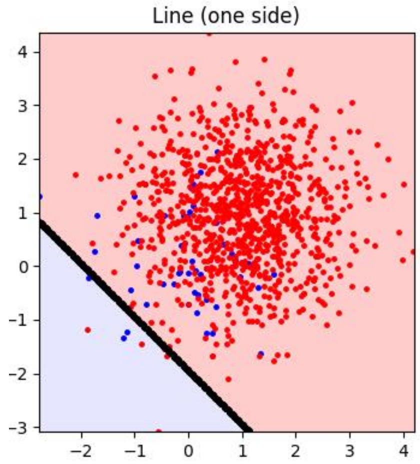
(b) An extension of the above model is to classify K classes by fitting a Gaussian distribution for each class· · ·

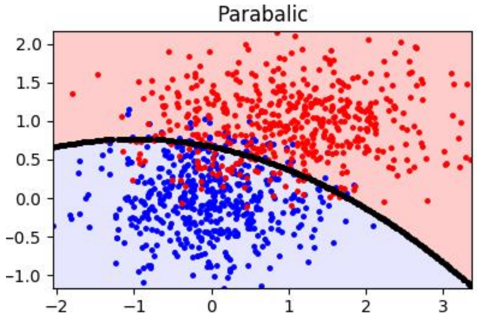
Answer: See the code part.

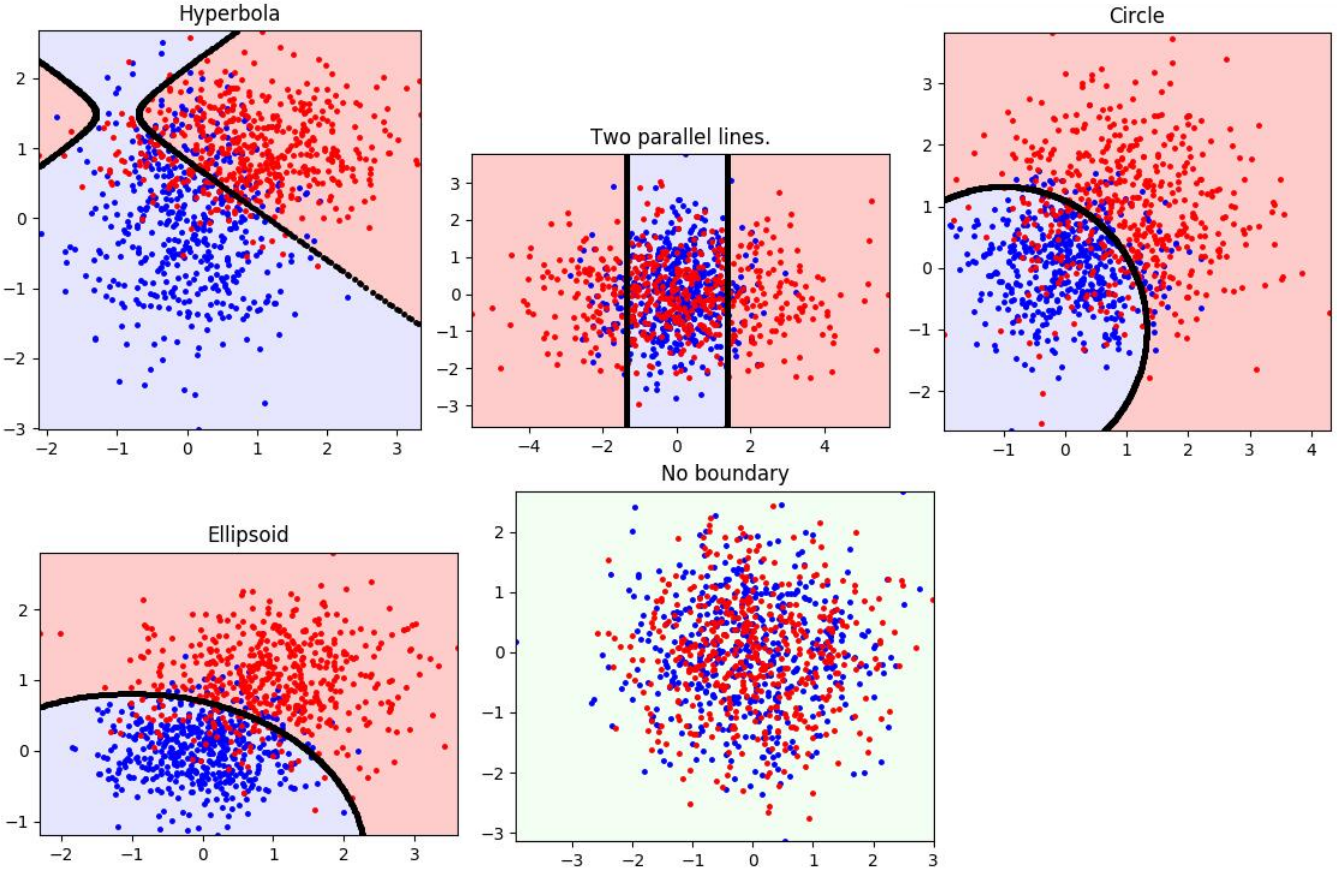
(c) Now let us do some field work – playing with the above 2-class Gaussian discriminant model.

Answer: Here are what I've got in the output:









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

Answer: Suppose there are totally m samples: $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(m)}, y^{(m)})$, and there are totally k classes: $c_1, c_2, ..., c_k$. If one sample $(x^{(1)}, y^{(1)})$ belongs to class k, then $y^{(i)} = k$.
Thus we can divide the whole samples into k groups, such that y values of the samples in the same group are all the class number.

$$
\begin{aligned}
&c_1 : (x_1^{(1)}, y_1^{(1)}), (x_1^{(2)}, y_1^{(2)}), ..., (x_1^{(m_1)}, y_1^{(m_1)})\\
&c_2 : (x_2^{(1)}, y_2^{(1)}), (x_2^{(2)}, y_2^{(2)}), ..., (x_2^{(m_2)}, y_2^{(m_2)})\\
&...\\
&c_k : (x_k^{(1)}, y_k^{(1)}), (x_k^{(2)}, y_k^{(2)}), ..., (x_k^{(m_k)}, y_k^{(m_k)})
\end{aligned}
$$

Obviously, $\sum_{i=1}^{k} m_k = m$. Now consider a single sample and its joint probability:

$$
\begin{aligned}
P(x_j^{(i)}, y_j^{(i)} | \mu_j, \Sigma_j, \phi_j) &= P(x_j^{(i)} | y_j^{(i)}; \mu_j, \Sigma_j, \phi_j) * P(y_j^{(i)} | \mu_j, \Sigma_j, \phi_j)\\
&= P(x_j^{(i)} | y_j^{(i)}; \mu_j, \Sigma_j) * P(y_j^{(i)} | \phi_j)\\
&= N(\mu_j, \Sigma_j) * \phi_j
\end{aligned}
$$

And its log-likelihood is:

$$\begin{aligned} l_{ij} &= \ln P(x_j^{(i)}, y_j^{(i)} | \mu_j, \Sigma_j, \phi_j) \\ &= \ln N(\mu_j, \Sigma_j) + \ln \phi_j \\ &= const - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2}(x_j^{(i)} - \mu_j)^T \Sigma_j^{-1}(x_j^{(i)} - \mu_j) + \ln \phi_j \end{aligned}$$

Then calculate the gradients of the sum-log-likelihood (Note: this sum only includes data in the same group. This makes derivation more succinct.):

$$\begin{aligned} \nabla_{\mu_j} \sum_{i=1}^{m_j} l_{ij} &= \sum_{i=1}^{m_j} \Sigma_j^{-1}(x_j^{(i)} - \mu_j) \\ \nabla_{\Sigma_j} \sum_{i=1}^{m_j} l_{ij} &= -\frac{1}{2}\Sigma_j^{-T} + \frac{1}{2} \sum_{i=1}^{m_j} (x_j^{(i)} - \mu_j)(x_j^{(i)} - \mu_j)^T - \Sigma_j^{-2T} \\ \nabla_{\phi_j} \sum_{i=1}^{m_j} l_{ij} &= \frac{m_j}{\phi_j} \end{aligned}$$

For μ_j and Σ_j, we can solve by letting the gradients be zero:

$$\begin{aligned} \nabla_{\mu_j} \sum_{i=1}^{m_j} l_{ij} = 0 &\Rightarrow \widehat{\mu_j} = \frac{1}{m_j} \sum_{i=1}^{m_j} x_j^{(i)} \\ \nabla_{\Sigma_j} \sum_{i=1}^{m_j} l_{ij} = 0 &\Rightarrow \widehat{\Sigma_j} = \sum_{i=1}^{m_j} (x_j^{(i)} - \widehat{\mu_j})(x_j^{(i)} - \widehat{\mu_j})^T \end{aligned}$$

For ϕ_j, we can observe that l_{ij} stays increasing with ϕ_j, so we need to solve it in another way.

Consider the sum-log-likelihood of all the data (including the ones in different groups):

$$\begin{aligned} L &= \sum_{j=1}^{k} \sum_{i=1}^{m_j} l_{ij} \\ &= constant + \sum_{j=1}^{k} m_j \ln \phi_j \end{aligned}$$

Here the constant means that the factor is nothing to do with any ϕ's. Note that

$$\sum_{j=1}^{k} m_j = m, \sum_{j=1}^{k} \phi_j = 1$$

Replace ϕ_k with $1 - \phi_1 - \phi_2 - ... - \phi_{k-1}$, and then we can rewrite the sum-log-likelihood function like this:

$$L = constant + m_1 \ln \phi_1 + ... + m_{k-1} \ln \phi_{k-1} + m_k \ln (1 - \sum_{i=1}^{k-1} \phi_i)$$

Then calculate the gradients of $\phi_1, \phi_2, ..., \phi_{k-1}$ using MLE:

$$\begin{aligned}
&\nabla_{\phi_1} L = \frac{m_1}{\phi_1} - \frac{m_k}{1 - \sum\limits_{i=1}^{k-1} \phi_i} = 0 \Rightarrow \frac{m_1}{\widehat{\phi_1}} = \frac{m_k}{\widehat{\phi_k}} \\
&\nabla_{\phi_2} L = \frac{m_1}{\phi_2} - \frac{m_k}{1 - \sum\limits_{i=1}^{k-1} \phi_i} = 0 \Rightarrow \frac{m_2}{\widehat{\phi_2}} = \frac{m_k}{\widehat{\phi_k}} \\
&... \\
&\nabla_{\phi_{k-1}} L = \frac{m_{k-1}}{\phi_{k-1}} - \frac{m_k}{1 - \sum\limits_{i=1}^{k-1} \phi_i} = 0 \Rightarrow \frac{m_{k-1}}{\widehat{\phi_{k-1}}} = \frac{m_k}{\widehat{\phi_k}}
\end{aligned}$$

Therefore, we know that ϕ_i must be proportion to m_i:

$$\frac{\widehat{\phi_1}}{m_1} = \frac{\widehat{\phi_2}}{m_2} = ... = \frac{\widehat{\phi_k}}{m_k} = \frac{\sum\limits_{j=1}^{k} \widehat{\phi_j}}{\sum\limits_{j=1}^{k} m_j} = \frac{1}{m}$$

$$\widehat{\phi_j} = \frac{m_j}{m}$$

To sum up, the optimal μ's, Σ's and ϕ's are:

$$\begin{aligned}
&\widehat{\mu_j} = \frac{1}{m_j} \sum_{i=1}^{m_j} x_j^{(i)} \\
&\widehat{\Sigma_j} = \sum_{i=1}^{m_j} (x_j^{(i)} - \widehat{\mu_j})(x_j^{(i)} - \widehat{\mu_j})^T \\
&\widehat{\phi_j} = \frac{m_j}{m}
\end{aligned}$$

respectively, where $j = 1, 2, ..., k$.

Specially in 2-class case where $k = 2$, the optimal parameters are:

$$\begin{aligned}
\widehat{\mu_0} &= \frac{1}{m_0} \sum_{i=1}^{m_0} x_0^{(i)} \\
\widehat{\mu_1} &= \frac{1}{m_1} \sum_{i=1}^{m_1} x_1^{(i)} \\
\widehat{\Sigma_0} &= \sum_{i=1}^{m_0} (x_0^{(i)} - \widehat{\mu_0})(x_0^{(i)} - \widehat{\mu_0})^T \\
\widehat{\Sigma_1} &= \sum_{i=1}^{m_1} (x_1^{(i)} - \widehat{\mu_1})(x_1^{(i)} - \widehat{\mu_1})^T \\
\widehat{\phi_0} &= \frac{m_0}{m} \\
\widehat{\phi_1} &= \frac{m_1}{m}
\end{aligned}$$

Problem 1-4. Text Classification with Naive Bayes

(a) List the top 10 words.

Answer: Top-10 words that are most indicative of the SPAM class:

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

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

Answer: $accuracy = 98.6\%$

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

Answer: No. If the ratio of spam and ham email is 1:99, then we will meet about 1 spam email in 100 emails. If the accuracy of the model is 99%, that means the model will classfy about 1 email by mistake in 100 mails.
If the model definitely classifies all the 100 emails correctly, the probability is $0.99^{100} = 0.366$, so it is not a good model.

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

Answer:

$$\begin{aligned}
&N(TP) = 1093, N(FP) = 28\\
&N(FN) = 31, N(TN) = 2883\\
&recall = \frac{N(TP)}{N(TP) + N(FN)} = 97.2\%\\
&precision = \frac{N(TP)}{N(TP) + N(FP)} = 97.5\%
\end{aligned}$$

(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 span filter, recall is more important because we usually hope to use the model to find as many as possible spam emails, but we don't care much about the precision.

For a drugs and bombs detecter, precision is more important because the cost of mistake can be very large, so we hope to make dicision with higher accuracy.