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

Homework 1

Collaborators:
Name: Jessie Peng
Student ID: xxxxxxxxxx

Problem 1-1. Machine Learning Problems

(a) Choose proper word(s) from

Answer:

1. B) Unsupervised Learning F) Clustering
2. C) Not learning
3. A) Supervised Learning D) Classification
4. B) Unsupervised Learning G) Dimensionality Reduction
5. A) Supervised Learning E) Regression
6. A) Supervised Learning D) Classification
7. B) Unsupervised Learning F) Clustering
8. A) Supervised Learning E) Regression
9. B) Unsupervised Learning G) Dimensionality Reduction

(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 the whole dataset to train the model may achieve good performance on the training set, but the model is likely to be overfitted, which leads to bad performance on the test set. Instead, the model should be trained from part of the dataset and validated by the other. And we should choose the parameters that maximize performance on the validation set.

Problem 1-2. Bayes Decision Rule

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

Answer:

1. The prior probabilities of the prize contained in any of the three boxes are equal, so $P(B_1 = 1) = \frac{1}{3}$
2. Since there is only one prize, $P(B_2 = 0 \mid B_1 = 1) = 1$
3. Since the opened box B_2 must not contain the prize, $\mathrm{P(B_2 = 0) = 1}$, so we have

$$\mathrm{P(B_1 = 1 \mid B_2 = 0) = \frac{P(B_2 = 0 \mid B_1 = 1) \times P(B_1 = 1)}{P(B_2 = 0)}} = \frac{1}{3}$$

4. If the prize is not in B_2, then it's either in B_1 or in B_3, which means

$$P(B_1 = 1 \mid B_2 = 0) + P(B_3 = 1 \mid B_2 = 0) = 1$$

According to the Bayes decision rule, we should compare the following two posterior probabilities:

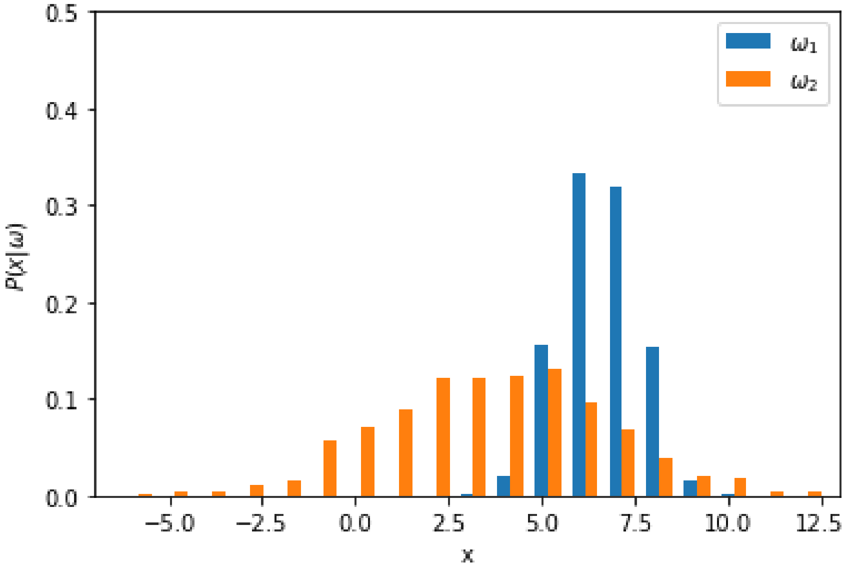
$$P(B_1 = 1 \mid B_2 = 0) = \frac{1}{3},\ P(B_3 = 1 \mid B_2 = 0) = \frac{2}{3}$$

Therefore, $P(B_1 = 1 \mid B_2 = 0) < P(B_3 = 1 \mid B_2 = 0)$ indicates that I should change my choice to B_3.

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

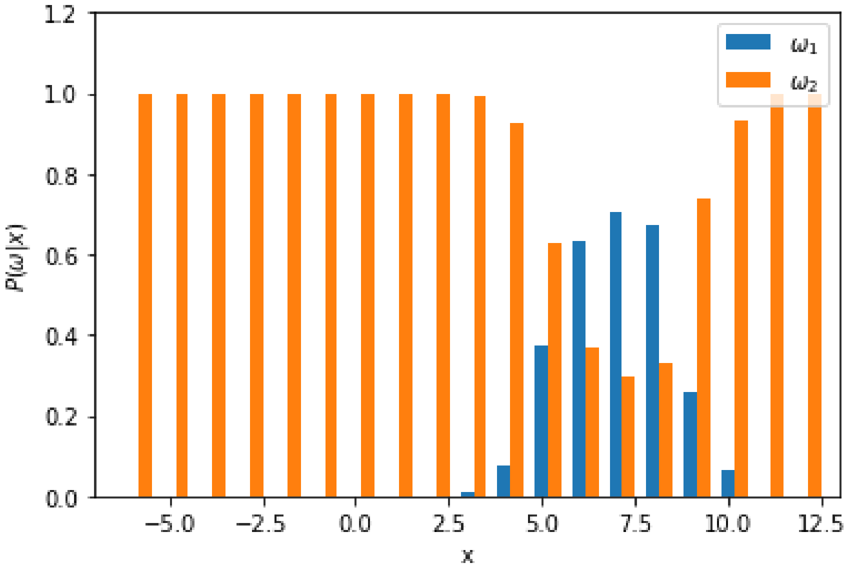
Answer:

1. The distribution of likelihood is shown below:



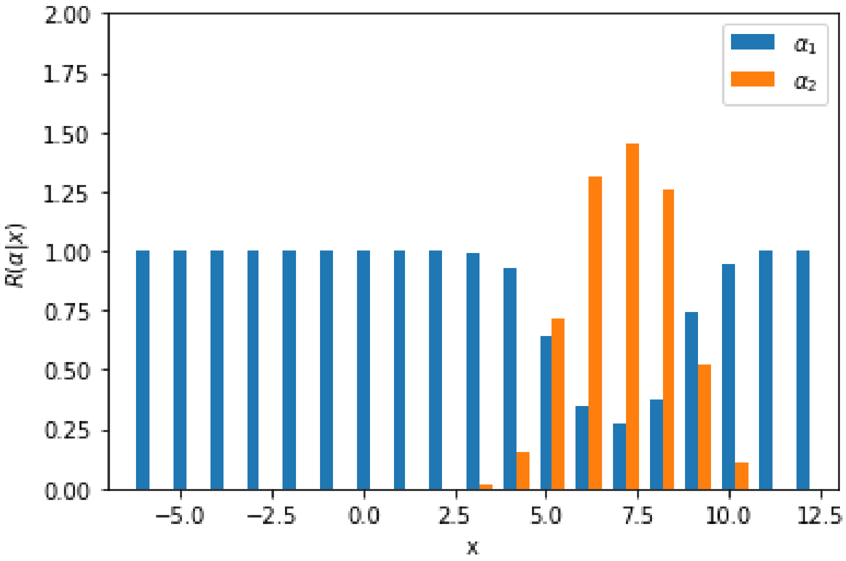
The number of misclassified samples is 64.
The test error using maximum likelihood decision rule is 0.2133.

2. The distribution of posterior probability is shown below:



The number of misclassified samples is 47.
The test error using optimal Bayes decision rule is 0.1567.

3. The distribution of risk is shown below:



The minimal total risk: $R = \sum_x \min_i R(\alpha_i \mid x) \times P(x) = 0.2427$.

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:

$$
\begin{aligned}
p(y=1) &= \theta = \frac{1}{2} \\
p(y=0) &= 1-\theta = \frac{1}{2} \\
p(x \mid y=0) &= N(\mu_0, \Sigma_0) = N\left((0,0)^T, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\
&= \frac{1}{2\pi\sqrt{|\Sigma_0|}} e^{-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0)} \\
&= \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} \\
p(x \mid y=1) &= N(\mu_1, \Sigma_1) = N\left((1,1)^T, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\
&= \frac{1}{2\pi\sqrt{|\Sigma_1|}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)} \\
&= \frac{1}{2\pi} e^{-\frac{1}{2}((x_1-1)^2+(x_2-1)^2)}
\end{aligned}
$$

To obtain the decision boundary, we need the posterior $p(y=0 \mid x)$ and $p(y=1 \mid x)$ to be equal, which can be written as

$$
\begin{aligned}
p(x \mid y=0) \times p(y=0) &= p(x \mid y=1) \times p(y=1) \\
\frac{1}{4\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)} &= \frac{1}{4\pi} e^{-\frac{1}{2}((x_1-1)^2+(x_2-1)^2)} \\
x_1^2 + x_2^2 &= (x_1-1)^2 + (x_2-1)^2 \\
x_1 + x_2 &= 1
\end{aligned}
$$

Therefore, the decision boundary is $x_1 + x_2 = 1$.

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

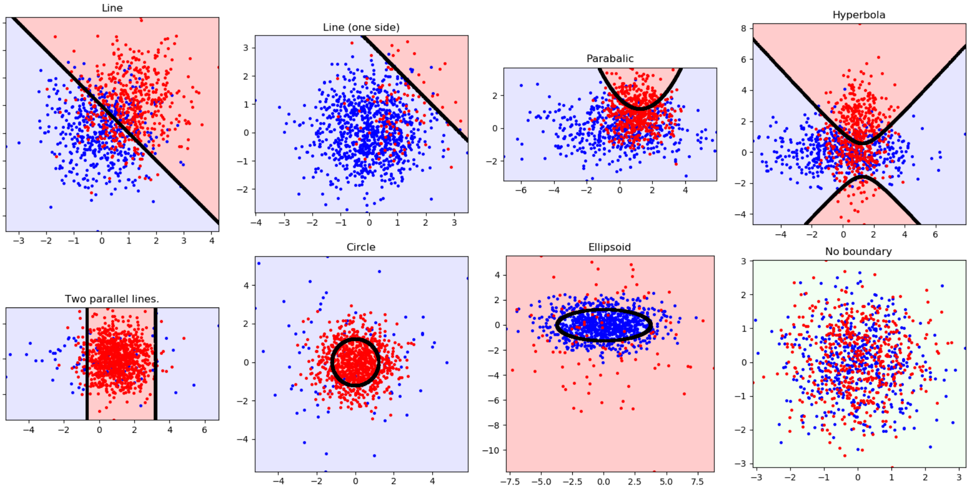
Answer:

$$\begin{aligned} p(y = k \mid x) &= \frac{p(x \mid y = k) \times p(y = k)}{p(x)} \\ &= \frac{p(x \mid y = k) \times p(y = k)}{\sum_{i=1}^{K} p(x \mid y = i) \times p(y = i)} \\ &= \frac{N(\mu_k, \Sigma_k) \times \phi_k}{\sum_{i=1}^{K} N(\mu_i, \Sigma_i) \times \phi_i} \\ &= \frac{\frac{\phi_k}{2\pi\sqrt{|\Sigma_k|}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}}{\sum_{i=1}^{K} \frac{\phi_i}{2\pi\sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}} \end{aligned}$$

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

Answer: The parameters I use and the resulting plots are shown below:

Type of Decision Boundary	$(\phi, \mu_0, \mu_1, \sigma_0, \sigma_1)$
Line	$\left(0.5, (0,0)^T, (1,1)^T, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
Line (one side)	$\left(0.1, (0,0)^T, (1,1)^T, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
Parabolic	$\left(0.5, (0,0)^T, (1,1)^T, \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
Hyperbola	$\left(0.5, (0,0)^T, (1,1)^T, \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\right)$
Two parallel lines	$\left(0.9, (0,0)^T, (1,0)^T, \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
Circle	$\left(0.9, (0,0)^T, (0,0)^T, \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
Ellipsoid	$\left(0.1, (0,0)^T, (0,0)^T, \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}\right)$
No boundary	$\left(0.5, (0,0)^T, (0,0)^T, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$



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

Answer: In terms of the general K-class Gaussian model, we assume the dataset $D = \{(x^{(i)}, y^{(i)}) \mid y \in \{0,1\}, i = 1, ..., m\}$ is partitioned into $D_1, D_2, ..., D_K$ in accordance with the K classes.
The parameter ϕ_k stands for the prior probability of class k, and thus can be estimated by the density of D_k:

$$\hat{\phi}_k = \frac{|D_k|}{|D|}, \forall k \in \{1, 2, ..., K\}$$

For class k ($\forall k \in 1, 2, ..., K$), we estimate the parameter μ_k and Σ_k by maximizing the following likelihood:

$$\begin{aligned} p(D_k \mid (\mu_k, \Sigma_k)) &= \prod_{x \in D_k} P(x \mid (\mu_k, \Sigma_k)) \\ &= \prod_{x \in D_k} \frac{1}{2\pi\sqrt{|\Sigma_k|}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)} \end{aligned}$$

This is equivalent to maximizing the log-likelihood:

$$\ln p(D_k \mid (\mu_k, \Sigma_k)) = \sum_{x \in D_k} -\frac{1}{2} \ln(4\pi^2 \Sigma_k) - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)$$

The gradient with respect to the parameters is given by

$$\triangledown_{(\mu_k, \Sigma_k)} \ln p(D_k \mid (\mu_k, \Sigma_k)) = \begin{bmatrix} \sum_{x \in D_k} \frac{1}{\Sigma_k}(x - \mu_k) \\ \sum_{x \in D_k} -\frac{1}{2\Sigma_k} + \frac{(x-\mu_k)^2}{2\Sigma_k^2} \end{bmatrix}$$

By setting the gradient to zero, we find the maximum likelihood estimation of the parameters as

$$\begin{aligned}
\hat{\mu_k} &= \frac{1}{|D_k|}\sum_{x\in D_k} x \\
\hat{\Sigma_k} &= \frac{1}{|D_k|}\sum_{x\in D_k}(x-\hat{\mu_k})^2 \\
&= \frac{1}{|D_k|}\sum_{x\in D_k}(x-\frac{1}{|D_k|}\sum_{x\in D_k} x)^2
\end{aligned}$$

In the multivariate scenario, assume $|D_1| = |D_2| = ... = |D_K| = n$, then the parameters can be written in vector form as follows:

$$\begin{aligned}
\hat{\boldsymbol{\mu}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x_i} \\
\hat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i}-\hat{\boldsymbol{\mu}})(\boldsymbol{x_i}-\hat{\boldsymbol{\mu}})^T
\end{aligned}$$

Problem 1-4. Text Classification with Naive Bayes

(a) List the top 10 words.

Answer: Here are the top 10 words indicative of SPAM (there is a tie at rank 10):

Rank	Index No.	Word	$\frac{P(word_i\|SPAM)}{P(word_i\|HAM)}$
1	30033	nbsp	1325
2	75526	viagra	1250
3	38176	pills	1102
4	45153	cialis	848
5	9494	voip	838
6	65398	php	769
7	37568	meds	673
8	13613	computron	652
9	56930	sex	614
10	9453	ooking	518
10	19957	width	518

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

Answer:

$TP = 1093$	$FP = 28$
$FN = 31$	$TN = 2983$

$$\begin{aligned} accuracy &= \frac{(TP + TN)}{(number\ of\ test\ samples)} \\ &= \frac{1093 + 2983}{1124 + 3011} \\ &= 0.9857 \end{aligned}$$

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

Answer: False. Accuracy is a relative measurement depending on specific situations. 99% can be high in some cases, but low in others. And accuracy alone cannot fully reflect the quality of a model. Take the spam filter as an example, if we have 100 emails, then 99% accuracy suggests $TP + TN = 99$, so it's possible that $FN = 1$ (and $FP = 0$). Note that the ratio of spam and ham email is 1:99, so the only spam is identified as a ham email, which means this filter does no help at all (it always outputs a negative prediction).

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

Answer:

$$\begin{aligned} precision &= \frac{TP}{TP+FP} \\ &= \frac{1093}{1093+28} \\ &= 0.9750 \\ recall &= \frac{TP}{TP+FN} \\ &= \frac{1093}{1093+31} \\ &= 0.9724 \end{aligned}$$

(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 spam filter, precision is more important than recall because falsely identifying a ham email as a spam will do greater harm than failing to filter out a spam.

However, for a drug or bomb classifier at airport, recall is more important than precision because it's better to identify more suspects (even though many of them can be proved innocent) than to leave out any potential risk.