Assignment 2_conceptual

Pong Yui Yi Monica SID: 20853295

2024-11-3

question 4.1

> Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

answer:

From (4.3), we have

$$\begin{aligned}\frac{p(X)}{1-p(X)} &= \exp(\beta_0+\beta_1 X)\\ p(X) &= \exp(\beta_0+\beta_1 X) - p(X)\exp(\beta_0+\beta_1 X)\\ p(X)(1+\exp(\beta_0+\beta_1 X)) &= \exp(\beta_0+\beta_1 X)\\ p(X) &= \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\end{aligned}$$

which is equivalent to (4.2).

question 4.2

> It was stated in the text that classifying an observation to the class for which (4.17) is largest is equivalent to classifying an observation to the class for which (4.18) is largest. Prove that this is the case. In other words, under the assumption that the observations in the kth class are drawn from a $N(\mu_k, \sigma^2)$ distribution, the Bayes classifier assigns an observation to the class for which the discriminant function is maximized.

answer:

the discriminant function is

$$\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma_2} + \log(\pi_k) \tag{4.18}$$

(4.17) is

$$
\begin{aligned}
p_k(x) &= \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)} \\
p_k(x) &= \frac{\pi_k \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \\
\log(p_k(x)) &= \log(\pi_k) - \frac{1}{2\sigma^2}(x-\mu_k)^2 - \log\left(\sum_{l=1}^{K} \pi_l \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)\right)
\end{aligned}
$$

Since we are maximizing over k, and the last term does not vary with k, which could be regarded as a constant. So we just need to maximize

$$
\begin{aligned}
f &= \log(\pi_k) - \frac{1}{2\sigma^2}(x^2 - 2x\mu_k + \mu_k^2) && (1) \\
&= \log(\pi_k) - \frac{x^2}{2\sigma^2} + \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} && (2)
\end{aligned}
$$

Since $\frac{x^2}{2\sigma^2}$ is also independent of k, so we could just maximize the equation as followed:

$$\log(\pi_k) + \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

Which is equivalent to the discriminant function (4.18).

question 4.3

> This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a classspecific mean vector and a class specific covariance matrix. We consider the simple case where $p = 1$; i.e. there is only one feature. Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma^2)$. Recall that the density function for the one-dimensional normal distribution is given in (4.16). Prove that in this case, the Bayes classifier is not linear. Argue that it is in fact quadratic.
> *Hint: For this problem, you should follow the arguments laid out in Section 4.4.1, but without making the assumption that* $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$.

answer:

Considering the Bayes classifier without assuming $\sigma_1^2 = ... = \sigma_K^2$
From Question 4.2, we have:

$$
\begin{aligned}
p_k(x) &= \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k}} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}{\sum_{l=1}^{k} \pi_l \frac{1}{\sqrt{2\pi\sigma_l}} \exp(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2)} \\
\log(p_k(x)) &= \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi\sigma_k}}\right) - \frac{1}{2\sigma_k^2}(x-\mu_k)^2 - \log\left(\sum_{l=1}^{k} \frac{1}{\sqrt{2\pi\sigma_l}} \pi_l \exp\left(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\right)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
f &= \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi\sigma_k}}\right) - \frac{1}{2\sigma_k^2}(x-\mu_k)^2 && (3)\\
&= \log(\pi_k) + \log\left(\frac{1}{\sqrt{2\pi\sigma_k}}\right) - \frac{x^2}{2\sigma_k^2} + \frac{x\mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} && (4)
\end{aligned}
$$

However, since we do not assume $\sigma_1^2 = ... = \sigma_K^2$, $\frac{x^2}{2\sigma_k^2}$ is not independent of k, so we retain the term with x^2, hence f, the Bayes' classifier, is a quadratic function of x.

question 5.1

Using basic statistical properties of the variance, as well as singlevariable calculus, derive (5.6). In other words, prove that α given by (5.6) does indeed minimize $Var(\alpha X + (1-\alpha)Y)$.

answer:

(5.6) is:

$$
\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \qquad (5.6)
$$

Then the risk is:

$$
\begin{aligned}
Var(\alpha X + (1-\alpha)Y) &= \alpha^2\sigma_X^2 + (1-\alpha)^2\sigma_Y^2 + 2\alpha(1-\alpha)\sigma_{XY}\\
&= \alpha^2\sigma_X^2 + \sigma_Y^2 - 2\alpha\sigma_Y^2 + \alpha^2\sigma_Y^2 + 2\alpha\sigma_{XY} - 2\alpha^2\sigma_{XY}
\end{aligned}
$$

Since we want to minimize the risk by variate the α, we set the partial derivative of the risk with respect to α to 0:

$$
\begin{aligned}
\frac{\partial}{\partial\alpha} &= 2\alpha\sigma_X^2 - 2\sigma_Y^2 + 2\alpha\sigma_Y^2 + 2\sigma_{XY} - 4\alpha\sigma_{XY} = 0\\
\alpha\sigma_X^2 + \alpha\sigma_Y^2 - 2\alpha\sigma_{XY} &= \sigma_Y^2 - \sigma_{XY}\\
\alpha &= \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}
\end{aligned}
$$

And the second partial derivative is as followed:

$$
\begin{aligned}
\frac{\partial^2}{\partial\alpha^2} &= 2\sigma_X^2 + 2\sigma_Y^2 - 4\sigma_{XY}\\
&= 2(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\\
&= 2\mathrm{Var}(X - Y)
\end{aligned}
$$

Since variance is positive, then this must be positive. $Var(\alpha X + (1-\alpha)Y)$ is the minimum with respect to α given by (5.6).

question 5.2

We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

(a). What is the probability that the first bootstrap observation is *not* the jth observation from the original sample? Justify your answer.
(b). What is the probability that the second bootstrap observation is *not* the jth observation from the original sample?
(c). Argue that the probability that the jth observation is *not* in the bootstrap sample is $(1 - 1/n)^n$.
(d). When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?
(e). When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?
(f). When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample?
(g). Create a plot that displays, for each integer value of n from 1 to 100,000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.
(h). We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

```
> store <- rep (NA, 10000)
> for (i in 1:10000) {
    store[i] <- sum(sample(1:100, rep = TRUE) == 4) > 0
}
> mean(store)
```

Comment on the results obtained.

(a). This is 1 - probability that it is the jth $= 1 - 1/n$.

(b). bootstrap observation is a random sample, this probability is the same $(1 - 1/n)$.

(c). For the jth observation to not be in the sample, it would have to *not* be picked for each of n positions, so not picked for $1, 2, ..., n$, thus the probability is $(1 - 1/n)^n$

(d).

```
n <- 5
1 - (1 - 1 / n)^n
```

```
## [1] 0.67232
```

$p = 0.67$

(e).

```
n <- 100
1 - (1 - 1 / n)^n
```

```
## [1] 0.6339677
```

$p = 0.64$

(f).

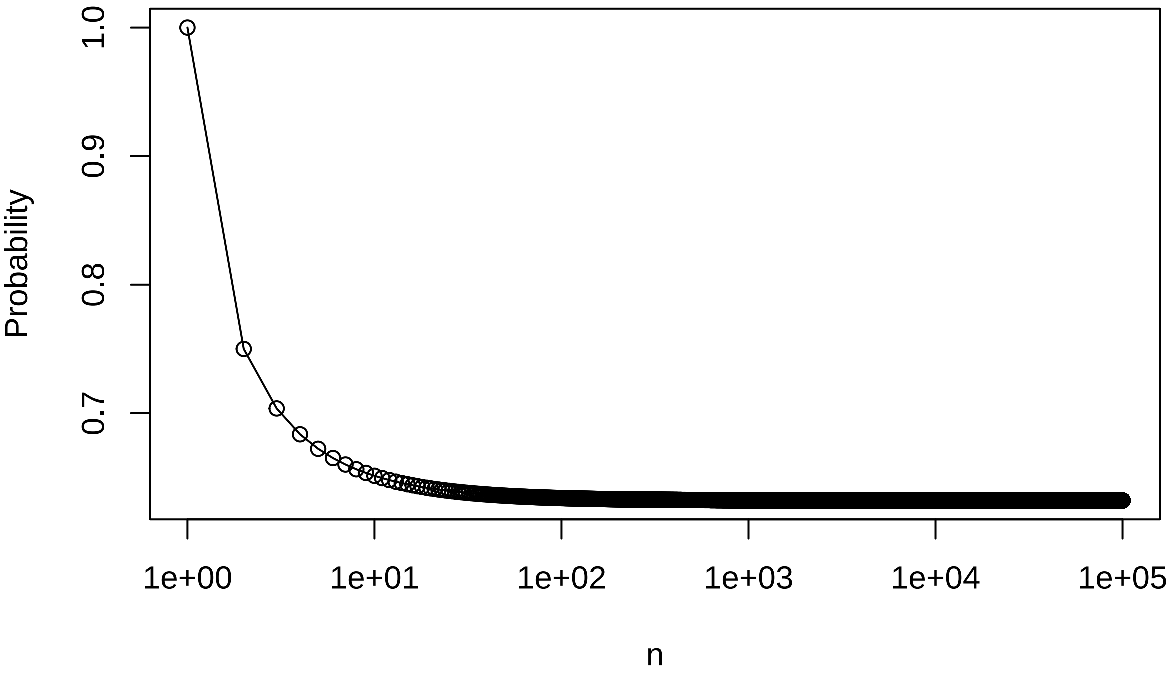
```r
n <- 100000
1 - (1 - 1 / n)^n
```

```
## [1] 0.6321224
```

$p = 0.63$

(g).

```r
x <- sapply(1:100000, function(n) 1 - (1 - 1 / n)^n)
plot(x, log = "x", type = "o", xlab = "n", ylab = "Probability")
```



The probability approaches 0.632 with increasing n.

Note that

$$e^x = \lim_{x \to \text{inf}} \left(1 + \frac{x}{n}\right)^n,$$

so with $x = -1$, we can see that our limit is $1 - e^{-1} = 1 - 1/e = 0.632$.

(h).

```r
store=rep(NA, 10000)
for(i in 1:10000){
  store[i] = sum(sample(1:100, rep=TRUE)==4)>0
}
mean(store)
```

```
## [1] 0.6387
```

The probability of including 4 when resampling numbers 1...100 is close to $1-(1-1/100)^{100}$, and the result is similar to the value calculated using the formula in 2(e).