HOMEWORK 4

Yohei Nishimura
ynishimura

Solution 1

Solution 1.1: Strategy 1

If we $\hat{x} \in argmax_x \theta_x$, and if let j be the x at which the largest probability is taken, then $E[1[\hat{x} = j]] = \theta_j$. Therefore, $E[1[\hat{x} \neq j]] = 1 - \theta_j$. By changing this notation j to x, we get the following equation.

$$E[1[\hat{x} \neq x]] = 1 - \theta_x$$

Solution 1.2: Strategy 2

Suppose $\hat{x}$ is generated from a multinomial distribution. $\hat{x} = x$ because the probability of x and $\hat{x}$ occurring is θ_x, which is θ_x^2 since the two events are independent. Therefore, we get below:

$$E[1[\hat{x} \neq x]] = 1 - \theta_x^2$$

Solution 2

$E[c_{ij}]$ can be formulated as follows.

$$\begin{aligned} E[c_{ij}] &= \frac{1}{k}(\theta_1 1[\hat{x} \neq 1] + \theta_2 1[\hat{x} \neq 2] + ... + \theta_k 1[\hat{x} \neq k]) \\ &= \frac{1}{k}((1-\theta_1)1[\hat{x} = 1] + (1-\theta_2)1[\hat{x} = 2] + ... + (1-\theta_k)1[\hat{x} = k]) \end{aligned}$$

In order to minimize $E[c_{ij}]$, we should choose $\hat{x}$ that minimizes the above equation. The strategy of the choice of $\hat{x}$, since we know θ, we can choose the i that minimizes $(1 - \theta_i), i = 1, ..., k$, that is, the largest θ_i in θ, as $\hat{x}$.

Solution 3

Solution 3.1

Let the mean of the distribution of Y_t be μ and the variance be σ^2. In the setting of this question, since we know these two variables, the optimal strategy is to set $x_t = \mu$ at any time. It is because $E[y_t] = \mu$, meaning that the value of y_t that is most likely to appear is μ.
At this time, the expected payment for the Tth time is as follows:

$$\begin{aligned} E[T(x_t - y_t)^2] &= T(\mu^2 - 2\mu E[y_t] + E[y_t^2])) = T(\mu^2 - 2\mu^2 + \sigma^2 + \mu^2) \\ &= T\sigma^2 \end{aligned}$$

During the development of the above equation, I used the following:

$$\begin{aligned} Var(y_t) = \sigma^2 = E[y_t^2] - (E[y_t])^2 = E[y_t^2] - \mu^2 \\ \therefore E[y_t^2] = \sigma^2 + \mu^2 \end{aligned}$$

Solution 3.2

Based on previous problem, we can set the "payment" function below:

$$argmin_{x_t} Payment = T(E[x_t]^2 - 2E[x_t]E[y_t] + E[y_t^2])] \tag{1}$$

At the end of the mth enforcement, calculate the following with the realized values.

$$E[x_t] = \frac{1}{m}\sum_{i=1}^{m} x_i = \bar{x}$$
$$E[x_t]^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2 + \bar{x}^2 = Var(x) + \bar{x}^2$$
$$E[y_t] = \frac{1}{m}\sum_{i=1}^{m} y_i = \bar{y}$$
$$E[y_t]^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})^2 + \bar{y}^2 = Var(y) + \bar{y}^2$$

Then, substitute the above four equations into equation (1).

$$\begin{aligned} argmin_{x_t} Payment &= T(\frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2 + \bar{x}^2 + \frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})^2 + \bar{y}^2 - 2\bar{x}\bar{y}) \\ &= T(Var(x) + Var(y) + (\bar{x} - \bar{y})^2) \end{aligned} \tag{2}$$

Calculate equation (2) above for the $m+1$th time as well, and if equation (2), or the benchmark, is lower than the mth time, we can conclude that the strategy we are currently using is correct. In other words, equation (2) is the benchmark.

Solution 4

Solution 4.1

From the setup of this question, $N = 30$, $K_L = 3$, and $\alpha = 0.5$. Using these, calculate the prior probabilities. The total number of data is 30, e0-e9, j0-j9, s0-s9, of which 10 are in English.

$$\begin{aligned} \hat{p}_\alpha(y = e) &= \frac{\sum_{i=1}^{N} 1[y^{(i)} = e] + \alpha}{N + K_L\alpha} \\ &= \frac{10 + 0.5}{30 + 3 * 0.5} = \frac{1}{3} \end{aligned}$$

Similarly, the two prior probabilities of ta are calculated and the answers are as follows.

$$\hat{p}_\alpha(y = e) = \hat{p}_\alpha(y = j) = \hat{p}_\alpha(y = s) = \frac{1}{3}$$

Solution 4.2

From the given conditions, calculate the following using $p(y = e) = \frac{1}{3}$.

$$\theta_{i,e} = \hat{p}(c_i|y = e) = \frac{\hat{p}(c_i \cap y = e)}{\hat{p}(y = e)} = 3\hat{p}(c_i \cap y = e)$$

Then, compute the θ for each i. θ_e is below.

Table 1: θ_e

character	probability
a	0.06
b	0.01
c	0.02
d	0.02
e	0.11
f	0.02
g	0.02
h	0.05
i	0.05
j	0.0
k	0.0
l	0.03
m	0.02
n	0.06
o	0.07
p	0.02
q	0.0
r	0.05
s	0.07
t	0.08
u	0.03
v	0.01
w	0.02
x	0.0
y	0.01
z	0.0
(space)	0.18

Solution 4.3

θ_j, θ_s,the class conditional probabilities for Japanese and Spanish are below.

Table 2: θ_j

character	probability
a	0.13
b	0.01
c	0.01
d	0.02
e	0.06
f	0.0
g	0.01
h	0.03
i	0.1
j	0.0
k	0.06
l	0.0
m	0.04
n	0.06
o	0.09
p	0.0
q	0.0
r	0.04
s	0.04
t	0.06
u	0.07
v	0.0
w	0.02
x	0.0
y	0.01
z	0.01
(space)	0.12

Table 3: θ_s

character	probability
a	0.1
b	0.01
c	0.04
d	0.04
e	0.11
f	0.01
g	0.01
h	0.0
i	0.05
j	0.01
k	0.0
l	0.05
m	0.03
n	0.05
o	0.07
p	0.02
q	0.01
r	0.06
s	0.07
t	0.04
u	0.03
v	0.01
w	0.0
x	0.0
y	0.01
z	0.0
(space)	0.17

Solution 4.4

The bag-of-words vector x is below.

Table 4: x of e10.txt

character	numbers
a	164
b	32
c	53
d	57
e	311
f	55
g	51
h	140
i	140
j	3
k	6
l	85
m	64
n	139
o	182
p	53
q	3
r	141
s	186
t	225
u	65
v	31
w	47
x	4
y	38
z	2
(space)	498

Solution 4.5

Taking the logarithm of the equation, we get $\log \hat{p}(x|y) = \sum_{i=1}^{d} x_i \theta_{i,y}$. Based on previous problems, $\log \hat{p}(x|y = e) = -7840.77$, $\log \hat{p}(x|y = j) = -8727.17$, and $\log \hat{p}(x|y = s) = -8478.20$, meaning $\hat{p}(x|y = e) = e^{-7840.77}$, $\hat{p}(x|y = j) = e^{-8727.17}$, and $\hat{p}(x|y = s) = e^{-8478.20}$

Solution 4.6

Taking the logarithm of the Bayes law, we get $\log \hat{p}(y|x) = \log \hat{p}(x|y) + \log \hat{p}(y)$. Then, based on previous problems, $\log \hat{p}(y = e|x) = -7841.87$, $\log \hat{p}(y = j|x) = -8728.27$, and $\log \hat{p}(y = s|x) = -8479.30$, meaning $\hat{p}(y = e|x) = e^{-7841.87}$, $\hat{p}(y = j|x) = e^{-8728.27}$, and $\hat{p}(y = s|x) = e^{-8479.30}$. In the calculation, the results of prior probability in problem 4.2 are used.
Because $\hat{p}(y = e|x) > \hat{p}(y = s|x) > \hat{p}(y = j|x)$, it is concluded that the label of x should be "English."

Solution 4.7

The answer table is below.

Table 5: Classification among English, Spanish and Japanese

	English	Spanish	Japanese
English	10	0	0
Spanish	0	10	0
Japanese	0	0	10

Solution 4.8

Shuffling the order of the letters does not affect the results of the estimation using this Naive Bayes classifier's prediction. This is because the order information has been omitted when creating the bag-of-words for each training and validation data, i.e. mathematically, when calculating the likelihood.

Solution 5

Solution 5.1

Set z as $z = W_2\sigma(W_1x)$ and $\delta_{i,j} = 1$(when $i = j$), 0 otherwise.

$$\begin{aligned}\frac{\partial L}{\partial W_2} &= \frac{\partial L}{\partial \hat{y}} \frac{\partial g(z)}{\partial z} \frac{\partial z}{\partial W_2} \\ &= -\sum_{i=1}^{k} \frac{y}{\hat{y}} y_i(\delta_{i,j} - y_j)\sigma(W_1x) \\ &= (\hat{y} - y)\sigma(W_1x)\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial W_1} &= \frac{\partial L}{\partial \hat{y}} \frac{\partial g(z)}{\partial z} \frac{\partial z}{\partial \sigma(W_1x)} \frac{\partial \sigma(W_1x)}{\partial W_1x} \frac{\partial W_1x}{\partial W_1} \\ &= -\sum_{i=1}^{k} \frac{y}{\hat{y}} y_i(\delta_{i,j} - y_j)W_2\sigma(W_1x)(1 - \sigma(W_1x)x \\ &= (\hat{y} - y)W_2\sigma(W_1x)(1 - \sigma(W_1x)x\end{aligned}$$

Solution 5.2

The number of epoch(= iteration / batch_size) is 30. The final test error (= 1 - accuracy) is 0.0839. Learning curve by loss and the history of error are below.

Figure 1: Learning curve by my model

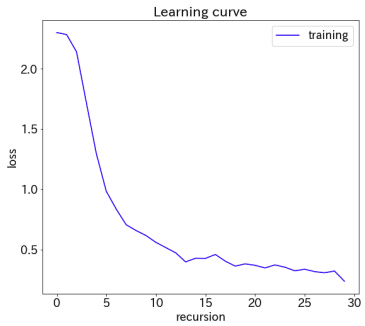
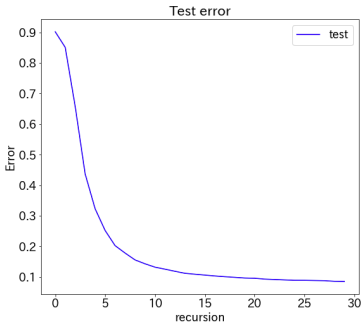


Figure 2: Test error by my model



Solution 5.3

The number of epoch is 30. The final test error is 0.0352. Learning curve by loss and the history of error are below.

Figure 3: Learning curve by pytorch model

Figure 4: Test error by pytorch model

Solution 5.4.a

The number of epoch is 30. The final test error is 0.8865. Learning curve by loss and the history of error are below.

Figure 5: Learning curve by my model (zeros)

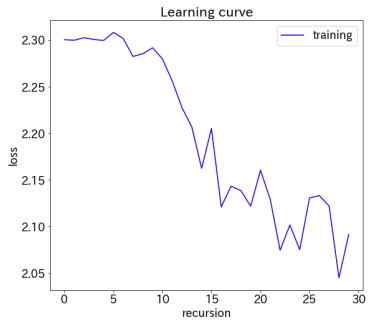
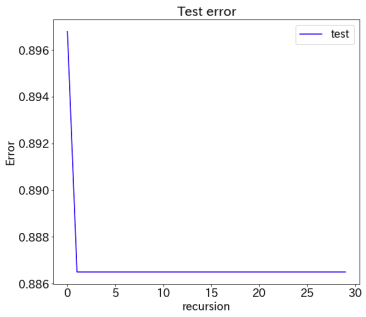


Figure 6: Test error by my model (zeros)



Solution 5.4.b

The number of epoch is 30. The final test error is 0.1102. Learning curve by loss and the history of error are below.

Figure 7: Learning curve by my model (random)

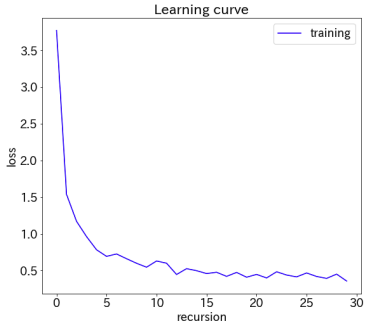


Figure 8: Test error by pytorch model (random)