HW2 SEUNG EUN LEE

1. (a)

K	Accuracy
1	0.751847
5	0.754889
11	0.764885
21	0.746632
41	0.752282
61	0.737505
81	0.726641
101	0.728814
201	0.731421
401	0.719687

1.(b)

K	Accuracy
1	0.856150
5	0.870056
11	0.878748
21	0.884398
41	0.885267
61	0.882660
81	0.877445
101	0.875272
201	0.860061
401	0.839635

1.(c)

	ID	1	5	11	21	41	61	81	101	201	401
0	t1	spam	spam	spam	spam	spam	spam	spam	spam	no	no
1	t2	spam	spam	spam	spam	spam	spam	spam	spam	no	no
2	t3	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
3	t4	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
4	t5	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
5	t6	spam	spam	no	spam	no	no	no	no	spam	spam
6	t7	spam	no	no	no	no	no	no	no	no	no
7	t8	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
8	t9	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
9	t10	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
10	t11	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
11	t12	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
12	t13	spam	spam	spam	spam	spam	spam	spam	no	no	no
13	t14	spam	spam	spam	spam	spam	spam	spam	spam	no	no
14	t15	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
15	t16	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam

16	t17	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
17	t18	spam	spam	spam	spam	spam	spam	spam	spam	spam	no
18	t19	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
19	t20	no	spam	spam	spam	spam	spam	spam	spam	spam	spam
20	t21	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
21	t22	spam	spam	spam	spam	spam	spam	spam	no	no	no
22	t23	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
23	t24	no	spam	spam	spam	spam	spam	spam	spam	spam	spam
24	t25	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
25	t26	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
26	t27	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
27	t28	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
28	t29	spam	spam	spam	spam	spam	spam	spam	spam	no	no
29	t30	spam	spam	spam	spam	no	no	no	no	no	no
30	t31	spam	no	no	no	no	no	no	no	no	no

31	t32	spam	spam	spam	spam	spam	spam	spam	spam	no	no
32	t33	spam	spam	spam	spam	spam	no	no	no	no	no
33	t34	spam	spam	spam	spam	spam	no	no	no	no	no
34	t35	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
35	t36	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
36	t37	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
37	t38	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
38	t39	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
39	t40	no	no	no	no	no	no	no	no	no	no
40	t41	no	no	no	no	no	no	no	no	no	no
41	t42	spam	spam	spam	spam	spam	spam	spam	spam	no	no
42	t43	no	no	no	no	no	no	no	no	no	no
43	t44	no	no	no	no	no	no	no	no	no	no
44	t45	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
45	t46	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
46	t47	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
47	t48	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
48	t49	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam
49	t50	spam	spam	spam	spam	spam	spam	spam	spam	spam	spam

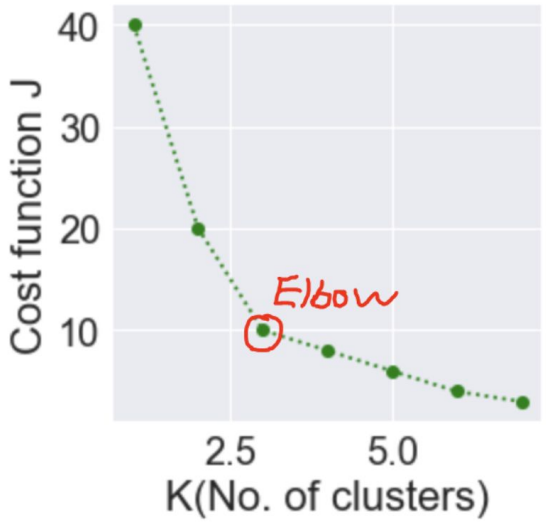
1.(d)

- **Z-score normalization increased the accuracy than the not-normalized one.**
- **For (a) K=11, for (b) K=41 have the best accuracies**

1.(e)

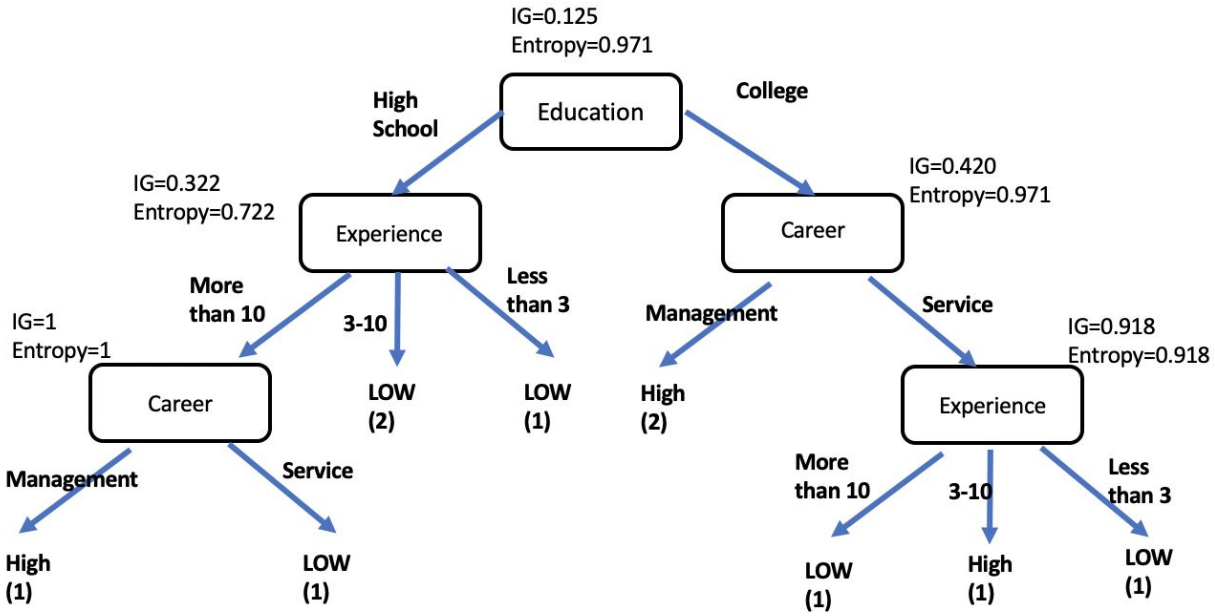
<Elbow method>

: Try various values of K, and if there is a point where the value changes significantly, where a point bends like an elbow bend. This point is judged as the optimal K value.



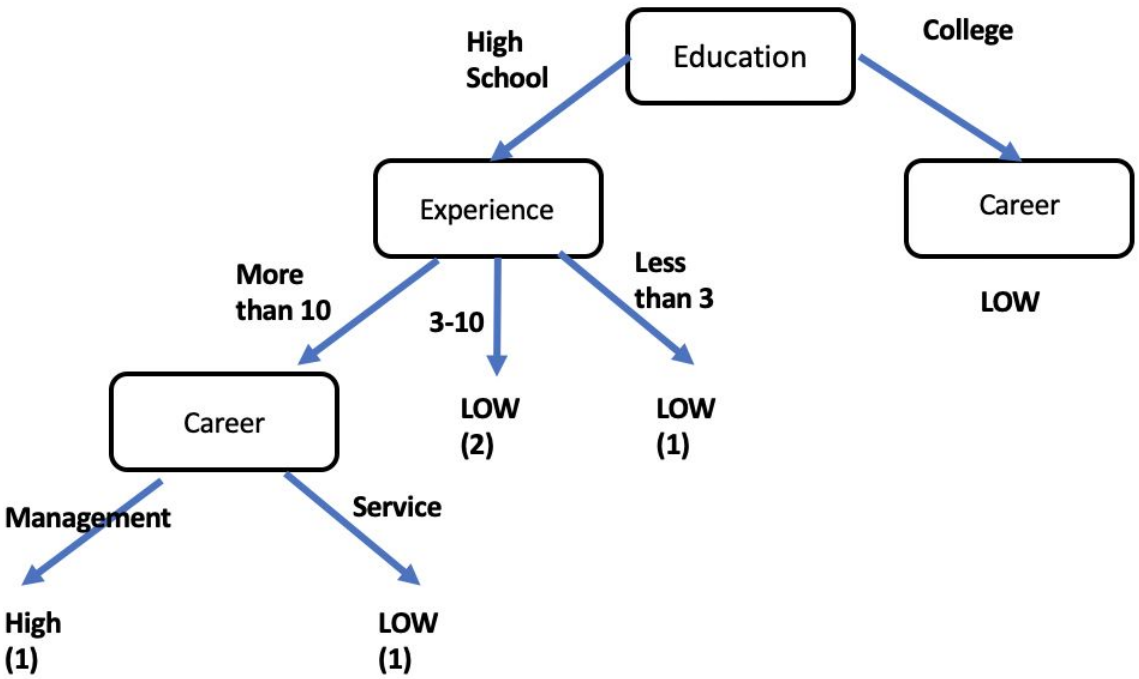
2.

<pre-prune>

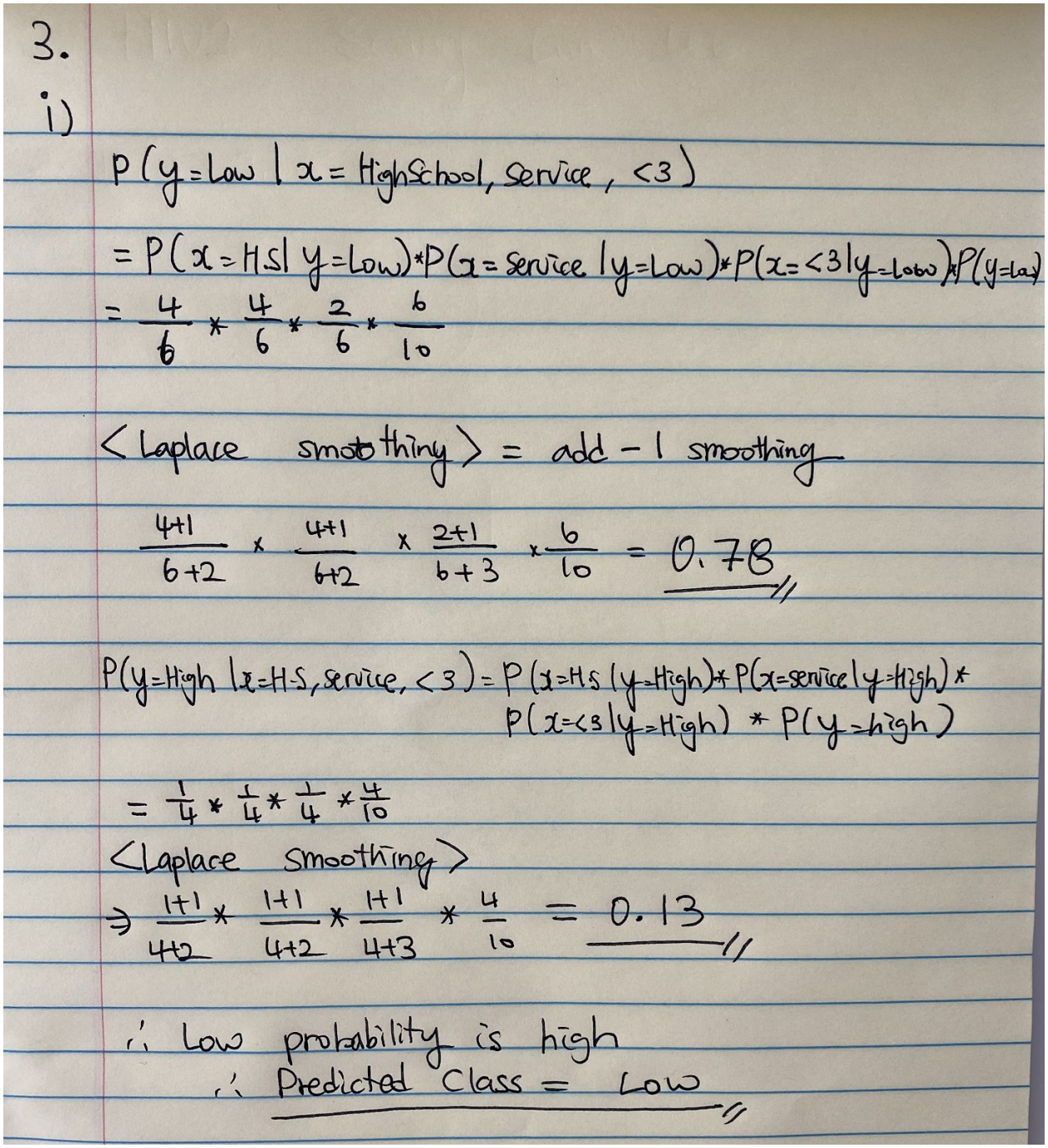


<post-prune>

College >> low


Education
High School
College
Experience
Career
LOW
More than 10
3-10
Less than 3
Career
LOW (2)
LOW (1)
Management
Service
High (1)
LOW (1)

3.


3.
i)
P(y=Low | x= HighSchool, Service, <3)
= P(x=H.S| y=Low)*P(x= service |y=Low)*P(x=<3|y=Low)*P(y=Low)
= 4/6 * 4/6 * 2/6 * 6/10
<Laplace smoothing> = add-1 smoothing
4+1/6+2 x 4+1/6+2 x 2+1/6+3 x 6/10 = 0.78
P(y=High |x=H.S, service, <3) = P(x=H.S |y=High)* P(x=service|y=High)*
P(x=<3|y=High) * P(y=high)
= 1/4 * 1/4 * 1/4 * 4/10
<Laplace Smoothing>
⇒ 1+1/4+2 * 1+1/4+2 * 1+1/4+3 * 4/10 = 0.13
∴ Low probability is high
∴ Predicted Class = Low

3(ii) $P(y=\text{low} \mid X=\text{College, Retail, <3}) = P(x=\text{college} \mid y=\text{low}) *$
$P(x=\text{Retail} \mid y=\text{low}) * P(x<3 \mid y=\text{low}) * P(y=\text{low})$

$$\Rightarrow \frac{2}{6} * \frac{0}{6} * \frac{2}{6} * \frac{6}{10}$$

<Laplace smoothing>

$$\frac{2+1}{6+2} * \frac{0+1}{6+3} * \frac{2+1}{6+3} * \frac{6}{10} = 0.0083$$

$P(y=\text{high} \mid x=\text{college, retail, <3}) = P(x=\text{college} \mid y=\text{high}) * P(x=\text{retail} \mid y=\text{high})$
$* P(x<3 \mid y=\text{high}) * P(y=\text{high})$

$$\Rightarrow \frac{3}{4} * \frac{0}{4} * \frac{1}{4} * \frac{4}{10}$$

<Laplace smoothing>

$$\frac{3+1}{4+2} * \frac{0+1}{4+3} * \frac{1+1}{4+3} * \frac{4}{10} = 0.011$$

$\therefore$ High

3(iii) $P(y=low \mid x=graduate, service, 3-10) = P(x=graduate \mid y=low) *$
$P(x=service \mid y=low) * P(x=3-10 \mid y=low) * P(y=low)$

$= \frac{0}{6} * \frac{4}{6} * \frac{2}{6} * \frac{6}{10}$

<Laplace smoothing>

$\frac{0+1}{6+3} * \frac{4+1}{6+2} * \frac{2+1}{6+3} * \frac{6}{10} = 0.014$ //

$P(y=high \mid x=graduate, service, 3-10)$
$= P(x=graduate \mid y=high) * P(x=service \mid y=high) * P(x=3-10 \mid y=high)$
$* P(y=high)$

$= \frac{0}{4} * \frac{1}{4} * \frac{1}{4} * \frac{4}{10}$

<Laplace smoothing>

$\frac{0+1}{4+3} * \frac{1+1}{4+2} * \frac{1+1}{4+3} * \frac{4}{10} = 0.0054$ //

$\therefore$ low //

4.

4. 25 students ⟶ 60% accuracy.

(a) take 3 models, M.V.C C_3

C_3's accuracy

error ⇒ 0.4 → p

$$\sum_{k>n/2}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

$$\sum_{k>1.5}^{3} \binom{3}{k} (0.4)^k (1-0.4)^{3-k}$$

$$\Rightarrow \sum_{k=2}^{3} \binom{3}{k} (0.4)^k (0.6)^{3-k}$$

$$\Rightarrow \binom{3}{2} 0.4^2\, 0.6^1 + \binom{3}{3} (0.4)^3 (0.6)^0$$

$$\Rightarrow 3(2\ 0.4^2\, 0.6^1 + 3(3\ 0.4^3 (0.6)^0$$

$$\Rightarrow 3 \times (0.4)^2 (0.6) + (0.4)^3 (0.6)^0$$

$$= 0.352$$

∴ accuracy = 0.648

(b) $$\sum_{k>2.5}^{5} \binom{5}{k} (0.4)^k (0.6)^{5-k}$$

$$\Rightarrow \sum_{3}^{5} \binom{5}{k} (0.4)^k (0.6)^{5-k}$$

$$= 5(3\ (0.4)^3 (0.6)^2 + 5(4\ (0.4)^4 (0.6)^1 + 5(5\ (0.4)^5 (0.6)^0$$

$$= 0.317$$

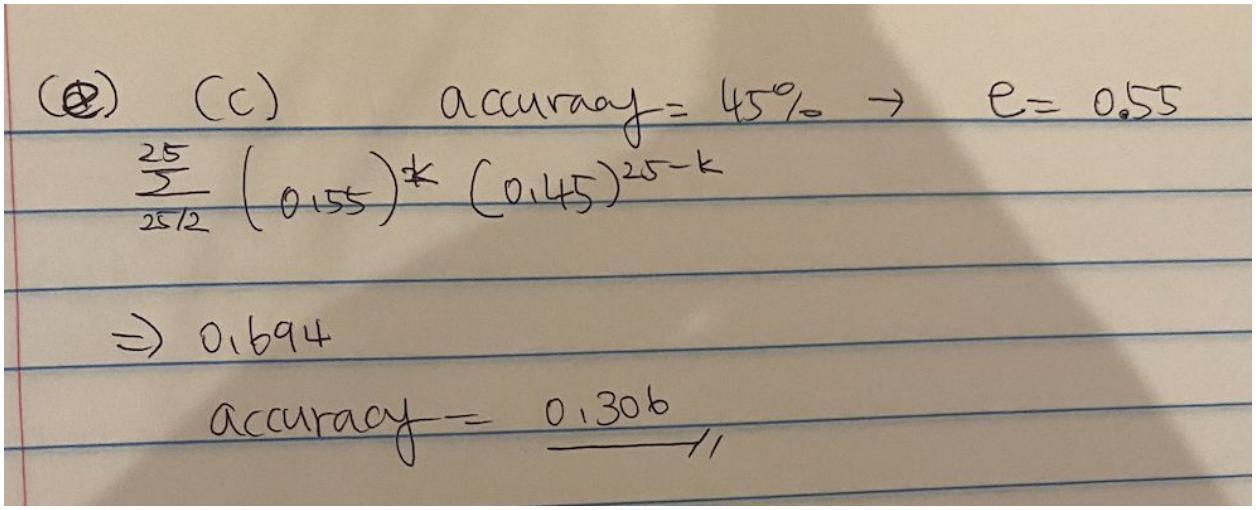
∴ accuracy = 0.683

(c) $$\sum_{k>25/2}^{25} \binom{25}{k} (0.4)^k (0.6)^{25-k} = 0.154$$

∴ accuracy = 0.846

(d)
Assumption of every classifier has the same probability will cause difference from the reality. Another assumption is that each classifier must be independent of each other.

(e)



(simple program to calculate the error probability)

```
from scipy.special import comb
import math

def ensemble_error(n_classifier, error):
    k_start = int(math.ceil(n_classifier / 2.))
    probs = [comb(n_classifier, k) * error**k * (1-error)**(n_classifier - k)
             for k in range(k_start, n_classifier + 1)]
    return sum(probs)

x=ensemble_error(n_classifier=25, error=0.55)
print(x)
```