6330300119
6330300038

1.)

ก. gini index = $1-\left(\left(\frac{10}{20}\right)^2+\left(\frac{10}{20}\right)^2\right)=\mathbf{\frac{1}{2}}$ Entropy = $-\left(\frac{5}{10}\right)\log_2(\frac{5}{10})-\left(\frac{5}{10}\right)\log_2(\frac{5}{10})=1$

ข. 1 – 20

Gini index = $1-(0+1^2)=0$ Entropy = $-0\log(0)-1\log(1)=0$

Information gain = $1-0=1$

ค. M

Gini index = $1-\left(\left(\frac{6}{10}\right)^2+\left(\frac{4}{10}\right)^2\right)=0.52$ Entropy = $-\left(\frac{6}{10}\right)\log_2(\frac{6}{10})-\left(\frac{4}{10}\right)\log_2(\frac{4}{10})=0.97$

Information gain = $1-0.97=0.03$

F

Gini index = $1-\left(\left(\frac{4}{10}\right)^2+\left(\frac{6}{10}\right)^2\right)=0.52$ Entropy = $-\left(\frac{4}{10}\right)\log_2(\frac{4}{10})-\left(\frac{6}{10}\right)\log_2(\frac{6}{10})=0.97$

Information gain = $1-0.97=0.03$

ง. Family

Gini index = $1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right)=0.625$ Entropy = $-\left(\frac{1}{4}\right)\log_2(\frac{1}{4})-\left(\frac{3}{4}\right)\log_2(\frac{3}{4})=0.81$

Information gain = $1-0.81=0.19$

Sport

Gini index = $1-\left(\left(\frac{8}{8}\right)^2+\left(\frac{0}{0}\right)^2\right)=0$ Entropy = $-\left(\frac{8}{8}\right)\log_2(\frac{8}{8})-\left(\frac{0}{0}\right)\log_2(\frac{0}{0})=0$

Information gain = $1-0=1$

Luxury

Gini index = $1-\left(\left(\frac{7}{8}\right)^2+\left(\frac{1}{8}\right)^2\right)=0.78$ Entropy = $-\left(\frac{7}{8}\right)\log_2(\frac{7}{8})-\left(\frac{1}{8}\right)\log_2(\frac{1}{8})=0.54$

Information gain = $1-0.54=0.46$

จ. Small

Gini index = $1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)=0.52$ Entropy = $-\left(\frac{3}{5}\right)\log_2(\frac{3}{5})-\left(\frac{2}{5}\right)\log_2(\frac{2}{5})=0.97$

Information gain = $1-0.97=0.03$

Medium

Gini index = $1-\left(\left(\frac{3}{7}\right)^2+\left(\frac{4}{7}\right)^2\right)=0.51$ Entropy = $-\left(\frac{3}{7}\right)\log_2(\frac{3}{7})-\left(\frac{4}{7}\right)\log_2(\frac{4}{7})=0.985$

Information gain = $1-0.985=0.015$

Large

Gini index = $1-\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)=0.5$ Entropy = $-\left(\frac{2}{4}\right)\log_2(\frac{2}{4})-\left(\frac{2}{4}\right)\log_2(\frac{2}{4})=1$

Information gain = $1-1=0$

Extra large

Gini index = $1-\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)=0.5$ Entropy = $-\left(\frac{2}{4}\right)\log_2(\frac{2}{4})-\left(\frac{2}{4}\right)\log_2(\frac{2}{4})=1$

Information gain = $1-1=0$

ฉ. Customer ID

ช. เหมาะ เพราะ มีความบริสุทธิ์มากที่สุด

6330300119
6330300038

2. จงอธิบายความหมายของ underfitting และ overfitting

ตอบ **Model underfitting** เกิดเมื่อโมเดลมีความซับซ้อนไม่พอที่จะเรียนรู้รูปแบบที่แฝงอยู่ในข้อมูลได้ สามารถสังเกตได้จากการที่ทั้ง training error และ test error มีค่าสูง

Model overfitting เกิดเมื่อโมเดลมีความซับซ้อนมากเกินไปทำให้เรียนรู้รูปแบบปลอม (spurious patterns) แทนที่จะเป็นรูปแบบที่แท้จริงตามธรรมชาติของข้อมูล สามารถสังเกตได้จากการที่ training error มีค่าต่ำหรือลดลงในขณะที่ test error มีค่าสูงหรือเพิ่มขึ้น

3. กำหนดผลการทดสอบประสิทธิภาพของโมเดลการจำแนกประเภทแบบสองคลาสดัง confusion matrix ต่อไปนี้ จงประเมินประสิทธิภาพของโมเดลดังกล่าวโดยคำนวณหาค่า Accuracy, Error rate, Recall, Precision, Specificity, False Positivie Rate และ F1-score

Model 1		Predicted Class	
		Positive Class	Negative Class
Actual Class	Positive Class	180	20
	Negative Class	60	340

	Accuracy	Error rate	Recall	Precision	Specificity	FPR	F1-Score
Model 1	$\frac{180+340}{180+20+60+340}$	$\frac{20+60}{180+20+60+340}$	$\frac{180}{180+20}$	$\frac{180}{180+60}$	$\frac{340}{60+340}$	$\frac{60}{60+340}$	$\frac{2(0.75)(0.9)}{0.75+0.9}$
	0.866	0.133	0.9	0.75	0.85	0.15	0.818

6330300119
6330300038

4. Cost(internal) = log m Cost(leaf) = log 3

Tree1 Tree2

Cost(tree1) = 2*log m + 3*log 3 Cost(tree2) = 4*log m + 5*log 3

= log m^2 + log 3^3 = log $m^4 3^5$

= log $m^2 3^3$

Cost(D|tree1) = 7*log n Cost(D|tree2) = 4*log n

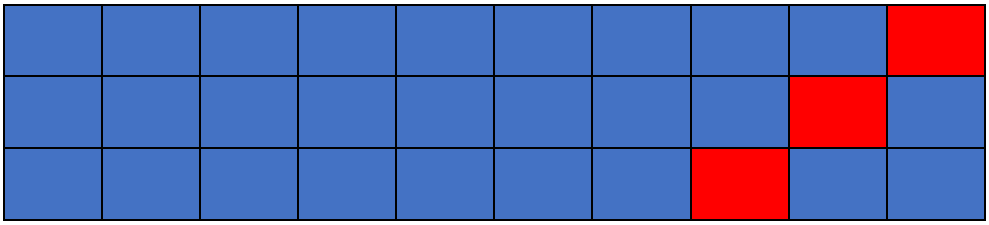
Cost(tree1,D) = log $m^2 3^3$ + 7*log n Cost(tree2,D) = log $m^4 3^5$ + 4*log n

= log $m^2 3^3 n^7$ = log $m^4 3^5 n^4$

Cost(tree1,D) – Cost(tree2,D)$= \log m^2 3^3 n^7 - \log m^4 3^5 n^4 = \log m^{-2} 3^{-2} n^3 = 3\log n - 2\log 3 - 2\log m$

∴ ควรใช้ tree2 เมื่อ 2n >= 3 + m และควรใช้ tree1 เมื่อ 2n <= 3 + m

5. แบ่งข้อมูลออกเป็น 10 ส่วน แล้วทำการหาความแม่นยำ 10 รอบโดยแต่ละรอบจะใช้ test set 1 ส่วนที่ไม่ซ้ำกัน

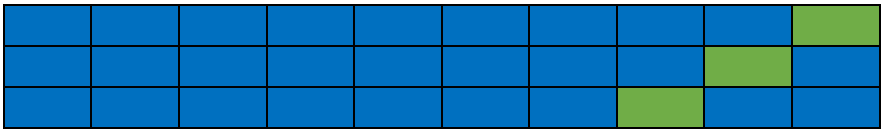


.

.

Train set Test set

ค่าความแม่นยำจะใช้ค่าเฉลี่ยจากทั้ง 10 รอบ และ train set (กล่องสีฟ้า) จะมีการคัดเลือก model จากการแบ่งเป็น 10 ส่วน โดยใช้ 1 ส่วนที่ไม่ซ้ำกัน 10 รอบ เป็น validation set เพื่อหาค่า error และเลือก model ที่มีค่า error รวมต่ำที่สุด



.

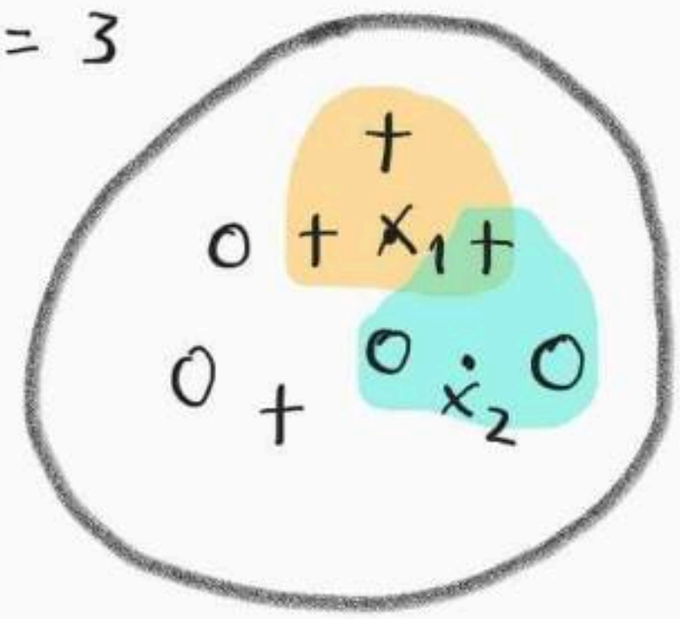
.

Train set validation set

6330300119
6330300038

K-NN แบ่งกลุ่ม

1. Majority voting แบ่งแบบประชาธิปไตยใครเยอะกว่าชนะ



- ดูรอบ x_1 3 ตัว
  จะเห็นว่า + อยู่รอบx_1 3 ตัว
  กลุ่ม x_1 เป็น +

* ถ้า x_1 มี O 2 ตัว + 2 ตัว
  จะแบ่งกลุ่มไม่ได้

- ดูรอบ x_2 3 ตัว
  จะเห็นว่า O อยู่รอบx_2 2 ตัว กับ + 1 ตัว
  กลุ่ม x_2 เป็น O

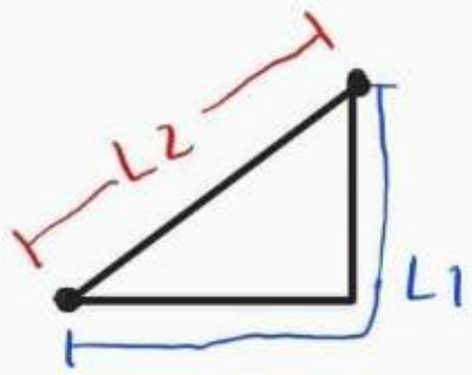
2. Weighted distance พวกระยะทางกลุ่มใกล้ใกล้จุดกึ่งกลาง
   มี 3 แบบ คือ Euclidean, Manhattan และ Cosine

- Euclidean distance

$$L2 = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{\frac{1}{2}}$$

- Manhattan

$$L1 = \sum_{i=1}^{n}|x_i - y_i|$$



แล้วแต่จะเลือกใช้

- Cosine Similarity เหมาะกับเอกสาร, text, seach engine วัดความคล้ายกัน

$$dist(x,y) = cos(x,y) = \frac{x \cdot y}{|x||y|}$$

การกำหนด ค่า K ต้องลองทดลองเอาว่า K ที่เท่าไรดีที่สุด

Gini Index ,Entropy และ Misclassification Error ต่างกันอย่างไร

Gini Index ,Entropy และ Misclassification Error เป็นการหาความบริสุทธิ์ของข้อมูลโดยจะมีผลลัพธ์ที่ใกล้เคียงกันคือ มีค่าเป็น 0 เมื่อข้อมูลมีเพียงคลาสเดียว และมีค่ามากที่สุดเมื่อ คลาสทุกคลาสมีปริมาณที่เท่าๆ กัน โดยค่าที่ได้จากทั้ง 3 แบบสามารถเรียงค่าจากผลลัพธ์ ดังนี้ Entropy >= Gini Index >= Misclassification Error