Tree-based Methods

Petchara Pattarakijwanich

Introduction to Data Science, 4 November 2022

Goal of this week

- Binary Tree
  - Regression Tree
  - Classification Tree
- Purity Metrics
  - Classification Error Rate
  - Gini Index
  - Entropy
- Methods to Improve Binary Tree
  - Bagging
  - Random Forest
  - Boosting

About the Final Project

Approaches

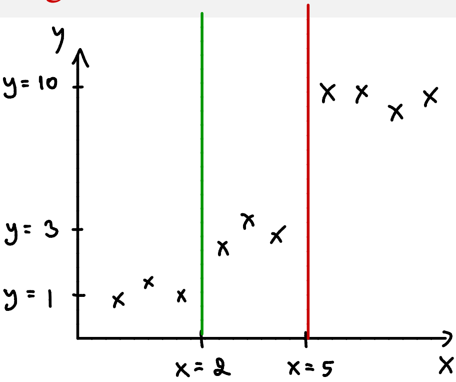
- Answer specific questions from data, using data science techniques
- Implement, invent, or test some data science techniques or algorithms

Data Source

- Your own research
- Publicly available datasets
  https://geekflare.com/open-datasets-for-data-science/
  https://www.dataquest.io/blog/free-datasets-for-projects/
- Simulated data (?)

Presentation (some time around final week or a bit later)

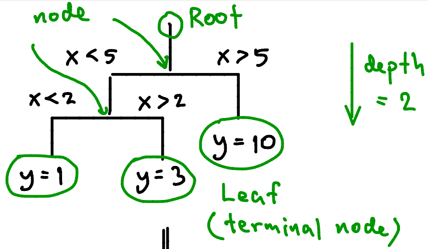
Regression Tree

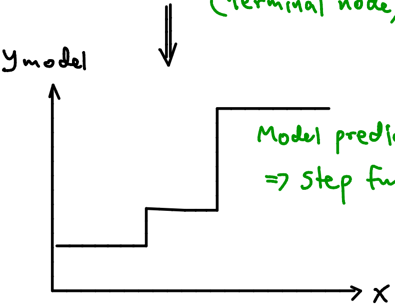


ตัดที่ไหน ?

=> ตัดแล้ว “ดี” ที่สุด (minimize MSE)

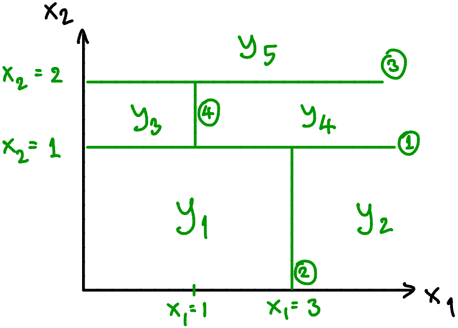
Binary Tree

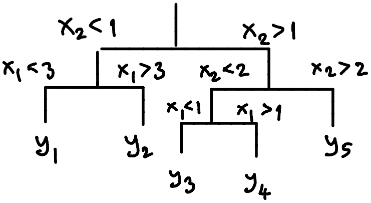




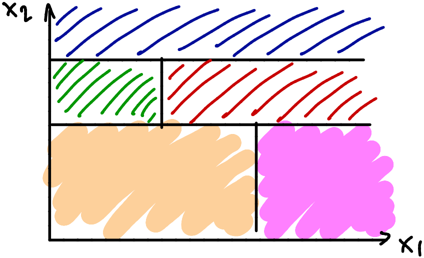
Model prediction => step function

Regression Tree กรณี 2 มิติ





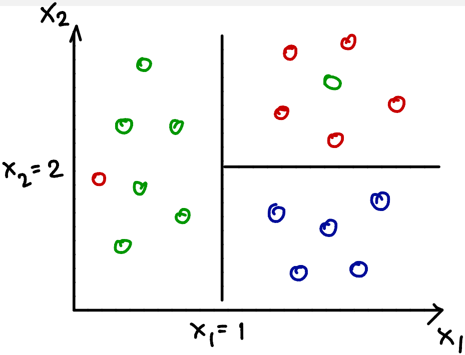
y_i = ค่าเฉลี่ยของจุดที่อยู่
ในช่วงนั้นๆ

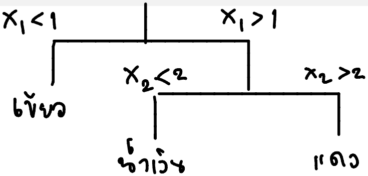


Model Prediction

Step function ที่ซับซ้อนขึ้น

Classification Tree





คำทำนายของแต่ละ node
= mode ของ node นั้น

ตัดที่ตัวแปรไหน & ค่าไหน ? ตัดยังไงถึง "ดี"

Purity metric

- Error Rate ← ไม่ค่อยใช้
- Gini Index
- Entropy

} ใช้บ่อย

How to Grow Trees

Input $\{x_1, x_2, \dots, x_n\}$ output $\{y\}$

- ตัวเลข → regression
- กลุ่ม → classification

① เริ่มจาก Root

② หาว่าตัดที่ตัวแปรไหน & ค่าเท่าไหร่

- ลองตัด x_1 → หาค่าที่ตัดดีสุด
- ลองตัด x_2 → หาค่าที่ตัดดีสุด

⋮

- สรุปได้ว่าตัดที่ x_i ที่ค่า $x_i = ?$

ทำซ้ำ จนพอใจ

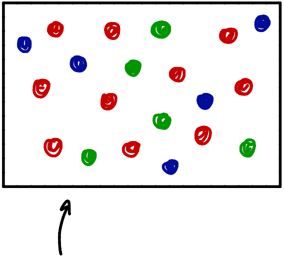
ดีสุด

- Regression ⇒ MSE ลดลง มากสุด
- Classification ⇒ Purity เพิ่มขึ้น มากสุด

③ หยุดเมื่อไหร่ ?

- ไม่หยุดตัด ⇒ หนึ่ง node มี 1 จุดข้อมูล ⇒ overfit
- หยุดเมื่อความ "ดี" เพิ่มไม่มาก
- หยุดที่ depth ที่ต้องการ

Purity Metrics



node นึงของ tree

แดง 10

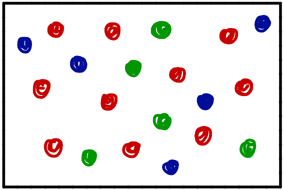
เขียว 5

น้ำเงิน 5

วัดความ "pure" ของกล่องได้ยังไง?

- Error Rate
- Gini Index
- Shannon Entropy

Classification Error Rate



$$\boxed{E = 1 - P_{k,max}}$$

ทำนายว่าทุกอย่างเป็น mode

⇒ ทายผิดกี่เปอร์เซ็นต์

แดง	10
เขียว	5
น้ำเงิน	5

⇒

$$P_{\text{แดง}} = \frac{1}{2}$$

$$P_{\text{เขียว}} = \frac{1}{4}$$

$$P_{\text{น้ำเงิน}} = \frac{1}{4}$$

ทายว่าทุกอย่าง = แดง

$$\Rightarrow \text{ทายถูก} = \frac{10}{20}$$

$$\text{ทายผิด} = \frac{10}{20} = \boxed{0.5}$$

↑

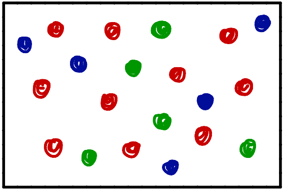
Error Rate

$$P_k = \begin{cases} P_{\text{แดง}} \\ P_{\text{เขียว}} \\ P_{\text{น้ำเงิน}} \end{cases}$$

$$P_{k,max} = P_{\text{แดง}} = 0.5$$

$$E = 1 - P_{k,max} = \boxed{0.5}$$

Gini Index



$$\boxed{G = \sum_k P_k(1-P_k) = 1 - \sum_k P_k^2}$$

สุ่มจุด 1 จุด + สุ่มสีมา 1 สี

⇒ โอกาสผิดเยอะกี่เปอร์เซ็นต์

$P_{\text{แดง}} = \frac{1}{2}$

$P_{\text{เขียว}} = \frac{1}{4}$

$P_{\text{น้ำเงิน}} = \frac{1}{4}$

$$G = P_{\text{แดง}}(1-P_{\text{แดง}}) + P_{\text{เขียว}}(1-P_{\text{เขียว}}) + P_{\text{น้ำเงิน}}(1-P_{\text{น้ำเงิน}})$$

↑ prob ที่จะได้สีนั้นๆ

↖ prob ที่จะได้ label คนละสีกัน

Gini Index

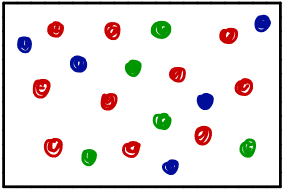
$$G = P_{\text{แมว}}(1 - P_{\text{แมว}}) + P_{\text{เสือ}}(1 - P_{\text{เสือ}}) + P_{\text{ม้าลาย}}(1 - P_{\text{ม้าลาย}})$$

$$= P_{\text{แมว}} - P_{\text{แมว}}^2 + P_{\text{เสือ}} - P_{\text{เสือ}}^2 + P_{\text{ม้าลาย}} - P_{\text{ม้าลาย}}^2$$

$$= 1 - P_{\text{แมว}}^2 - P_{\text{เสือ}}^2 - P_{\text{ม้าลาย}}^2$$

$$G = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \dots$$

Shannon Entropy



$P_{\text{แดง}} = \frac{1}{2}$

$P_{\text{เขียว}} = \frac{1}{4}$

$P_{\text{น้ำเงิน}} = \frac{1}{4}$

$$\boxed{D = -\sum_k P_k \log_2 P_k}$$

"Information" ของ Event ที่มี prob p

$$I = \log_2\left(\frac{1}{p}\right) = -\log_2 p$$

"Information" p น้อย → I มาก

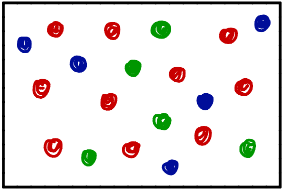
$$I_A + I_B = \log_2\left(\frac{1}{P_A}\right) + \log_2\left(\frac{1}{P_B}\right)$$

$$= \log_2\left(\frac{1}{P_A P_B}\right)$$

$$= \log_2\left(\frac{1}{P(A \cup B)}\right)$$

Information ของ 2 event

Shannon Entropy



$P_{\text{แดง}} = \frac{1}{2}$

$P_{\text{เขียว}} = \frac{1}{4}$

$P_{\text{น้ำเงิน}} = \frac{1}{4}$

Entropy = Expected value ของ Informatn

$$= \langle I \rangle$$

$$= \sum_k P_k I_k$$

$$= \sum_k P_k \log_2 \left(\frac{1}{P_k}\right)$$

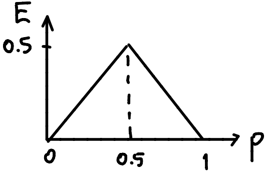
$$= - \sum_k P_k \log_2 P_k$$

$$D = - P_{\text{แดง}} \log_2 P_{\text{แดง}} - P_{\text{เขียว}} \log_2 P_{\text{เขียว}} - P_{\text{น้ำเงิน}} \log_2 P_{\text{น้ำเงิน}} = \ldots$$

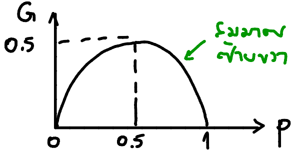
Purity Metrics

สมมติมี 2 class $\begin{cases} + & , \text{prob } p \\ - & , \text{prob } 1-p \end{cases}$

$$E = 1 - p_{k,\max} = \begin{cases} 1-p, & p \geq 0.5 \\ p, & p < 0.5 \end{cases}$$

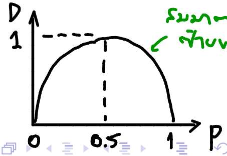


$$G = 1 - \sum_k p_k^2 = 1 - p^2 - (1-p)^2$$



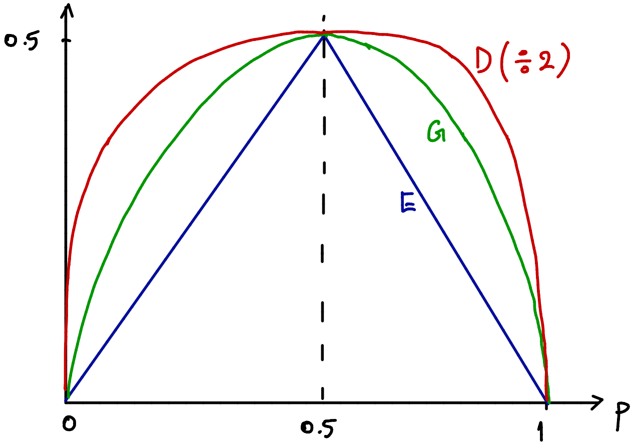
สมมาตร
ซ้ายขวา

$$D = -\sum_k p_k \log_2 p_k = -p \log_2 p - (1-p)\log_2(1-p)$$

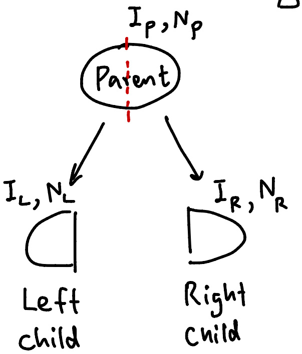


สมมาตร
ซ้ายขวา

Purity Metrics



Information Gain



$$\Delta I = I_p - \underbrace{\frac{N_L}{N_p} I_L - \frac{N_R}{N_p} I_R}_{\text{ค่าเฉลี่ยถ่วงน้ำหนักของ } I_L,\, I_R}$$

$\Delta I > 0$ ⇒ แบ่งแล้ว ปนกันน้อยลง

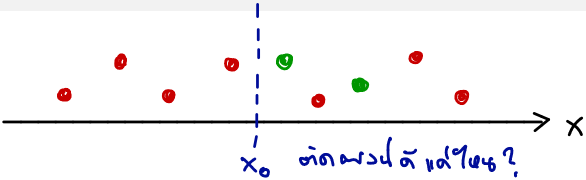
⇒ Purity สูงขึ้น

ΔI ยิ่งมากยิ่งดี

แบ่งที่ตัวแปรไหน? ค่าเท่าไหร่? ⇒ ให้ ΔI สูงสุด

Information Gain

ตัวอย่าง



ตัดตรงนี้ ดีแค่ไหน?

Parent $N_p = 9$, $P_{\text{แดง}} = \frac{7}{9}$, $P_{\text{เขียว}} = \frac{2}{9}$

$$G_p = 1 - \sum_k P_k^2 = 1 - P_{\text{แดง}}^2 - P_{\text{เขียว}}^2 = 1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2 = 0.35$$

Left child $N_L = 4$, $P_{\text{แดง}} = 1$, $P_{\text{เขียว}} = 0$

$$G_L = 1 - P_{\text{แดง}}^2 - P_{\text{เขียว}}^2 = 0$$

Right child $N_R = 5$, $P_{\text{แดง}} = \frac{3}{5}$, $P_{\text{เขียว}} = \frac{2}{5}$

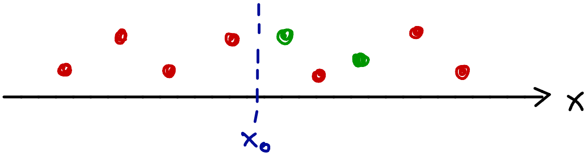
$$G_R = 1 - P_{\text{แดง}}^2 - P_{\text{เขียว}}^2 = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

Information Gain

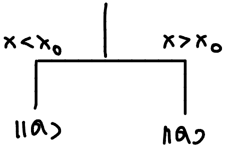
$$\Delta G = G_p - \frac{N_L}{N_p} G_L - \frac{N_R}{N_p} G_R$$

$$= 0.35 - \frac{4}{9} \times 0 - \frac{5}{9} \times 0.48$$

$$= 0.08 > 0$$ ⇒ แบ่ง ดีกว่าไม่แบ่ง



⇒



แบ่งทำไม

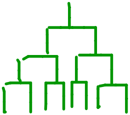
- เพิ่ม node purity
- แบ่งเพื่อ node ย่อย ต่อได้

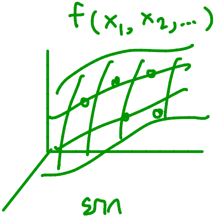
Pros and Cons of Classification Tree

ข้อดี

- ใช้กับ Qualitative variable ง่าย
- เข้าใจง่าย อธิบายง่าย

(ใกล้เคียงกับ กระบวนการ ตัดสินใจ ของมนุษย์?)





ข้อเสีย

- Variance สูง
- Prediction Accuracy ต่ำ

} over fit ง่าย

- แก้โดย Bagging
  Random Forest
  Boosting