HW4 电计1701吴瑞林 201785072

No.

Date . .

(a) ID3

total : 14, + : 9 − : 5

$S(9+, 5-)$ $P(c_1) = 9/14$ $P(c_2) = 5/14$

$\text{Entropy}(S) = -\sum_{i} P(w_i) \log_2 P(w_i)$

$= -(\frac{9}{14} \times \log_2 \frac{9}{14} + \frac{5}{14} \times \log_2 \frac{5}{14}) \approx 0.94$

(I) root:

① outlook: Sunny: 5^{2+}_{3-} Overcast: 4+, Rain: 5^{3+}_{2-}

$E(S) = -(2/5 \times \log_2 2/5 + \frac{3}{5} \log_5 3/5) = 0.971$

$E(O) = -(1 \times \log 1) = 0$

$E(R) = -(3/5 \times \log_2 3/5 + 2/5 \times \log_2 2/5) = 0.971$

$G(S, \text{outlook}) = E(S) - \frac{5}{14} E(\text{sunny}) - \frac{4}{14} E(\text{overcast}) -$

$5/14 \times E(\text{Rain}) = 0.94 - \frac{5}{7} \times 0.971 = 0.246$

② temp : hot: 4^{2+}_{2-} mid: 6^{4+}_{2-} cool: 4^{3+}_{1-}

$E(h) = -(\frac{2}{4} \log \frac{2}{4} + \frac{2}{4} \log \frac{2}{4}) = 1$

$E(m) = -(\frac{4}{6} \log_2 \frac{4}{6} + \frac{2}{6} \times \log_2 \frac{2}{6}) = 0.918$

$E(c) = -(\frac{3}{4} \times \log_2 \frac{3}{4} + \frac{1}{4} \log_2 \frac{1}{4}) = 0.811$

$G(S, \text{temp}) = 0.94 - \frac{4}{14} \times 1 - \frac{6}{14} \times 0.918 - \frac{4}{14} \times 0.81$

$= 0.029$

③ Humidity : H: 7^{3+}_{4-} N: 7^{6+}_{1-}

$E(H) = -(3/7 \log 3/7 + 4/7 \log 4/7) = 0.985$

$E(N) = -(6/7 \log 6/7 + 1/7 \log 1/7) = 0.592$

$G(S, H) = 0.94 - 7/14 \times 0.985 - 7/14 \times 0.592 = 0.151$

④ wind: W:8$^{6+}_{2-}$ S:6$^{3+}_{3-}$

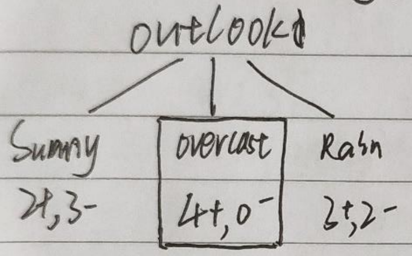
E(w) = $-(6/8 \log 6/8 + 2/8 \log 2/8) = 0.811$

E(S) = $-(3/6 \log 3/6 + 3/6 \log 3/6) = 1$

G(S,w) = $0.94 - 8/14 \times 0.811 - 6/14 \times 1 = 0.048$

$\because$

G(S, O) is the biggest, outlook is root



(II) Next Node IN sunny

E(sunny) = 0.971

① Temp: H: 2-, M: 2$^{1+}_{1-}$, C: 1+

E(H) = 0, E(C) = 0, E(M) = 1

G(S, temp) = $0.971 - 2/5 \times 1 = 0.571$

② Humidity: H: 3-, N: 2+

E(H) = 0, E(N) = 0

G(S, H) = $0.971 - 0 = 0.971$

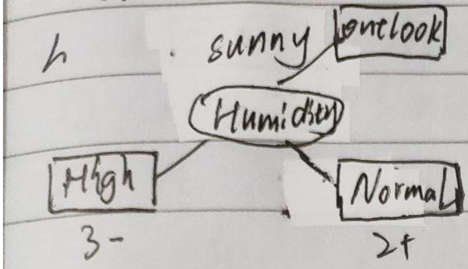
③ Wind: W: 3$^{1+}_{2-}$, S: 2$^{1+}_{1-}$

E(w) = $-(1/3 \log_2 1/3 + \frac{2}{3} \log_2 2/3) = 0.918$, E(S) = 1

G(S, w) = $0.971 - 3/5 \times 0.918 - 2/5 \times 1 = 0.0202$

∵ G(sunny, Humidity) is the larggest,

∴



(III), next node in Rain

$E(Rain) = 0.971$

① temp: $M: 3^{2+}_{1-}$ $C: 2^{1+}_{1-}$

$E(mild) = -(2/3 \log_2 2/3 + 1/3 \log_2 1/3) = 0.918$, $E(C) = 1$

$G = 0.971 - \frac{2}{5} - 3/5 \times 0.918 = 0.0202$

② Humidity: $H\,2^{1+}_{1-}$ $N: 3^{2+}_{1-}$

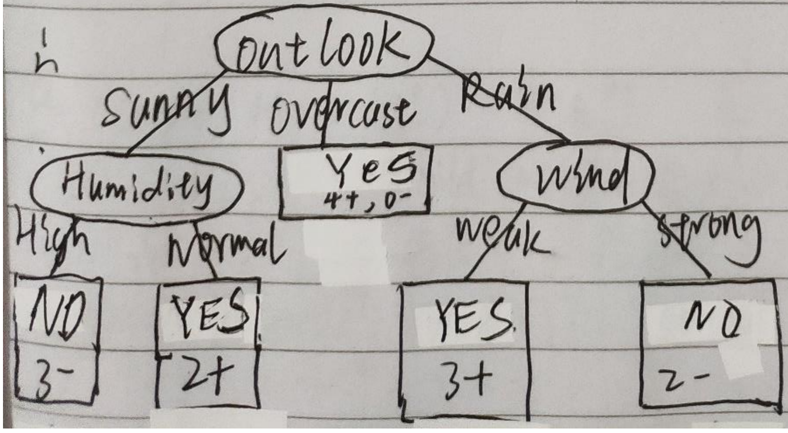
$E(H) = 1$, $E(N) = 0.918$

$G = 0.971 - 2/5 - 3/5 \times 0.918 = 0.0202$

③ wind: W: 3+, S: 2-

$E(W) = 0$, $E(S) = 0$, $G = 0.971$, (biggest)

∴



(b) CART.

No.

Date

(I) root, D is the sample

A. outlook sunny: $5\,^{2+}_{3-}$ overcast: 4^{+} Rain: $5\,^{3+}_{2-}$

① {Sunny} | {overcast, Rain}

A_1 (2+,3-) A_2 (7+,2-)

$\text{Gini}(A_1) = 1-(\frac{2}{5})^2-(\frac{3}{5})^2 = \frac{12}{25}$

$\text{Gini}(A_2) = 1-(\frac{7}{9})^2-(\frac{2}{9})^2 = \frac{28}{81}$

$\text{Gini}(D,\{A_1,A_2\}) = \frac{5}{14}\times \text{Gini}(A_1) + \frac{9}{14}\times \text{Gini}(A_2)$

$= \frac{5}{14}\times\frac{12}{25} + \frac{9}{14}\times\frac{28}{81} = 0.394$

② {overcast} | {Sunny, Rain}, $A_3(4^+)$ | $A_4(5^+5^-)$

$G(A_3) = 0$, $G(A_4) = 1-0.5^2-0.5^2 = 0.5$

$G(D,\{A_3,A_4\}) = \frac{1}{2}\times\frac{10}{14} = 0.357$

③ {Rain} | {sunny, overcast}, A5|A6

$G(A5) = 1-(\frac{3}{5})^2-(\frac{2}{5})^2 = 12/25$

$G(A6) = 1-(6/9)^2-(3/9)^2 = \frac{4}{9}$

$G(D,\{A5.A6\}) = 5/14\times\frac{12}{25} + 9/14\times\frac{4}{9} = 16/35 = 0.457$

B. temp: $H=4\,(^{2+}_{2-})$ $M=6\,^{4+}_{2-}$ $C: 4\,^{3+}_{1-}$

① $\text{Gini}(D,\{\text{Hot}\}|\{\text{mild},\text{cool}\}) = \frac{4}{14}\times(1-\frac{1}{2}^2-\frac{1}{2}^2) + \frac{10}{14}\times$
$(1-\frac{7}{10}^2-\frac{3}{10}^2) = \frac{4}{14}\times\frac{1}{2} + \frac{10}{14}\times\frac{21}{50} = 0.443$

② $\text{Gini}(D,\{\text{mild}\}|\{\text{Hot},\text{cool}\}) = (1-(\frac{4}{6})^2-(\frac{2}{6})^2)\times\frac{6}{14} + \frac{8}{14}\times(1-$
$(\frac{5}{8})^2-(\frac{3}{8})^2) = \frac{6}{14}\times\frac{4}{9} + \frac{8}{14}\times\frac{15}{32} = \frac{11}{24} = 0.458$

③ $\text{Gini}(D,\{\text{cool}\}|\{\text{Hot},\text{mild}\}) = \frac{4}{14}\times(1-(\frac{3}{4})^2-(\frac{1}{4})^2) + \frac{10}{14}\times(1-(\frac{6}{10})^2-(\frac{4}{10})^2)$

$= \frac{4}{14}\times\frac{3}{8} + \frac{10}{14}\times\frac{12}{25} = \frac{9}{20} = 0.45$

C. Humidity: $H:7(^{3+}_{4-})$ $N:7(^{6+}_{1-})$

$G(D, \{Gini(H), Gini(N)\}) = \frac{1}{2} \times (1-(\frac{3}{7})^2-(\frac{4}{7})^2) + \frac{1}{2} \times$

$(1-(\frac{6}{7})^2-(\frac{1}{7})^2) = \frac{24+12}{49 \times 2} = \frac{18}{49} = 0.367$

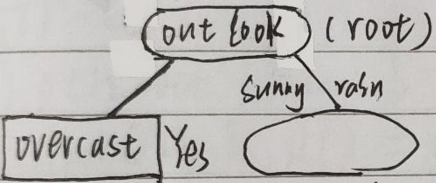
D. Wind: $W: 8^{6+}_{2-}$ $S: 6^{3+}_{3-}$

$Gini(D, \{Gini(W), Gini(S)\}) = \frac{8}{14}(1-(\frac{6}{8})^2-(\frac{2}{8})^2 + \frac{6}{14}(1-(\frac{3}{6})^2-(\frac{3}{6})^2)$

$= 8/14 \times 3/8 + 6/14 \times 1/2 = 0.2429$

1. $Gini(D; \{overcast\}|\{Sunny; Rain\})$ has the smallest

2. Gini : 0.357



4+ $D = D\{1,2,4,5,6,8,9,10,13,14\}$

(II) second node

Ⓐ temp: Hot=2−, mid: 5^{3+}_{2-}, cool: 3^{2+}_{1-}

① $Gini(D, \{G(H) | G(mild, coo)\}) = \frac{2}{10} \times 0 + \frac{8}{10} \times \frac{15}{32} = 0.375$

② $Gini(D, \{G(m) | G(c, H)\}) = \frac{5}{10} \times (\frac{12}{25}) + \frac{5}{10} \times \frac{12}{25} = 0.48$

③ $Gini(D, \{G(c) | G(m,H)\}) = \frac{3}{10} \times \frac{4}{9} + \frac{7}{10} \times \frac{24}{49} = 0.476$

Ⓑ Wind $W: 6^{4+}_{2-}$ $S: 4^{1+}_{3-}$

$G(D, \{W, S\}) = \frac{6}{10} \times \frac{16}{36} + \frac{4}{10} \times \frac{6}{16} = 0.417$

Ⓒ Humidity $H: 5^{1+}_{4-}$ $N: 5^{4+}_{1-}$

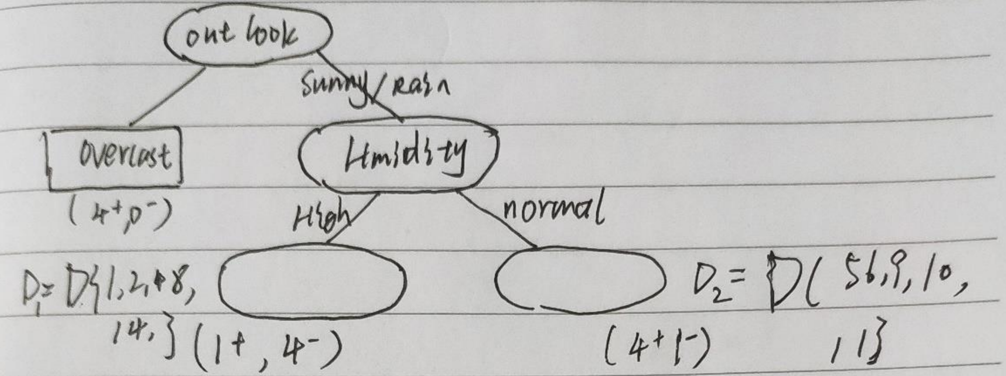
$G(D, \{H, N\}) = \frac{1}{2}\frac{8}{25} + \frac{1}{2}\frac{8}{25} = \frac{8}{25} = 0.32$

Ⓓ Out Look sunny 5^{2+}_{3-} Rain 5^{3+}_{2-}

$G(D, \{S, R\}) = \frac{1}{2}\frac{12}{25} \times 2 = 0.48$

∴ the smallese is Gini(D, {High, Normal}) = 0.32

∴



(III) the left, $D = D_2$

(A) outlook: Sunny: 3^-, Rain: 2^{1+}_{1-}

Gini(D, {S, R}) = $\frac{3}{5}\times 0 + \frac{2}{5}\times\frac{1}{2} = 0.2$

(B) temp: H: 2^- mild: 3^{1+}_{2-}

Gini(D, {H, m}) = $\frac{2}{5}\times 0 + \frac{3}{5}\times(1-(\frac{1}{3})^2-(\frac{2}{3})^2) = \frac{4}{15} = 0.267$

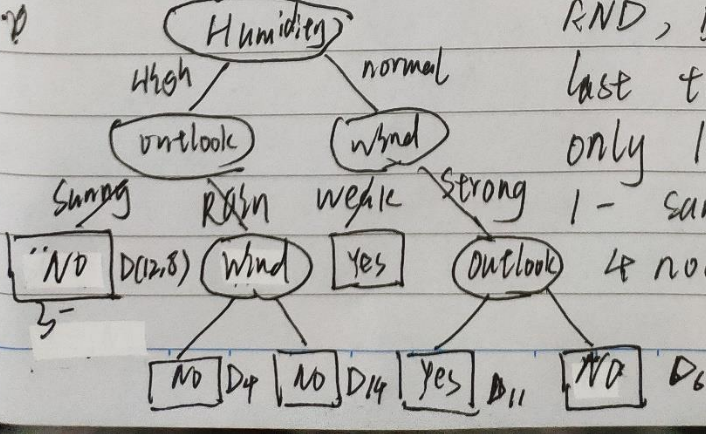
(C). wind: W: 3^{1+}_{2-} S: 2^-

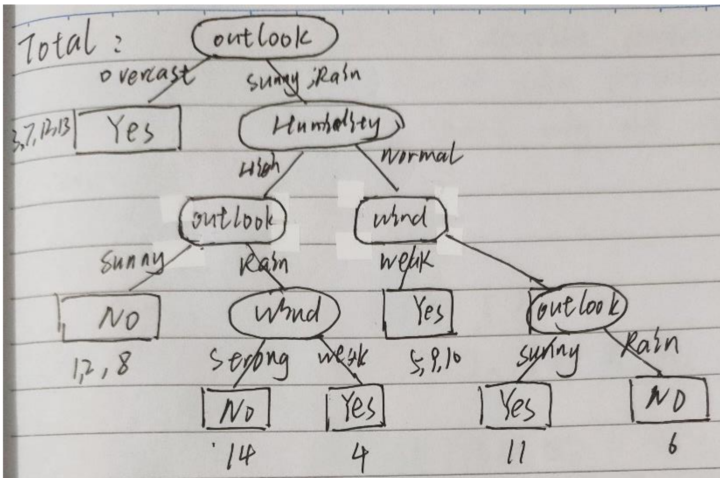
Gini(D, {w, S}) = $\frac{3}{5}\times\frac{4}{9} + 0 = 0.267$

∵ Gini(D, {S, R}) = 0.2 is the smallest

~~and~~ ∴ left is outlook, right is wind

∴ AND, Because the last two node have only 1+ sample and 1- sample, which creat 4 nodes.





(c) ① CART use minimization of squared error, ID3 use log, when calutate, CART is faster than ID3

② CART tree is a binary tree, ID3 is not. The branching factor of ID3 equas to the number of decision category

③ CART has a deeper tree than ID3, so when decition, ID3 is faster.

④ ID3 use information gain to describe choosing, so choose the laggest one which can make the remaining samples have more purity

CART use Gini index to describe decition, The goal is to minimize the probability of missclassification. That's why we choose the smallest one.

HW4 附加题

K聚类如何选初始点?

① 随机选取，多次平均。较果一般，也是最原始的选择方法

② K-means++. 选尽可能远的K个点。对于第 n (1<n≤k) 个点，选离前n-1个点的中心最远的点

③ ISODATA: 当某一类点过少，去除之。当某一类点过多、分散较大，分裂为两个聚类