: Information gain ابتدا (1)

"yes" = 9
"No" = 4 $\Rightarrow$ $H(S) = -\left(\frac{9}{13}\log_2\left(\frac{9}{13}\right) + \frac{4}{13}\log_2\left(\frac{4}{13}\right)\right)$

$= -(0.4\log_2(0.4) + 0.6\log_2(0.6))$
$= -(0.4 \times -0.529 + 0.6 \times -1.322)$
$= 0.971$

: Attributes حال انتروپی هر کدام از

1) Age : $H(\text{young}) = -(0.6\log_2(0.6) + 0.4\log_2(0.4)) = 0.971$

$H(\text{Middle}) = -(0.4\log_2(0.4) + 0.6\log_2(0.6)) = 0.971$

$H(\text{Old}) = -(0.8\log_2(0.8) + 0.2\log(0.2)) = 0.722$

میانگین وزنی دار : $\frac{5}{13} \times 0.971 + \frac{5}{13} \times 0.971 + \frac{3}{13} \times 0.722 = 0.888$

$\text{Gain(Age)} = 0.971 - 0.888 = 0.083$

2) Has Job :

$H(\text{True}) = -(0.5\log_2(0.5) + 0.5\log_2(0.5)) = 0$

$H(\text{false}) = -(0.4\log_2(0.4) + 0.6\log_2(0.6)) = 0.971$

میانگین وزنی دار : $\frac{5}{13} \times 0 + \frac{10}{13} \times 0.971 = 0.647$

$\text{Gain(Has Job)} = 0.971 - 0.647 = 0.323$

3) Own house

$$H(false) = -\left(\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\left(\frac{3}{6}\right)\right) = 0.918$$

$$H(True) = -\left(1\log_2(1) + 0\log_2(0)\right) = 0$$

میانگین وزنی : $\frac{9}{15}(0.918) + \frac{6}{15}(0) = 0.551$

$$Gain(Own\ house) = 0.971 - 0.551 = 0.420$$

4) Credit Rating

$$H(fair) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) = 0.722$$

$$H(Good) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 0.918$$

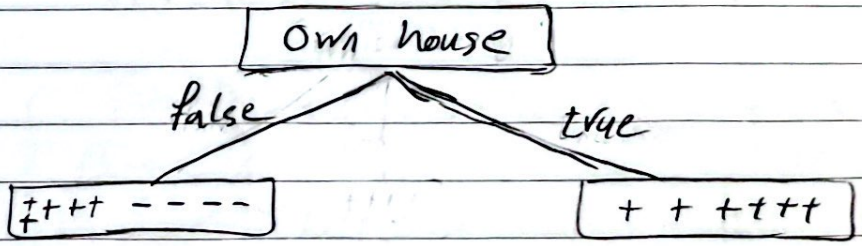
$$H(Excellent) = -\left(1\log_2 1 + 0\log_2 0\right) = 0$$

$$Remainder(Credit\ Rating) = \frac{5}{15} \times 0.722 + \frac{6}{15} \times 0.918 + 0 = 0.608$$

$$Gain(Credit\ Rating) = 0.971 - 0.608 = 0.363$$

⇐ بهترین انتخاب :

مرحله ۱ :



پس نیاز به حساب جدید هستیم . برای شاخه false .

1) false بله :

۱- محاسبه انتروپی نود : $H(S) = -\left(\frac{۶}{۱۰}\log_۲\left(\frac{۶}{۱۰}\right) + \frac{۴}{۱۰}\log_۲\left(\frac{۴}{۱۰}\right)\right)$

$= ۰/۹۱۸$

۲- محاسبه انتروپی و Info gain برای هر Att :

1) Age :

$H(young) = -(۰/۷۵\log_۲(۰/۷۵) + ۰/۲۵\log_۲(۰/۲۵)) = ۰/۸۱۳$

$H(middle) = -(۱\log_۲(۱) + ۰\times\log_۲(۰)) = ۰$

$H(old) = -(۰/۳۳۳\log_۲(\frac{۱}{۳}) + \frac{۲}{۳}\log_۲(\frac{۲}{۳})) = ۰/۹۱۸$

$Remainder(Age) = \frac{۴}{۹}\times ۰/۸۱۳ + \frac{۲}{۹}\times ۰ + \frac{۳}{۹}\times ۰/۹۱۸ = ۰/۴۴۷$

$Gain(Age) = ۰/۹۱۸ - ۰/۴۴۷ = ۰/۴۷۱$

2) Has Job :

$H(false) : -(۱\log_۲(۱) + ۰\log_۲(۰)) = ۰$

$H(True) : -(۱\log_۲(۱) + ۰\log_۲(۰)) = ۰$

$Remainder(Has\ Job) = ۰ + ۰ = ۰$

$Gain(Has\ Job) = ۰/۹۱۸ - ۰ = ۰/۹۱۸$

3) Credit Rating :

$$H(\text{fair}) = -(1 \log_2(1) + 0 \log_2(0)) = 0$$

$$H(\text{good}) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = 1$$

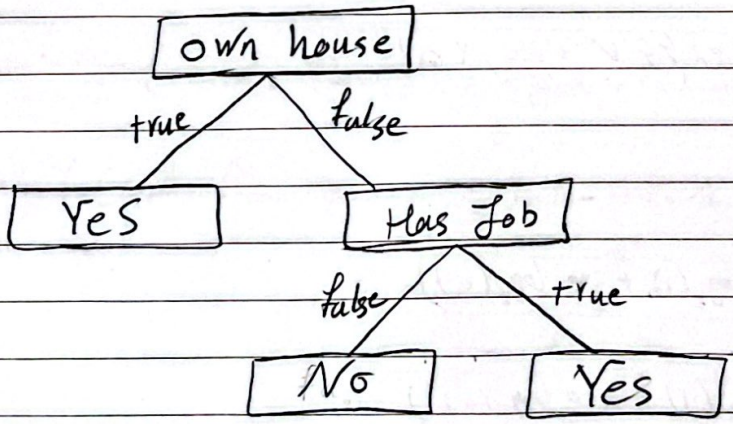
$$H(\text{Excellent}) = -(1 \log_2(1) + 0 \log_2(0)) = 0$$

$$\text{Remainder}(\text{Credit Rating}) = \frac{4}{9} \times 0 + \frac{4}{9} \times 1 + \frac{1}{9} \times 0 = 0.444$$

$$\text{Gain}(\text{Credit Rating}) = 0.918 - 0.444 = 0.474$$

=> بهترین انتخاب : Has Job

=> تفکیک کامل => ادامه دادن لازم نیست.



② با استفاده از Gini :

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2$$

۱- محاسبه Gini کل :

$$Gini(S) = 1 - (0.4)^2 + (0.6)^2 = 0.48$$

۲- محاسبه Gini برای همه Att ها و انتخاب بهترین :

1) Age :

$$Gini(young) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$$
$$Gini(middle) = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$$
$$Gini(old) = 1 - (\frac{4}{5})^2 - (\frac{1}{5})^2 = 0.32$$
میانگین وزن دار (Age) $= \frac{5}{15} \times 0.48 + \frac{5}{15} \times 0.48 + \frac{5}{15} \times 0.32 = 0.427$
$$Gain(Age) = 0.48 - 0.427 = 0.053$$

2) Has Job :

$$Gini(false) = 1 - (\frac{6}{10})^2 - (\frac{4}{10})^2 = 0.48$$
$$Gini(True) = 1 - (\frac{10}{10})^2 - 0^2 = 0$$
میانگین وزن دار (Has Job) $= \frac{10}{15} \times 0.48 + 0 = 0.32$
$$Gain(Has\ Job) = 0.48 - 0.32 = 0.16$$

3) Own house :

$$Gini(true) = 1 - (\frac{6}{6})^2 - 0 = 0$$
$$Gini(false) = 1 - (\frac{3}{9})^2 - (\frac{6}{9})^2 \simeq 0.44$$
میانگین وزن دار (Own house) $= 0 + \frac{9}{15} \times 0.44 = 0.264$
$$Gain(own\ house) = 0.48 - 0.264 = 0.216$$

4) Credit Rating

$Gini(fair) = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.32$

$Gini(good) = 1 - (\frac{4}{6})^2 - (\frac{2}{6})^2 \simeq 0.44$

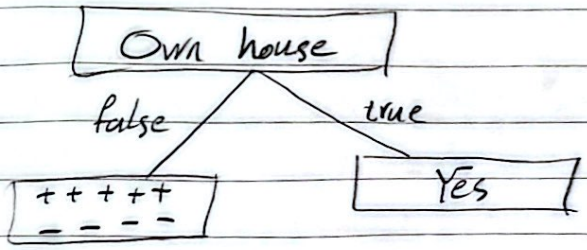
$Gini(excellent) = 1 - (\frac{4}{4})^2 - 0 = 0$

میانگین وزن دار (Credit rating) $= \frac{5}{15} \times 0.32 + \frac{6}{15} \times 0.44 + 0 = 0.282$

$Gain(Credit\ Rating) = 0.48 - 0.282 = 0.198$

Own house = بهترین انتخاب = Gain بیشترین =>

مرحله 1 :



حال تمام مراحل بالا را برای شاخه چپ دوباره انجام می دهیم :

۱- ابتدا Gini اولیه یا کل را حساب می کنیم : $Gini(S) = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.444$

۲- حساب برای Attribute ها :

1) Age :

$Gini(young) = 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375$

$Gini(middle) = 1 - (0)^2 - (\frac{2}{2})^2 = 0$

$Gini(old) = 1 - (\frac{2}{3})^2 - (\frac{1}{3})^2 \simeq 0.444$

میانگین وزن دار (Age) $= \frac{4}{9} \times 0.375 + 0 + \frac{3}{9} \times 0.444 = 0.314$

$Gain(Age) = 0.444 - 0.314 = 0.13$

2) Has Job

$Gini(false) = 1 - 0^2 - (\frac{6}{6})^2 = 0$

$Gini(true) = 1 - 0^2 - (\frac{3}{3})^2 = 0$

میانگین وزنی دار $= 0$

$Gain(Has\ Job) = 0/444 - 0 = 0/444$

3) Credit Rating

$Gini(fair) = 1 - (\frac{1}{1})^2 - 0 = 0$

$Gini(good) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0/5$

$Gini(excellent) = 1 - (1)^2 - 0 = 0$

میانگین وزنی دار gini $= 0 + \frac{4}{9} \times 0/5 + 0 = 0/222$

$Gain(Credit\ Rating) = 0/444 - 0/222 = 0/222$

⇐ با انتخاب Has Job که بیشترین Information gain را هم دارد درخت تشکیل می‌شود: