人工知能入門
第13回

藤田 悟

黄　潤和

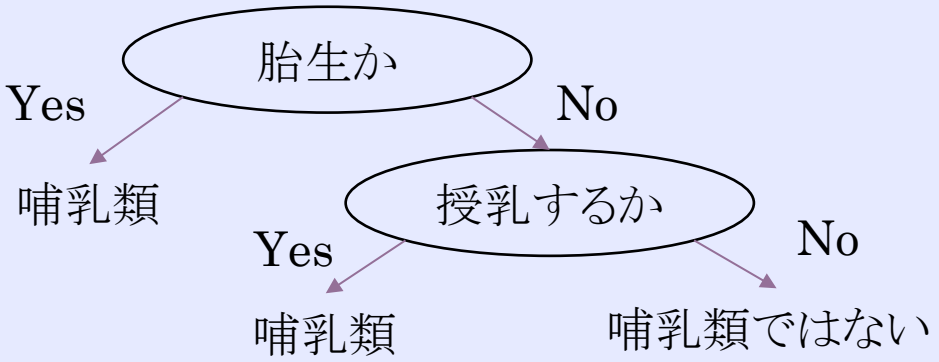
今回学ぶこと

- 分類木の学習
  - ID3

分類木の学習

- 特徴を順に質問して、最終的に対象の分類を行う構造を、分類木と呼ぶ。
  - 例えば、哺乳類の分類木を作ってみよう

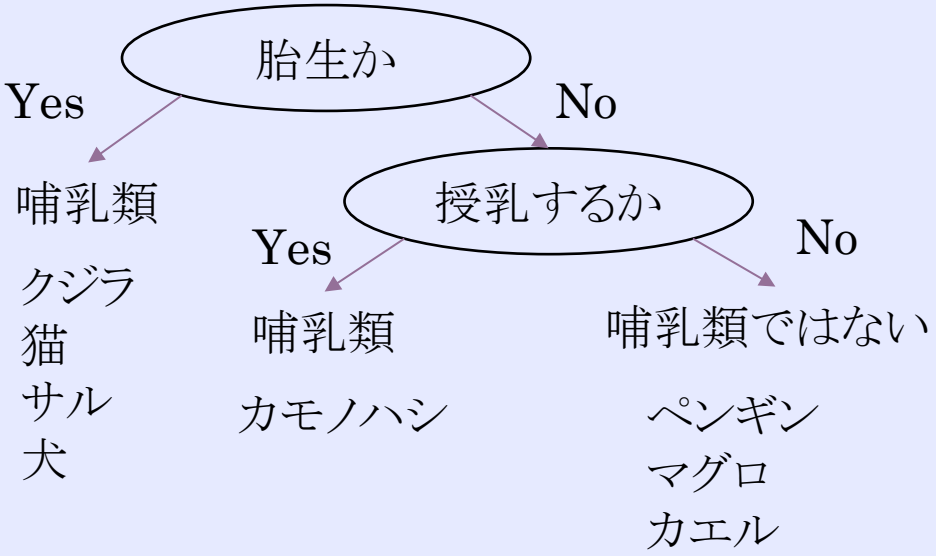
	泳ぐ	嘴を持つ	胎生である	授乳する	哺乳類
猫	no	no	yes	yes	yes
クジラ	yes	no	yes	yes	yes
ペンギン	yes	yes	no	no	no
カモノハシ	yes	yes	no	yes	yes

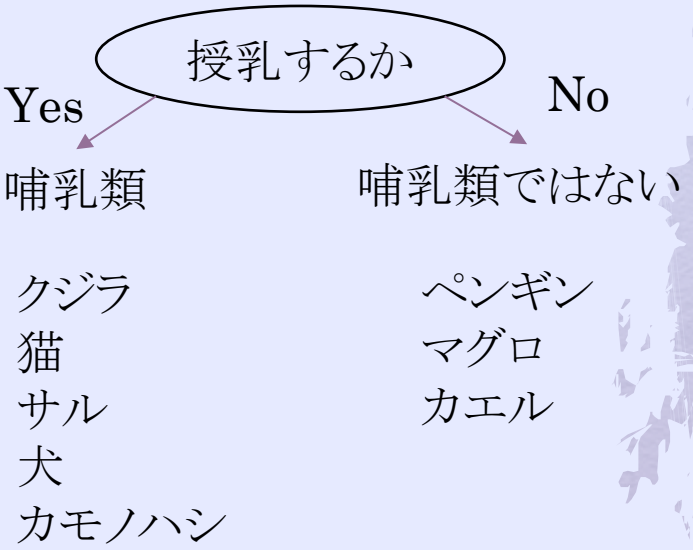


少し冗長な気が…

分類木の良しあし

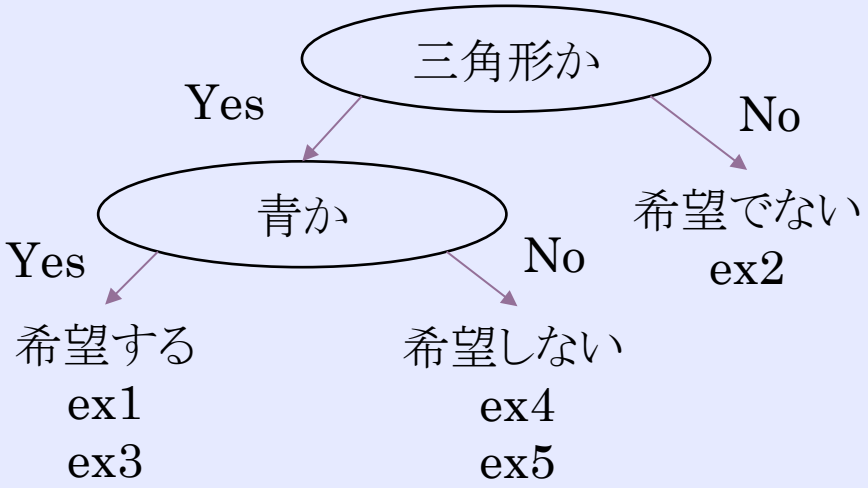
- 適当に作った分類木は判定するための多くの質問が必要になる。
  - 効率の良さはどのように判定できるか？

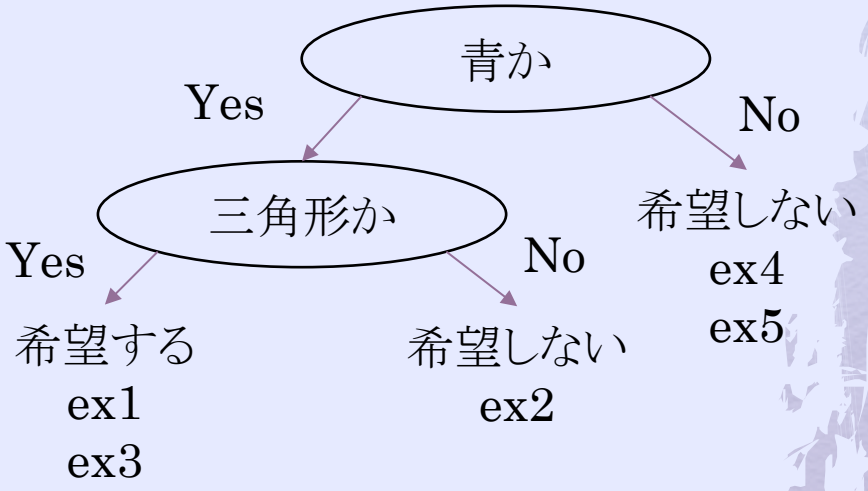




別の例: タイルの選択

	形	色	サイズ	材質	正/負
ex1	三角形	青	大	木	yes
ex2	正方形	青	小	プラスティック	no
ex3	三角形	青	小	プラスティック	yes
ex4	三角形	緑	大	プラスティック	no
ex5	三角形	黄	大	木	no





どちらの分類木が良いか？

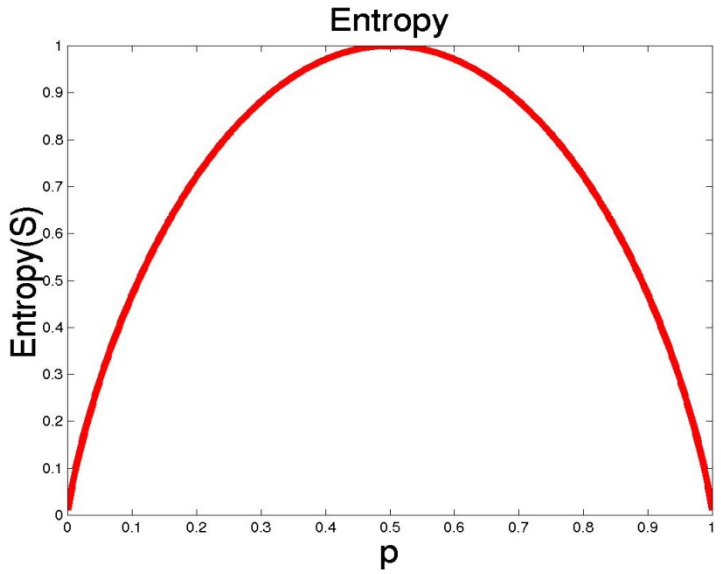
情報量

- 確率 p の事象の情報量

$$情報量(bit) = -\log_2 p$$

- コインを1枚投げて表が出る
  - $-\log_2\left(\frac{1}{2}\right) = 1$
- コインを2枚投げて、表表が出る
  - $-\log_2\left(\frac{1}{4}\right) = 2$
- サイコロで1の目が出る
  - $-\log_2\left(\frac{1}{6}\right) = 2.58$

Entropy



- S is a sample of training examples
- p_+ is the proportion of positive examples
- p_- is the proportion of negative examples
- Entropy measures the impurity of S

$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

情報エントロピー

- 確率 p の事象の情報量

$$情報量(bit) = -\log_2 p$$

- 情報エントロピー

$$情報エントロピー = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- 4個のデータ中に3個の正例、1個の負例

$$情報エントロピー = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= 0.75 \times 0.415 + 0.25 \times 2 = 0.811$$

情報エントロピーとは、情報のランダムさ・雑然さ、そして、情報を得た時の期待情報量

男性か、女性かを知る

- 公立小学校で、質問する(女性が1/2だとする)

$$\text{情報エントロピー} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$
$$= 0.5 \times 1 + 0.5 \times 1 = 1.0$$

- 情報科学部で、質問する(女性が1/4だとする)

$$\text{情報エントロピー} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$
$$= 0.75 \times 0.415 + 0.25 \times 2 = 0.811$$

- ある年度の某大学の電気電子工学科で、質問する(女性が1/89だとする)

$$\text{情報エントロピー} = -\frac{88}{89}\log_2\frac{88}{89} - \frac{1}{89}\log_2\frac{1}{89}$$
$$= 0.989 \times 0.0163 + 0.0112 \times 6.48 = 0.0889$$

同じ二者択一の質問をしても、状況により得られる情報量が違う！

タイルの選択

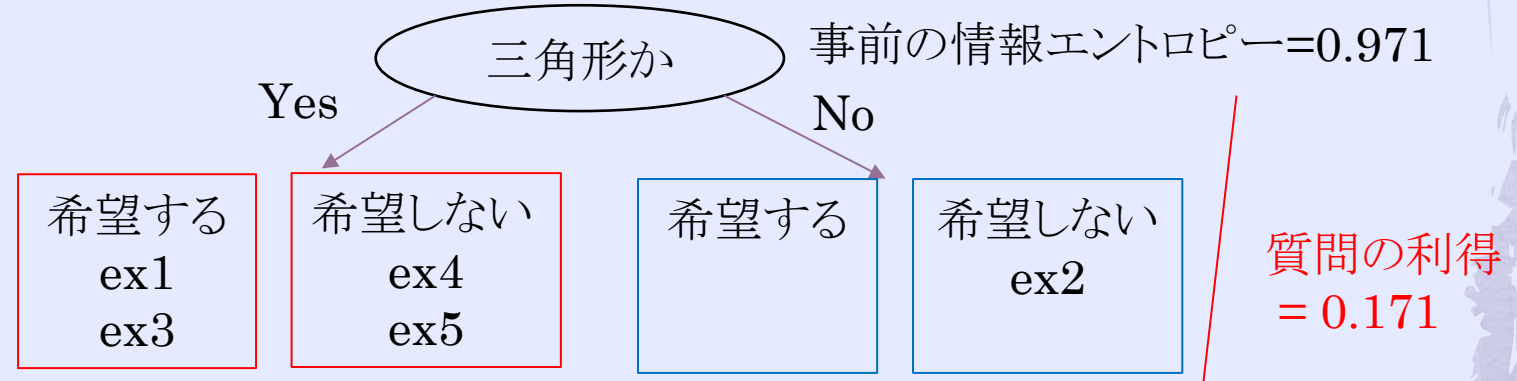
	形	色	サイズ	材質	正/負
ex1	三角形	青	大	木	yes
ex2	正方形	青	小	木	no
ex3	三角形	青	小	プラスティック	yes
ex4	三角形	緑	大	プラスティック	no
ex5	三角形	黄	大	木	no

初期状態の情報エントロピー

$$\text{情報エントロピー} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$
$$= 0.4 \times 1.32 + 0.6 \times 0.737 = 0.971$$

タイルの選択

	形	色	サイズ	材質	正/負
ex1	三角形	青	大	木	yes
ex2	正方形	青	小	木	no
ex3	三角形	青	小	プラスティック	yes
ex4	三角形	緑	大	プラスティック	no
ex5	三角形	黄	大	木	no



事前の情報エントロピー=0.971

質問の利得
= 0.171

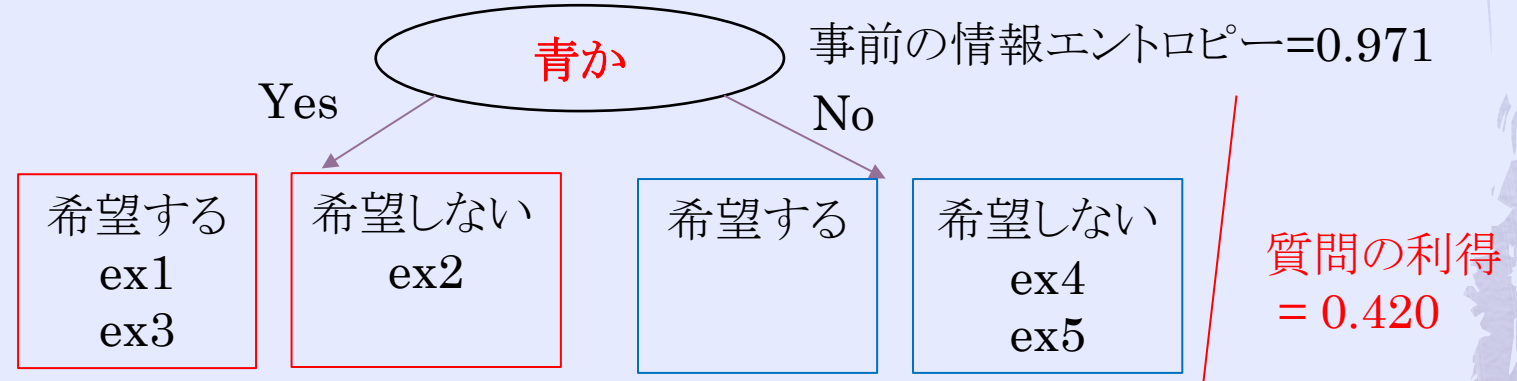
情報エントロピー $= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$
$= 1.0$

情報エントロピー $= -\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}$
$= 0.0$

判断後の期待情報エントロピー $= \frac{4}{5} \times 1.0 + \frac{1}{5} \times 0.0 = 0.8$

タイルの選択

	形	色	サイズ	材質	正/負
ex1	三角形	青	大	木	yes
ex2	正方形	青	小	木	no
ex3	三角形	青	小	プラスティック	yes
ex4	三角形	緑	大	プラスティック	no
ex5	三角形	黄	大	木	no



事前の情報エントロピー=0.971

質問の利得
= 0.420

情報エントロピー $= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$
$= 0.918$

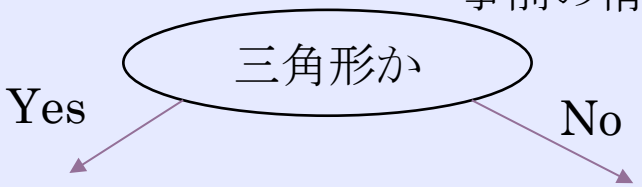
情報エントロピー $= -\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}$
$= 0.0$

判断後の期待情報エントロピー $= \frac{3}{5} \times 0.918 + \frac{2}{5} \times 0.0 = 0.551$

タイルの選択

	形	色	サイズ	材質	正/負
ex1	三角形	青	大	木	yes
ex2	正方形	青	小	木	no
ex3	三角形	青	小	プラスティック	yes
ex4	三角形	緑	大	プラスティック	no
ex5	三角形	黄	大	木	no

事前の情報エントロピー=0.971



利得= 0.171

判断後の期待情報エントロピー

$$= \frac{4}{5} \times 1.0 + \frac{1}{5} \times 0.0 = 0.8$$

利得= 0.420

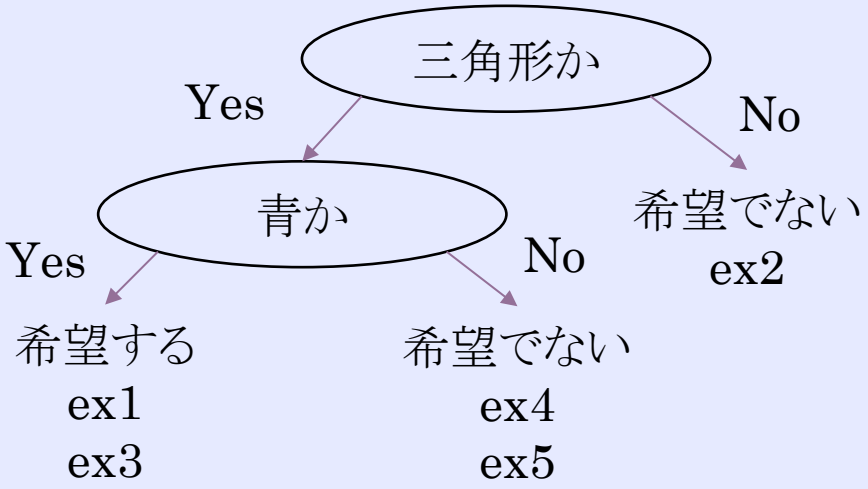
判断後の期待情報エントロピー

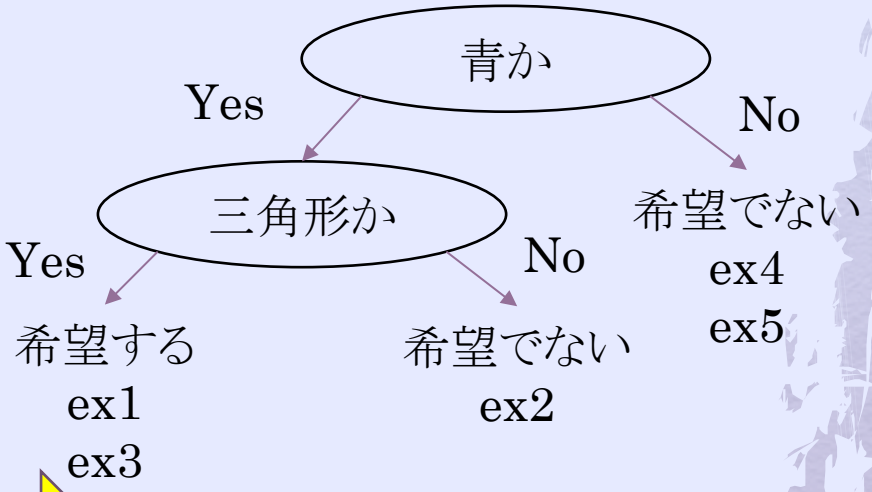
$$= \frac{3}{5} \times 0.918 + \frac{2}{5} \times 0.0 = 0.551$$

こちらの選択肢の方が利得が大きい

タイルの選択

	形	色	サイズ	材質	正/負
ex1	三角形	青	大	木	yes
ex2	正方形	青	小	プラスティック	no
ex3	三角形	青	小	プラスティック	yes
ex4	三角形	緑	大	プラスティック	no
ex5	三角形	黄	大	木	no





平均質問数 $= \frac{1}{5} \times 1 + \frac{4}{5} \times 2 = \frac{9}{5}$

良い

平均質問数 $= \frac{2}{5} \times 1 + \frac{3}{5} \times 2 = \frac{8}{5}$

分類木

- 分類をするための質問は、
  - 情報エントロピーの減少が大きい質問を、優先して質問する方が良い
    - 平均利得が最大の質問を行う
  - 情報エントロピーの減少が大きい質問を優先して行うと、全体の平均質問数が減少する
    - 少ない質問で、分類を完了できる
- このアルゴリズムを考案したのが、Quinlan であり、アルゴリズム名を ID3 と呼ぶ

(演習1) タイルの選択

	形	色	サイズ	材質	正/負
ex1	三角形	青	大	木	yes
ex2	正方形	青	小	プラスティック	no
ex3	三角形	青	小	プラスティック	yes
ex4	三角形	緑	大	プラスティック	no
ex5	三角形	黄	大	木	no

質問として、サイズや材質を選択した場合について、
質問後の平均情報エントロピーを計算せよ

(演習2) 弁当の選択

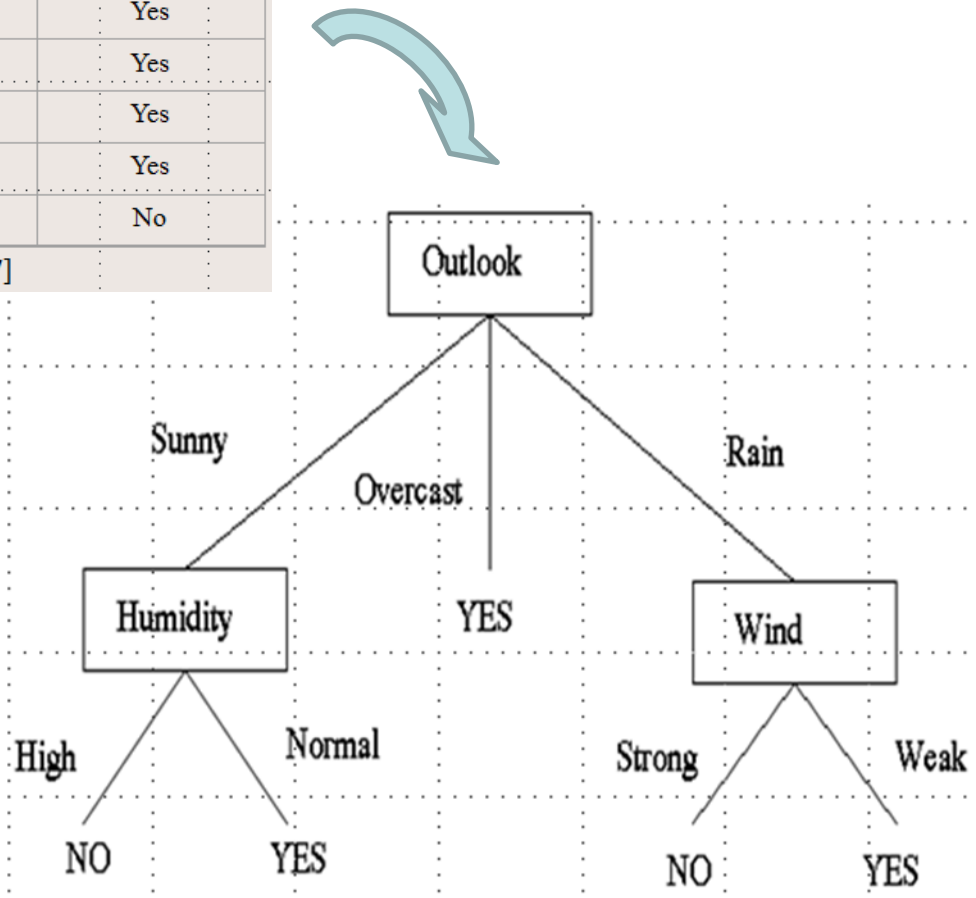
	主菜	野菜	量	値段	正/負
ex1	肉	キャベツ	大	安い	no
ex2	魚	キャベツ	小	安い	yes
ex3	肉	トマト	小	高い	no
ex4	魚	トマト	大	高い	no
ex5	魚	トマト	大	安い	yes

質問後の平均情報エントロピーを計算し、最適な質問の順番を決定せよ。

An example:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

[See: Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997]



http://www.cs.csi.cuny.edu/~imberman/ai/Entropy%20and%20Information%20Gain.htm

Outlook	Play Tennis
Sunny	No
Sunny	No
Overcast	Yes
Rain	Yes
Rain	Yes
Rain	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rain	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rain	No

P(Play Tennis = Yes) = 9/14
P(Play Tennis = No) = 5/14

Entropy(Play Tennis) = - 9/14 $\log_2$ (9/14) - 5/14 $\log_2$ (5/14) = .940

conditional entropy for rain

P(Outlook = Rain and Play Tennis = yes) = 3/5
P(Outlook = Rain and Play Tennis = no) = 2/5

$$Entropy(S_{rain}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = .971$$

$$Entropy(S_{overcast}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - 0\log_2(0) = 0$$

$$Entropy(S_{sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = .971$$

P(rain) = 5/14 P(overcast) = 4/14 P(sunny) = 5/14

$$\text{Entropy (Play Tennis | Outlook)} = -\frac{5}{14}(.971) - \frac{4}{14}(0) - \frac{5}{14}(.971) = .694$$

By knowing Outlook, how much information have I gained?

Entropy (Play Tennis) - Entropy (Play Tennis | Outlook) = .940 - .694 = .246

$$E(S) = \Sigma^{c}_{i=1} -p_i \log_2 p_i ,$$

Ball	Size	Color	Weight	Rubber?	Result (Bounces?)
1	Small	green	Light	yes	yes
2	Small	blue	Medium	no	no
3	Medium	red	Medium	no	no
4	Small	red	Medium	yes	yes
5	Large	green	Heavy	yes	yes
6	Medium	blue	Heavy	yes	no
7	Medium	green	Heavy	yes	no
8	Small	red	Light	no	no

Figure C1: Identification Tree Training Data

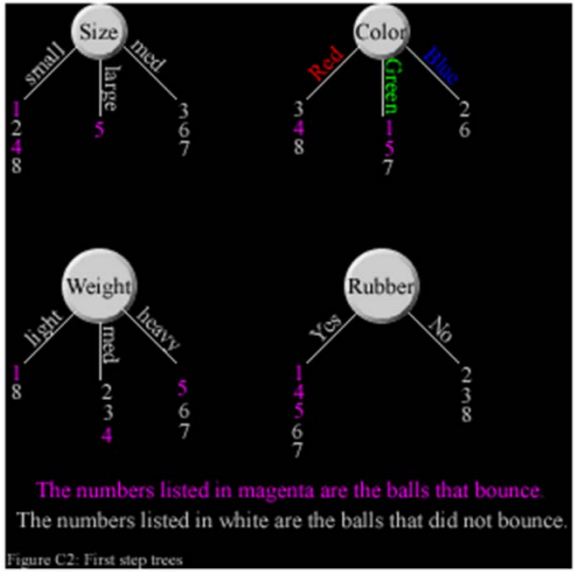


Figure C2: First step trees

$$Gain(S,F) = Entropy(S) - \sum_{v \in Values(F)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Disorder

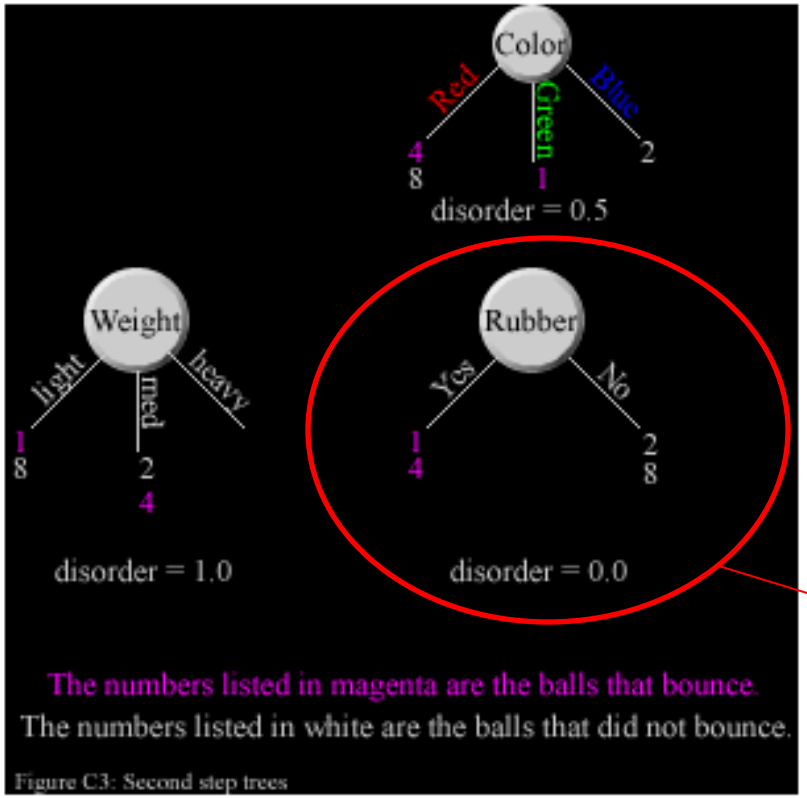
$$Size\_Disorder = \Sigma_b\ (nb/nt) * (\Sigma_c\ -(nbc/nb)\log_2(nbc/nb))$$

$$= (4/8) * ((-(2/4)\ \log_2\ (2/4)) + (-(2/4)\ \log_2\ (2/4))) + ((1/8) * 0) + ((3/8) * 0)$$

$$= 0.5$$

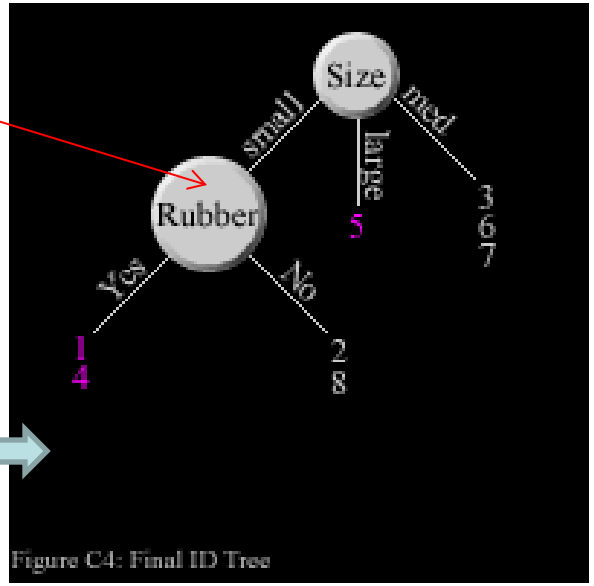
Color_Disorder = 0.69
Weight_Disorder = 0.94
Rubber_Disorder = 0.61

For the case of Size = small, continue to split this note



How about other two cases?
Split or not? Why?
- medium?
- large?

Finish splitting?
Why?



ID Trees to Rules

Once an ID tree is constructed successfully, it can be used to generate a rule-set, which will serve to perform the necessary classifications of the ID tree. This is done by creating a single rule for each path from the root to a leaf in the ID tree.

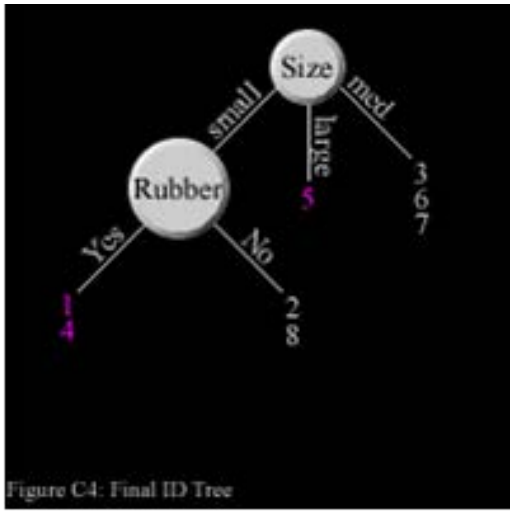


Figure C4: Final ID Tree

R1: if (size = large)
then (ball does bounce)
R2: if (size = medium)
then (ball does not bounce)
R3: if (size = small)
(rubber = no)
then (ball does not bounce)
R4: if (size = small)
(rubber = yes)
then (ball does bounce)

(課題) 合格者の選択

	性別	課題	試験	出席	合否
Aさん	男	提出	70	良好	合格
Bさん	男	提出	50	良好	不合格
Cさん	女	提出	70	良好	合格
Dさん	男	提出	80	良好	合格
Eさん	女	提出	80	良好	合格
Fさん	女	未提出	70	不良	不合格

質問後の平均情報エントロピーを計算し、最適な質問の順番を決定せよ。