知的情報処理論　第二回レポート

28G23027
川原尚己

1.

(a)　「ゲームを買う」ラベルが持つエントロピーをI_0とすると，

$$
\begin{aligned}
I_{g_0} &= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) \\
&= -\frac{2}{5}\log_2 2 - \frac{3}{5}\log_2 3 + \log_2 5 \\
&= -0.4 - 0.6 \times 1.58 + 2.32 \\
&= \mathbf{0.972}
\end{aligned}
$$

(b)　「評判」で分岐した場合：

「良い」：$(\mathrm{Yes}, \mathrm{No}) = (1, 1)$

「普通」：$(\mathrm{Yes}, \mathrm{No}) = (1, 1)$

「悪い」：$(\mathrm{Yes}, \mathrm{No}) = (0, 1)$

「時間」で分岐した場合：

「有」：$(\mathrm{Yes}, \mathrm{No}) = (2, 1)$

「無」：$(\mathrm{Yes}, \mathrm{No}) = (0, 2)$

「お金」で分岐した場合：

「有」：$(\mathrm{Yes}, \mathrm{No}) = (2, 2)$

「無」：$(\mathrm{Yes}, \mathrm{No}) = (0, 1)$

(c)　「評判」について，「良い」,「普通」,「悪い」である場合のエントロピーをI_{10}, I_{11}, I_{12}，また,「評判」におけるエントロピーの期待値$\mathrm{E}[\mathrm{I}_1]$は，

$$
\mathrm{I}_{10} = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1,
$$

$$
\mathrm{I}_{11} = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1,
$$

$$
\mathrm{I}_{12} = -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0,
$$

$$
\mathrm{E}[\mathrm{I}_1] = \frac{2}{5} \times 1 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0 = 0.8
$$

となる．

「時間」について，「有」,「無」である場合のエントロピーをI_{20}, I_{21},「時間」におけるエントロピーの期待値$\mathrm{E}[\mathrm{I}_2]$ とし，「お金」について，「有」,「無」である場合のエントロピーをI_{30}, I_{31},「お金」におけるエントロピーの期待値$\mathrm{E}[\mathrm{I}_3]$とすると，

$$\mathrm{I}_{20} = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = -\frac{2}{3} + \log_2 3 = 0.914,$$

$$\mathrm{I}_{21} = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0,$$

$$\mathrm{E}[\mathrm{I}_2] = \frac{3}{5} \times 0.914 + \frac{2}{5} \times 0 = 0.548,$$

$$\mathrm{I}_{30} = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1,$$

$$\mathrm{I}_{31} = -\frac{0}{1}\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0$$

$$\mathrm{E}[\mathrm{I}_3] = \frac{4}{5} \times 1 + \frac{1}{5} \times 0 = 0.8$$

となる．このとき，「評判」,「時間」,「お金」の情報利得$Gain\left(天気\right), Gain\left(時間\right), Gain($

$お金)$は，

$$Gain\left(天気\right) = I_0 - \mathrm{E}[\mathrm{I}_1] = 0.972 - 0.8 = 0.172$$

$$Gain\left(時間\right) = I_0 - \mathrm{E}[\mathrm{I}_2] = 0.972 - 0.548 = 0.424$$

$$Gain\left(お金\right) = I_0 - \mathrm{E}[\mathrm{I}_3] = 0.972 - 0.8 = 0.172$$

となり，「時間」によって分岐させると情報利得が最大となる．

(d) (a)から(c)で行なったものと同様の処理を各ノードにおいて繰り返し実行すると，図 1 のような決定木が得られる．

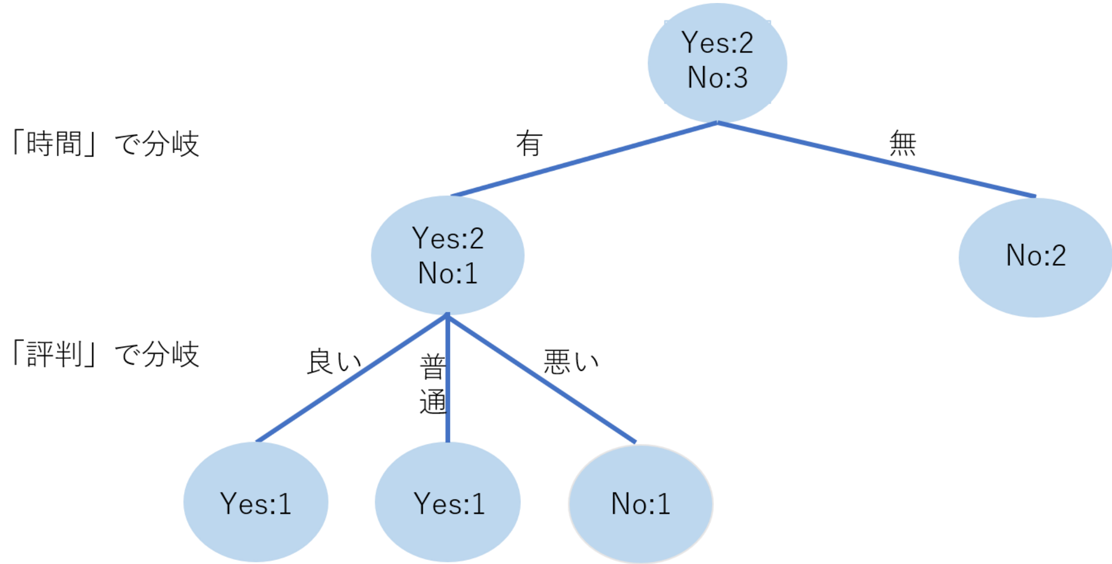

図 1　決定木の分類結果

２．

“Breast Cancer Wisconsin”という，乳房腫瘍に関するデータセット(WDBC)を用い，乳がんが陽性かどうかの分類を SVM を用いて行う．

WDBC データセットは 569 人のレコードからなり，属性数は11である．腫瘍半径・テクスチャ・周長・面積・平滑度・緻密さ・凹面度（輪郭の凹部の激しさ）・凹点（輪郭の凹部の数）・対称性・フラクタル次元及び診断結果からなり，診断結果は 1（陽性），0（陰性）であり，他の属性はすべて連続値を持つ．今回の測定では，尺度の差を是正するため，連続値を持つ各属性の値を区間[0,1]の値へと正規化を行ってから測定を行っている．

SVM として使用したライブラリは，sklearn.svm.SVC であり，カーネルは”rbf”，ハイパーパラメータは$\mathrm{C} = 1.3$，$\mathrm{gamma} = 0.5$とした．測定は学習と計測に同じデータを使用する closed test，異なるデータを使用する open test，学習と計測に使用するデータを逐次変更しながら測定する交差検定の三種類で行い，推定の正しかった割合を測定した．open test には WDBC データセットの前半284レコードを学習に，後半 285 レコードを計測に使用し，交差検定には．sklearn.model_selection. KFold を用い，分割数Kは，$\mathrm{K} = 10$とした．測定結果を以下の表 1 に示す．

closed test	open test	交差検定
0.984	0.968	0.972

表 1　WDBC データセットに対する測定結果

表 1 を見ると，わずかな差ではあるが，open test よりも closed test の方が高い有用性を示している．問題文の設定に従うと，「(b) 学習データとは異なるデータを使用した場合」よりも「(a) 学習データをそのまま使用した場合」の方が高い有用性を示しているといえる．この理由としては，closed test は open test と比べて「学習に用いるデータが多いこと」に加えて「評価を既知のデータで行うことができること」が影響しているものと思われる．

しかし，実際のところ今回の実験では closed test と open test の結果は非常に近い値を出力しており，上記のような考察は妥当であるとは言いがたい．二者の間で近い値が出力されたのはモデルの学習に対してデータセットのレコード数が十分多かったためであると考え，以下のような実験を行った：

closed test と open test に対して，学習及び計測に用いるデータ数の割合，測定方法は先ほどの実験と等しいまま，使用するデータを$x \in \{20, 100,300,569\}$個だけランダムに使用し，測定する．この処理を各データ使用数に対して100回行い，有用性の平均を出力した．その結果を以下の表 2 及び図 2 に示す．

データ数	closed test	open test
20	0.988	0.853
100	0.984	0.939
300	0.985	0.962
569	0.985	0.971

表 2　有用性平均

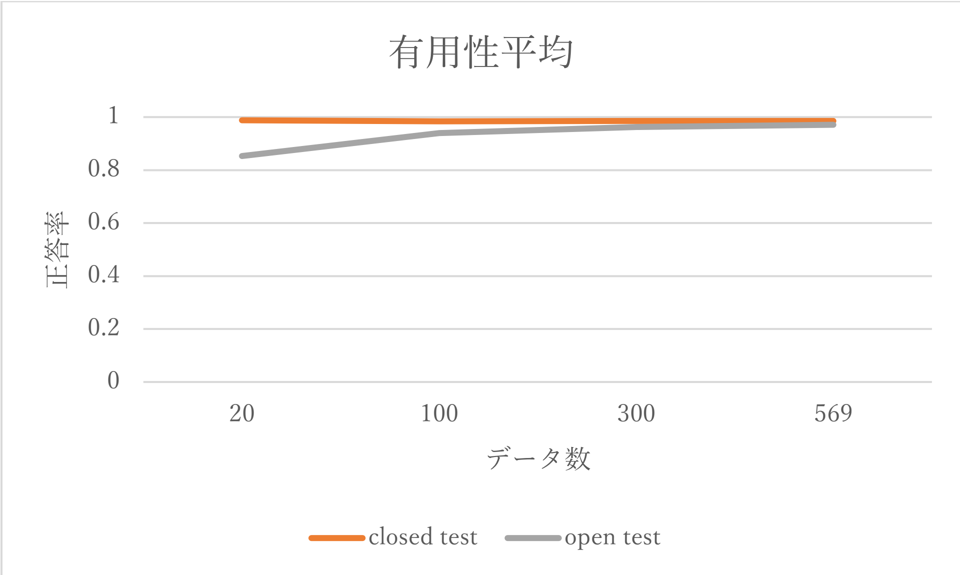

図 2　有用性平均のグラフ

この結果より，

- closed test はデータ数にはほとんど依存せず，正答率98%程度を示している．
- open test はデータ数が小さくなるほどに，正答率が下がっている．

ことが読み取れる．前者からは評価を既知のデータを用いて行っていることにより，少ないデータ数でも高い有用性を得ることができており，後者からはデータ数の多寡が有用性に小さくない影響を与えると考察できる．