知的情報処理論　第二回レポート

28G23027

川原尚己

1. 「ゲームを買う」ラベルが持つエントロピーをとすると，
2. 「評判」で分岐した場合：

「良い」：

「普通」：

「悪い」：

「時間」で分岐した場合：

「有」：

「無」：

「お金」で分岐した場合：

「有」：

「無」：

1. 「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，また，「評判」におけるエントロピーの期待値は，

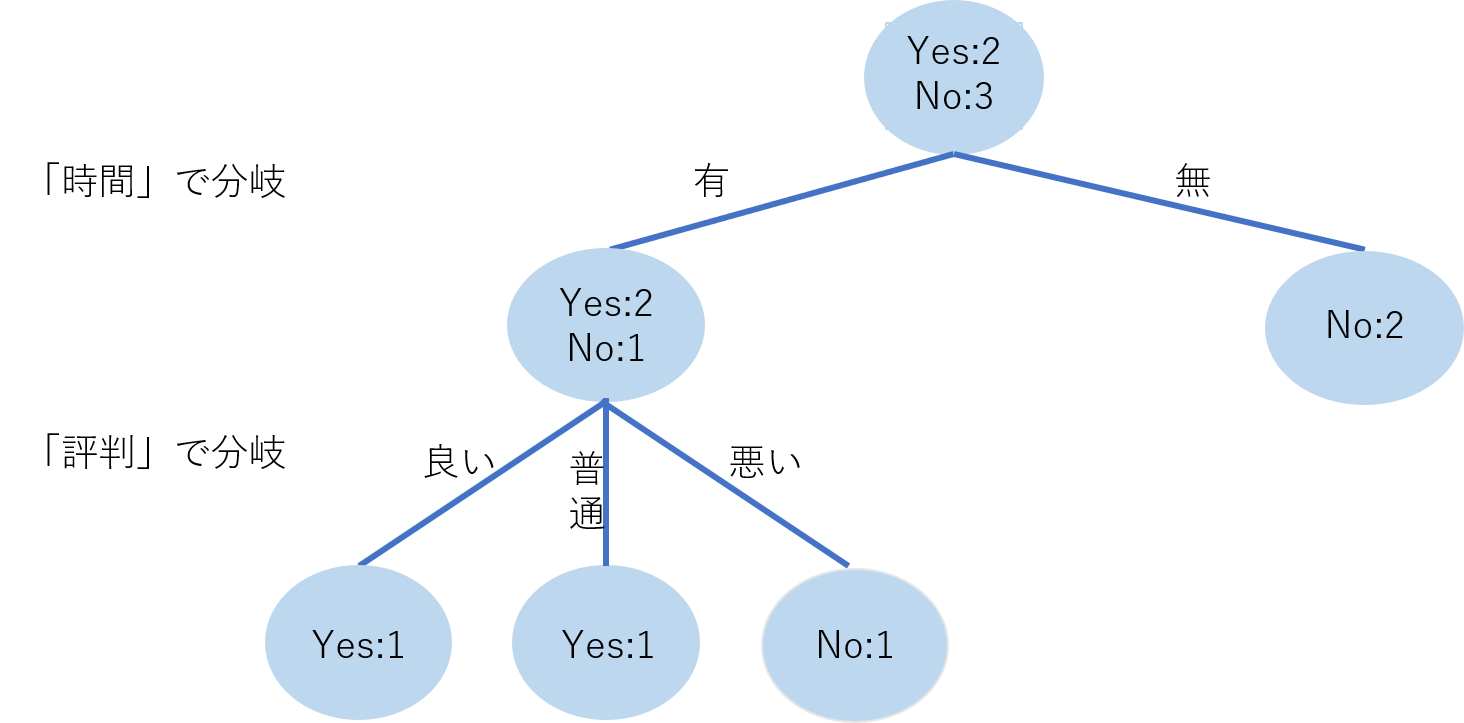
となる．

「時間」について，「有」，「無」である場合のエントロピーを，「時間」におけるエントロピーの期待値 とし，「お金」について，「有」，「無」である場合のエントロピーを，「お金」におけるエントロピーの期待値とすると，

となる．このとき，「評判」，「時間」，「お金」の情報利得は，

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．

1. (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，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，分割数は，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 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

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．