第2回レポート課題

2023/5/16 駒谷 和範

課題：以下の2題に取り組み，その結果分かったことや自分で考えたことを報告せよ．

1. 表1中の「ゲームを買う」ラベルを予測する決定木を作成することを考える．つまり．表1のデータを全て学習に用いて決定木を作成する．このとき以下の各問いに答えよ．ただし解答だけでなく導出過程を示すこと．

   木の分岐には，評判（ゲームの出来に関する評判），時間（自分の余裕時間），お金（自分の所持金）の3つの属性を用いる．必要に応じて $\log_2 3 = 1.58$，$\log_2 5 = 2.32$ を使うとよい．

表 1: ゲームを買うか否かに関するデータ

評判	時間	お金	ゲームを買う
良い	無	有	No
普通	有	有	Yes
普通	無	無	No
良い	有	有	Yes
悪い	有	有	No

   (a) 「ゲームを買う」ラベルが持つエントロピーを求めよ．

   (b) 評判，時間，お金の3つの属性それぞれで，最初に分岐を作成する場合を考える．このとき，この3つの属性の値で分岐を作成した際の,「ゲームを買う」ラベルの分布を書き下せ．

   (c) 上記 (b) の結果に基づいて，属性「評判」で最初に分岐を作成した後のエントロピーの期待値を求めよ．同様に，時間，お金で最初に分岐を作成した後のエントロピーの期待値も求めよ．また，これらの結果から，最初にどの属性で分岐させるのがよいか答えよ．情報利得が最大となる属性を選択することとする．

   (d) この手続きを繰り返して得られる決定木を示せ．

2. 何らかの分類問題または回帰問題を設定し，それを機械学習により解け．機械学習ライブラリを使用して構わない．

   - 評価データとして，以下の2つの場合の性能を比較し，そのような性能が得られた理由について考察せよ．

     (a) 学習データをそのまま使用した場合
     (b) 学習データとは異なるデータを使用した場合

     説明には以下の単語を含めるとよい．

     closed test, open test, 交差検定 (cross validation)

   - 実験の詳細を必ず示すこと．例えば以下が含まれる．
     - 問題の定式化（少なくとも入力と出力の仕様）
     - 用いたデータの詳細．学習データや評価データの量
     - 実験設定（用いた場合は機械学習ライブラリの名前を含む）
     - 用いた機械学習アルゴリズムの概要

  - 指定可能なハイパーパラメータ (hyperparameter) の値.
    何を指定できるか，その意味，デフォルト値，など．全てを書く必要はない．
- ハイパーパラメータチューニングを行った場合はその内容も書くとよい．また検証用データセット（validation dataset）の必要性についても議論するとよい．
- Weka などに同梱されているサンプルデータなどを使ってもよい（当然出典は明記すること）．ただしその場合でも，分類の目的や入力の仕様などは説明すること．

- 注意：
  - 考察や説明など考えたことを，他人が読んでわかる日本語（または英語）で書くこと．**プログラムや実行結果のみを送りつけてきた場合は，極めて低く評価する，または受理しない**．ソースコードをレポートに含める必要はない．
  - 出典や参考にした情報源がある場合は明記すること．剽窃や盗用が疑われる場合は相応の処置を取る．自分が理解したことを書くこと．
- その他，講義に対する感想や要望など，何かあれば書いてください．
- 本レポートは情報通信工学演習（情報通信工学コース必修）の一部である．知的情報処理論の成績にも加味する．
- 提出期限： 2023 年 6 月 12 日 (月)
- 提出方法： CLE 上にて PDF で提出（Word ファイルでも可）