集合知プログラミング
第7章
決定木によるモデリング(3)

06T4007A　伊藤輝将

数値による帰結への対処

- 数値的帰結を持つデータセットにbuildtreeをかけることは可能だがあまりいい結果は得られない

  →数値のすべてが異なる「ｶﾃｺﾞﾘｰ」として扱われると、ある数値とある数値が近く、他の数値は遠い、という事実を考慮に入れないため

- 帰結が数値的な場合のスコアリング関数には分散(variance)を使う

住宅価格のモデリング

・Zillow　API

-不動産価格を追跡し、その情報を使って他の住宅の推測価格を生成する無料のウェブサービス

-問い合わせの検索パラメータをすべて含んだURLにリクエストをかけ、返されたXMLをパースして詳細、つまりベッドルーム数などと推測価格を取得する

"HOTNESS"のモデル化

- Hot or Not
  - -ユーザーが自分の写真をアップロードできるサイト
  - -APIによりメンバーのデモグラフィーデータと"hotness"レーティングを取得できる

- Hot or Not API
  - -クエリのパラメータをURLに渡して返ってきたXMLを解釈する

決定木を使うべき場面

- 決定木の利点
  - -トレーニング後のモデルが非常に簡単に解釈できる
  - -データの確率的な振り分けをも可能

- 決定木の不利な点
  - -多数の帰結を含むデータセットにはあまり効果がない
  - -単純な数値データが扱えるものの、判別ポイントとしては、ある数以上/以下というものしか持つことができない

決定木を使うべき場面(2)

- 決定木が強いのは、区切りのある数値的データやカテゴリー的なデータといったものを多数持つものに対して

  →判断を下すプロセスの理解が重要であれば決定木がベストチョイス

- 多数の数値的な入力と出力をもつ問題や、数値的入力同士に多くの複雑な関係が存在する問題には、おそらく決定木はよい選択ではない

  例：財務データの解釈、画像分析