第7章
決定木によるモデリング

発表者
04T4027T　斉藤　章久

はじめに

- この章で『決定木学習』という手法を学ぶ

決定木は見るだけでその推論家庭が理解できるし、if-then文の単純ならびに変換することも可能

- この章では決定木を使った3つの例を紹介する。

①どのユーザがサイトのプレミアム会員になりそうか予測する方法

②住宅価格のモデリング

③"hotness"のモデリング

サインアップを予測する（1）

無料アカウントと登録アカウントがある新しいアプリケーション　←（好奇心で訪れるユーザが多い）

サインアップした人にくまなくメールを出すような手段に訴えるサイトが多い

↓

　あるユーザが有料顧客になる見込みの予測を可能にできると便利である

サインアップを予測する（2）

サインアップを可能な限り早く行わせるため、サイトではユーザにあまり質問しない

↓

代わりにサーバーログを見て、リンク元サイト、ユーザーの所在地、サインアップ前に見たページ数などを収集する。

決定木

決定木とは

- 観測を分類する完全に透明な手法。
- トレーニング後はツリー状に配置したif-then文のようになる。
- 質問に正しく答えながら枝をたどって下りていけば、答えに達する。
- 逆向きにたどっていくことで、その最終分類に達した理由もわかる。

ツリーのトレーニング

・決定木の構築はまずルートノードの作成

↓

すべての変数を調べ、どの条件で帰結を分割するとユーザーの行動が予測しやすくなるか決定する

・最高の分割をするには、ある集合がどの程度混合するかを計測する手段が必要。

・混合の度合いを計算する尺度がある。

ここで紹介するのは、ジニ不純度とエントロピーである。

ジニ不純度

- 集合中のアイテムのひとつに、帰結のひとつをランダムに当てはめる場合の期待誤差率のこと。

- まず可能な帰結それぞれがおきる確率を計算（帰結の総数を、集合にある行の総数で割ったもの）

  その確率同士をすべて乗じ、合計する

  ↓

  無作為な予測が誤った帰結を与える場合の総確率が得られる。

エントロピー

エントロピーとは集合中の無秩序の量である。

（求め方）
それぞれの要素の頻度を計算
p(i)＝頻度（帰結）＝度数（帰結） / 度数（行）
それを次式に当てはめる
エントロピー＝すべての帰結のp(i) x log(p(i))の合計

これは帰着同士がどれだけ違っているかの尺度である。
帰結がすべて等しければ0。混同したグループになるほどエントロピーは増大する

エントロピーとジニ不純度の違い

- 大きな違いとしては、エントロピーの方がピークに達するのが遅いこと

↓

混合した集合に対するペナルティがわずかに重くなる傾向がある

この章では、より普及しているエントロピーの方を尺度として使用する

情報ゲイン

- ある属性がどれほど良いか調べるとき

①グループ全体のエントロピーを計算する

②各属性の取りうる値によってグループを分割

③分割後の両方のグループについてサイドエントロピーを計算

ここで、最も優れた分割になる属性を決定するために計算されるのが情報ゲインである。

情報ゲインはグループ全体から、分割後の2グループのエントロピー加重平均を引いたものである。

枝の分割

条件に合致するかしないかで、まず分割される

↓

両方の枝についてさらなる分割が行えるか判断する

↓

新しい枝がさらに分割可能であれば、上と同じ手順に従いどの変数を使うか判断する

【分割をやめるのは、そのノードの分割による情報ゲインが正の数にならなくなったとき】

再帰的ツリー構築

①ループをかけ、各項目が値をすべて見つけ出す。

②データをサブセットの対に分割する。

③それぞれのサブセットのエントロピーに要素数の割合を掛けて、
　エントロピーの加重平均を計算する。

④どの対のエントロピーが最も低いか記録していく。

↓

加重平均エントロピーが、最良のサブセットのよりも現在の集合
　の方が低くなれば、その枝は終了となる。

そうでなければ、各サブネットに対してbuildtreeがコールされる。

こうして決定木全体が構築される