【決定木】

機械学習にも統計学にも多く用いられる手法

【決定木とは】

物事を木構造でデータを分類していく手法である。

（決定木にはCARTというアルゴリズム名がある。）

【例】

目的変数＝会員と非会員の分類分け

顧客（ルートノード）

∟男性（ターミナルノード）

∟女性

　　　∟大人（ターミナルノード）

　　　∟子供（ターミナルノード）

※各項目をノードといい、ノードの末端の事をターミナルノードと言う。

【決定木を利用する場面】

・予測における初期分析

・現状のデータの構造把握　　などの場面で使用されている。

【例】

目的変数＝顧客の購買予測。

・使用データ

行動データ（ネットの閲覧歴や購買履歴などから分ける個人の趣向など）

属性データ（性別、年齢、居住地域などのパーソナルデータ）

上記を木構造で分類分けを行い目的変数を求める。

また、以下の2点についての特徴があるため、多く用いられている。

・モデルの簡便さ

・アウトプットのわかりやすさ

【決定木を応用した手法群】

決定木を作成し、最適な分類を行う事を目的として機械学習に利用されている。

しかし、決定木だけを使用した場合の予測精度はとても高いとは言えない。

そのため、アンサンブル学習を用いた2つの主な手法が用いられている。

【ランダムフォレスト】

訓練データをランダムにサンプリングする→複数の決定木を作成する

→結果をそれぞれ出力する→分類問題なら多数決、回帰問題なら平均を取る。

最終的な多数決結果や平均を出したモデルを最終的に採用する。

**この並列して決定木モデルを作成する手法をバギングという。**

【Xgboost】

訓練データから決定木を作る→結果を出力する→決定木を作る→結果を出力する

→決定木を作る→結果を出力する。

**出力したデータを元に決定木を作成する直列的に最適モデルを作成する事をブースティングという。**