決定木 (Decision Tree)

決定木 (Decision Tree)

決定木とは木構造を用いて分類や回帰を行う教師あり学習の手法の 1 つである．再帰的にデータ分割し, データから木構造の分類器を生成する．訓練データ $\{(\mathbf{x}_i, y_i), i = 1, 2, \ldots n\}$ について, 目的変数 (ラベル) が y_i が連続変数 (量的変数) の場合は回帰木, 離散変数 (質的変数) の場合は分類木と呼ばれる．

利点と欠点

決定木の利点は, 分類器が木構造で与えられるため可読性が高く, 各特徴量がどれだけ予測に寄与したかを評価できる．さらに, 使用するデータが質的データと量的データどちらも扱えることなどがある．また, データを分割して, 学習を進める事から外れ値に対して頑健であることも利点．欠点は, それほど他の分類 (SVM など) と比べて分類性能が低いこと, 木が過度に分岐することがある (過学習)．他にも, データの変化に弱くデータの些細な変化で全く異なる木が生成されることや, 線形性のあるデータには適していないこと, 同時に 2 変数を扱えないことなどがある．

基本構造

決定木の基本構造は, 根ノード, 葉ノードの 2 つに分かれている．根ノードから 2 つの葉ノードを追加することでを枝分かれさせ, 木を成長させる．

分割基準

決定木では, 各ノードで与えられたデータをうまく分離してくれるような特徴量としきい値でデータを分割し, 枝分かれさせる必要がある．ではどのように枝分かれするのか．特徴量としきい値とのすべての組み合わせにおいて, 評価関数 E の値を計算し, それが最も小さい組み合わせを採用する．代表的な評価関数として, データの混ざり具合である不純度を表す指標であるジニ係数と, 情報量を表す指標であるエントロピーが挙げられる．

ジニ係数は, データからランダムに 2 つの要素を抜き出したときに, その 2 つのデータがそれぞれ別のクラスに属する確率である．よって不純度が高いほどジニ係数としては大きくなり, 逆に不純度が低いほどジニ係数が小さくなる．ジニ係数は計算が速く, 大きなデータセットでも効率的であること, 分類エラーの確率を直接的に表現できる利点がある．しかし, エントロピーと比べて情報の損失やデータの不確実性を直接的には表現できない欠点がある．

エントロピーは, 情報の不確実さを表したものである．ある確率分布が与えられたとする．結果が n 通りある場合, どれも等確率で起こるとき $(p_1, p_2, \ldots p_n) = (\frac{1}{n}, \frac{1}{n}, \ldots \frac{1}{n})$, エントロピーは最大になる．逆に 1 つの結果が必ず起こるとき $(p_1, p_2, \ldots p_n) = (1, 0, 0, \ldots, 0)$, エントロピーは最小になる．エントロピーはジニ係数はよりもバランスの取れた決定木を生成する傾向があること, データの不確実性を量化できる利点がある．しかし, ジニ係数に比べて計算コストが高くなる可能性があること, 小さな変更に対して過敏に反応する欠点がある．

K はクラスの数, p_{ik} は葉ノード i におけるクラス k のサンプルの割合のとき, 評価関数 E は

$$E = \sum_{i=1}^{n} E_i$$

ジニ係数

全体の確率からランダムに 2 つの要素を抜き出したときに, その 2 つのデータが同じクラスに属する確率を引く.

$$\begin{aligned} E_i &= \sum_{k=1}^{K} p_{ik} - \sum_{k=1}^{K} p_{ik}^2 \\ &= 1 - \sum_{k=1}^{K} p_{ik}^2 \end{aligned}$$

エントロピー

$$E_i = -\sum_{k=1}^{K} p_{ik} log(p_{ik})$$

過学習

決定木における過学習は, 木が深く広く枝分かれしたときの細かな枝分かれは必ずしも本質的な違いとは限らず, 収集したデータのみが持つ微妙な違いに応じた分岐でしかないときに汎化性能が低下することである. 過学習の対策として, 学習された木に対して細かく分岐した枝をある判断基準のもとで切る対策がある. これは枝刈りまたは剪定と呼ばれる. 他にも最大深さと最小分割数をあらかじめ設定して, 学習することで過学習を抑制できる. 決定木の精度と複雑度はトレードオフの関係性があるため, クロスバリデーションなどで検証するべきである.

アンサンブル学習

単独で精度が低い場合であっても複数の機械を用いることで精度が向上することがある. それぞれ並列に学習させた学習器を複数用いて多数決や平均によって最終的な分類予測を行うものをアンサンブル学習と呼ぶ. 複数の決定木を組み合わせて分類予測をするものはバギング (Bagging), ランダムフォレスト (Random Forest), ブースティング (Boosting) などがある. アンサンブル学習は決定木の課題である精度については向上するが, 複数の決定木を統合するため単一の決定木ほどの直感的な説明が難しくなるので, モデル全体の解釈性が低下する欠点がある.

バギング (Bagging)

訓練データから重複を許した m 個のデータを無作為抽出 (Bootstrap Sampling) し, 各データに対して決定木を作成する. 予測値が連続データの場合は予測値の平均をとり, 離散データの場合は確率が最も高いもの (多数決) によって予測値を統合する. 異なる木が同じ特徴量に過度に依存するため m 個の予測値は相関が高く, 予測精度が悪いという課題がある.

ランダムフォーレスト (Random Forest)

バギングの課題を改善するために, ランダムサンプリングされた訓練データとランダムに選択された特徴量を用いることで, 相関の低い決定木群を作成する. 過学習を防げること, 特徴量の重要度を計算できることも利点である.

ブースティング (Boosting)

ブースティングとは, 与えられたデータから決定木を生成し, 予測が正しくできなかったデータに重みをつけて再度, 決定木を生成する. これを繰り返すことで精度を高める方法. さらに, データに重みづけするのではなく予測値と実績値の誤差を計算し, 誤差を決定木で学習する方法は勾配ブースティングと呼ばれる.

chatGPT による試験対策問題

問題 1

大量のデータを分析し, 特定の属性値を基にデータを分類する手法として決定木が用いられる. 決定木においてノード分割を行う際, 以下の評価指標がよく使われる. それぞれの指標の特徴を説明し, どのような場合に有効か述べなさい.

1. ジニ不純度 (Gini Impurity)
2. エントロピー (Entropy)
3. 分類誤差 (Classification Error)

(回答) ジニ不純度はノード不純度を表す指標で, 分類エラーの確率を直接的に表現できる利点がある. 全体の確率からデータを 2 回連続でとってきたときに, 同じクラスに所属している確率引いた以下の式で表現される. 不純度が高いほど 1 に近づき, 低いほど 0 に近づく. 決定木では不純度を最小にするように分割するための基準として使用される.

$$\begin{aligned} E_i &= \sum_{k=1}^{K} p_{ik} - \sum_{k=1}^{K} p_{ik}^2 \\ &= 1 - \sum_{k=1}^{K} p_{ik}^2 \end{aligned}$$

エントロピーは情報量を表す指標で, 確率分布が与えられたときに得られる. すべての事象が等確率で起きる場合 $(p_1, p_2, \ldots, p_n) = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ は最大をなり, 1 つの事象が確定で起きる場合 $(p_1, p_2, \ldots, p_n) = (1, 0, \ldots, 0)$ は最小になる. 決定木では親ノードと子ノードの情報量の差がなるべく大きくなるように分割される. 以下の式で表現される.

$$E_i = -\sum_{k=1}^{K} p_{ik} log(p_{ik})$$

分類誤差は正解ラベルと各ノードでの分類結果との誤差を分割の指標にするもの. ノードに含まれるサンプルのある特定のクラスに分類される確率を計算して, それを全体の確率から引いて誤差を計算をする. 最頻クラスを使うことが多い. 以下の式で表現される.

$$E = 1 - max(p_{ik})$$

ジニ不純度は計算コストが安く, 大きいデータセットでも学習が可能になる. エントロピーは安定した木生成するが計算コストが高い欠点がある.

問題 2

以下のデータセットに基づいて決定木を構築する場合, 最初の分割基準としてジニ不純度を用いる場合, どの属性を選ぶべきか計算しなさい.

属性 A	属性 B	属性 C	クラス
高い	是	男性	○
低い	否	女性	×
中間	是	男性	○
高い	否	女性	×
低い	是	女性	○

(回答) 属性 B を選ぶべき

属性 A の時

$$\frac{2}{5}(1-(\frac{1}{2})^2-(\frac{1}{2})^2)+\frac{1}{5}(1-(\frac{1}{1})^2-(\frac{0}{1})^2)+\frac{2}{5}(1-(\frac{1}{2})^2-(\frac{1}{2})^2)=\frac{2}{5}$$

属性 B の時

$$\frac{3}{5}(1-(\frac{3}{3})^2-(\frac{0}{3})^2)+\frac{2}{5}(1-(\frac{0}{2})^2-(\frac{2}{2})^2)=0$$

属性 C の時

$$\frac{2}{5}(1-(\frac{2}{2})^2-(\frac{0}{2})^2)+\frac{3}{5}(1-(\frac{1}{3})^2-(\frac{2}{3})^2)=\frac{4}{15}$$

問題 3

分類木の過学習を防ぐためには, いくつかの方法が提案されている. 以下の問いに答えなさい.

1. 過学習が発生する原因を簡潔に説明しなさい.
2. 過学習を防ぐために有効な手法を 3 つ挙げ, それぞれを説明しなさい.
3. 分類木における「剪定 (Pruning)」の具体的な方法を述べなさい.

(回答) 分類木における過学習とは木が分割によって成長する際に, 過剰に複雑な分割が行われてしまうこと. 必要以上に多くの特徴量を含むことで、過学習をすることがある. また訓練データにノイズが含まれる場合や, データ数が不足している場合も過学習をしやすい.

最大深度を固定する. 木の成長の深さを固定することで細分化し過ぎること抑えて過学習を防ぐ. 深度を制限することで, ノイズやランダムな変動に基づいた分割を抑制できる. 最大分割数を固定する. ノードを分割する際の必要な最小データ数を固定することで, 少数のデータポイントのための分割を防ぎ, 過学習のリスクを減少させる. 剪定を行う. 成長した木に対して, 必要のないと思われる分岐 (枝) を枝刈りすることで複雑なモデル化を簡素化し, 過学習を防ぐ.

剪定 (Pruning) の具体的な方法として, 評価関数と葉ノードの数からなるペナルティ項の和によって枝刈りをするかを決定する. 木の精度と複雑さはトレードオフの関係があるので, ペナルティ項にどれほどの重みをつけるかはクロスバリデーションなどで判断する. 他にも分割後の評価関数の改善量が一定の量を下回った場合に, その枝の成長を止める方法もある.

問題 4

分類木を用いて以下のデータを分析し, ユーザの購買行動を予測する場合, 以下のデータセットを参考に決定木のルールを作成しなさい. また, 作成したルールを説明しなさい.

年齢	収入	学歴	購入
20 代	高い	大学卒	はい
30 代	中間	高校卒	いいえ
40 代	高い	大学卒	はい
50 代	低い	中学卒	いいえ
60 代	中間	高校卒	いいえ

計算の簡単さからジニ不純度を評価関数として採用する. データセットから収入と学歴だけで購買行動を分類できそうなので, 最大深度を 2 に固定する. ジニ不純度の比較から収入と学歴どちらも選んでも変わらないので, ここではまず収入で判断し高い場合は購入, それ以外は学歴によって判断するような木とする.

収入

$$\frac{2}{5}(1-(\frac{2}{2})^2-0)+\frac{2}{5}(1-0-(\frac{2}{2})^2)+\frac{1}{5}(1-0-(\frac{1}{1})^2)=0$$

学歴

$$\frac{2}{5}(1-(\frac{2}{2})^2-0)\frac{2}{5}(1-0-(\frac{2}{2})^2)\frac{1}{5}(1-0-(\frac{1}{1})^2)$$