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 $\{(\mathbf{x}_i, y_i), i = 1, 2, \dots n\}$ について, ラベル y_i が実数値の場合は回帰木, 分類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, 除法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 $(p_1, p_2, \dots p_n) = (\frac{1}{n}, \frac{1}{n}, \dots \frac{1}{n})$, 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 $(p_1, p_2, \dots p_n) = (1, 0, 0, \dots, 0)$, 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} 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

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

分割基準