機械学習勉強メモ --- 決定木編

決定木とは

- 意味解釈可能性(結果の意味が解釈しやすい)が特徴的な機械学習モデルである.
- データセットの特徴量に基づいて,クラスを分類するための一連の質問を学習し,分類する.
- 情報利得(親と子の不純度の差)が最大となる特徴量でデータを分割する.
- 情報利得IGは以下の式で定義される

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{n} \frac{N_j}{N_p} I(D_p)$$

- fは分割を行う特徴量，$I(D_p)$は親のデータ，D_jはj番目の子のデータを指す.
- Iは不純度，N_pは親ノードのデータの総数，N_jはj番目の子のデータの総数である.
- また，2値分類の場合は以下の式で定義される

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

- D_{left}は左の子のデータ，D_{right}は右の子のデータである.
- 不純度は，異なるクラスのデータがそのノードにどの程度混ざっているかを定量化した指標
- 不純度はジニ不純度，エントロピー，分類誤差の3種類ある.
    - ジニ不純度
        - ジニ不純度は誤分類の確率を最小化する条件
        - ジニ不純度(I_G)は以下の式で定義される

$$I_G(t) = \sum_{i=1}^{c} p(i|t) \times (1 - p(i|t))$$

        - $p(i|t)$は，特定のノードtにおいてクラスiに属しているデータの割合
    - エントロピー
        - 相互情報量が最大化するように試みる条件である
        - エントロピー(I_H)は以下の式で定義される

$$I_H(t) = -\sum_{i=1}^{c} P(i|t) \times log_2 P(i|t)$$

    - 分類誤差
        - 決定木の剪定に役立つ条件
        - ノードのクラス確率の変化に敏感でないため，決定木の成長には適さない

- 分類誤差(I_E)は以下の式で定義される

$$I_E(t) = 1 - maxP(I|t)$$