

scikit-learn ゼミ

1.8.4. Complexity

小野寺喜行

Run time cost

- ・一般的にバランスの取れたバイナリツリーを構築する実行時間

$$O\left(n_{samples} n_{features} \log\left(n_{samples}\right)\right) \text{ and } O\left(\log\left(n_{samples}\right)\right)$$

- ・ツリー構築アルゴリズムは常にはバランスが取れたツリーを生成できない

Run time cost

- ・サブツリーがバランスの取れたままだと仮定



各ノードのコストはエントロピー最大の削減を提供する特徴を見つけるための

$$O(n_{features})$$

がかかる探索で構成

Run time cost

・各ノードのコスト

$$O\left(n_{features} n_{samples} \log\left(n_{samples}\right)\right)$$



ツリー全体の総コスト

$$O\left(n_{features} n_{samples}^2 \log\left(n_{samples}\right)\right)$$

Decision trees on Scikit-learn

- Scikit-learnでは効率的な決定木構築の実装を提供
- 単純な実装では与えられた特徴に沿って、
新しい分割点のための
 クラスラベルヒストグラム(分類用)
 or
 手法(回帰用)
を再計算

Decision trees on Scikit-learn

- ・全ての関連サンプル上で特徴を事前ソート
及び、実行中のラベルカウントを保持



$O(n_{features} \log(n_{samples}))$ に各ノードでの複雑さを軽減

$O(n_{features} n_{samples} \log(n_{samples}))$ の総コストになる