多様な解の発見問題のAIと大規模データ解析への展開

A02班 有村博紀（北大）

最近の機械学習と大規模データ解析分野では…

- 複数の最適解を求めることに関心が高まっている (2010s～)
- 従来は，ただ一つの最適解を求める技術に興味(1990～2010)

文字列解析（ゲノム解析）

- 極大部分列問題（MCS）の多様解

機械学習

- 木編集距離(TED)に基づく最適決定木の多様解

定式化: 計算問題Πの多様性問題

入力:

計算問題Πの入力例 I, 正整数k

タスク:

解集合 $Sol_\Pi(I)$ 中のk個の解の組合せ $X_1, \ldots, X_k$ で相互距離 $Dist(X_1, \ldots, X_k)$ を最大化するものを見つけよ

問題の超パラメータ:

- 集合 $Sol_\Pi(I)$
- 二つの解の距離 $dist(\bullet, \bullet)$

解の相互距離の例

Max-Min

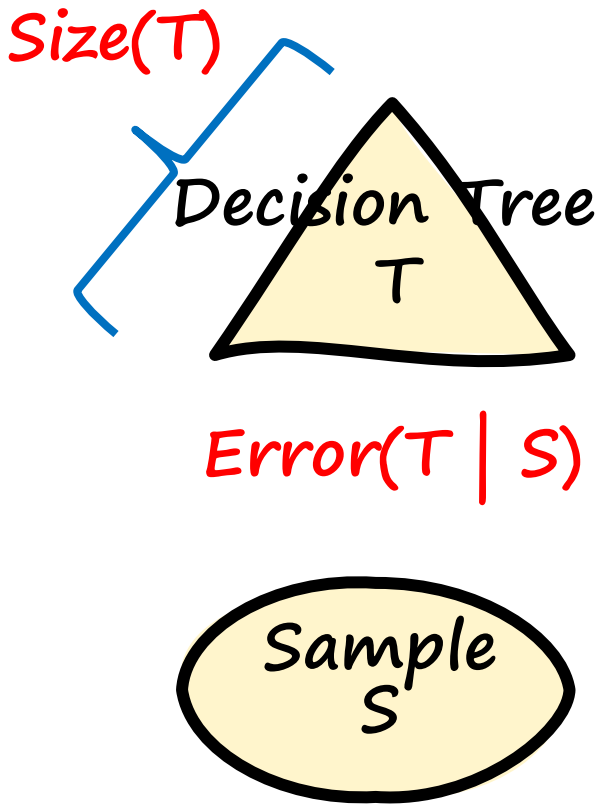
$$Dist(X_1, \ldots, X_k) = \min_{1 \le i < j \le k} dist(X_i, X_j)$$

Max-Sum

$$Dist(X_1, \ldots, X_k) = \sum_{1 \le i < j \le k} dist(X_i, X_j)$$

各問題領域ごとにどう決めるかが大事

多様性問題の例：最適決定木発見（機械学習）



入力：

入力例 $I = (X, Y, S, s, e)$,
where 変数集合X, ラベル集合, 分類例集合S, 整数$s \geqq 1$, $e \geqq 0$

問題の超パラメータ：

- 解集合 $Sol_{\Pi}(I)$: サンプルS上で，小さく精度の高い決定木の全体

$Sol_{\Pi}(I) = \{ T \text{ in } H_{DT} : size(T) \leqq s, err(T, S) \leqq e \}$

- 二つの解TとT'の距離 $dist(T, T')$: TとT'の間の木編集距離 (TED).
  ただし，ラベル付き二分木として．

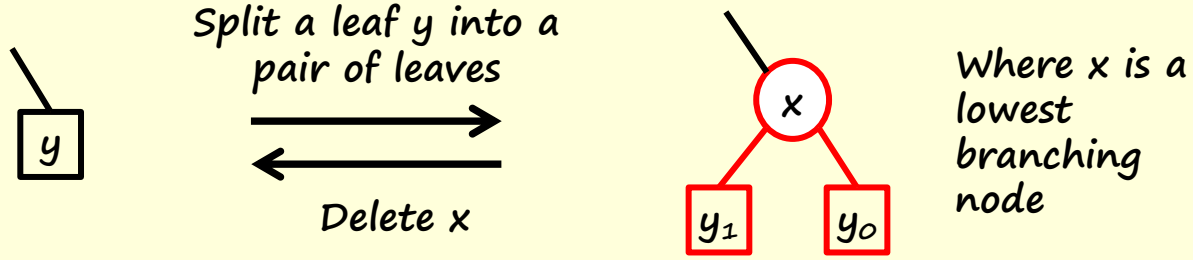
決定木 TとT'の間の木編集距離 (BinTED).

議論のたたき台

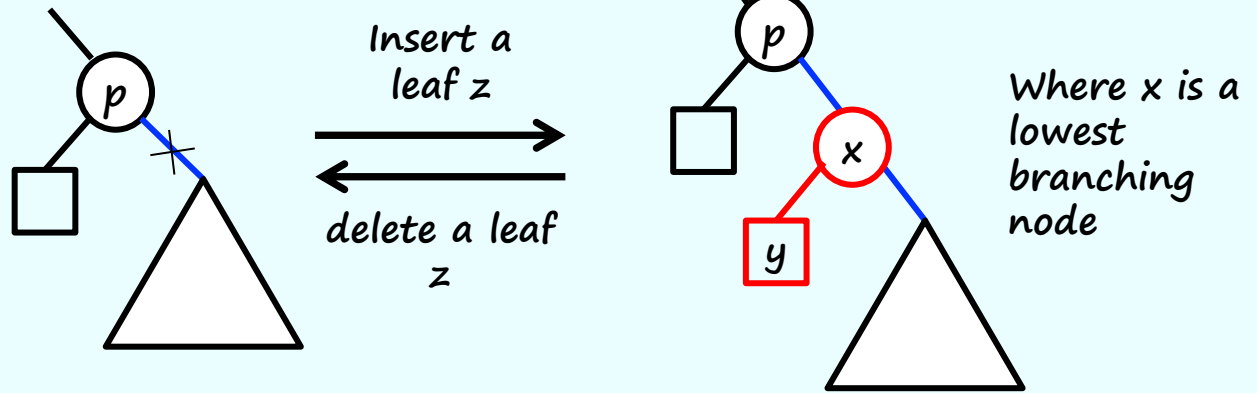
Notes: 演算Op 1とOp 3は，80sの決定木の貪欲学習法で広く用いられた木の精錬方法

Def. 二分木の編集演算（5個）

Op 1. Leaf Split / Opt 2. Leaves merge　Cost +1

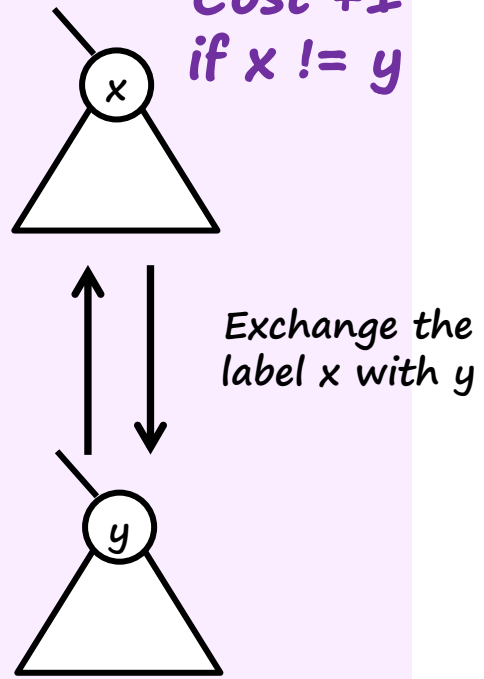


Op 3. Edge split / Opt 4. Edge contraction　Cost +1



Op 5. Label exchange

Cost +1 if x != y



- Def. 上記の演算をくり返し適用して，TからT'が得られるとき，T'はTの詳細化/refinement (TはT'の汎化generalization)であるという

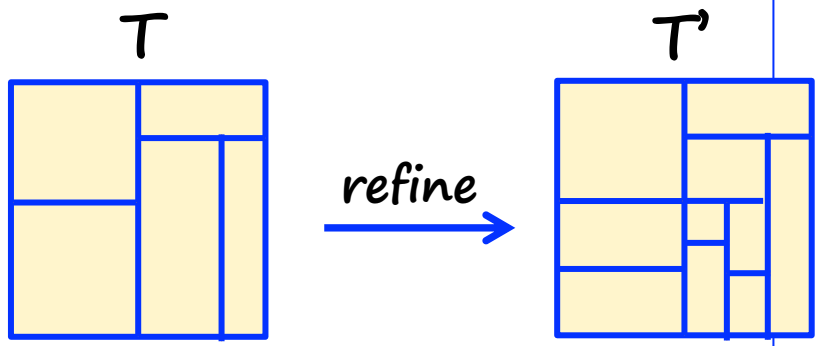
決定木 TとT'の間の木編集距離 (BinTED).

Def. 二分木TとT'の木編集距離 $d_{ED}(T, T')$

TをT'に変換する全ての編集操作の列 $s = (o_1, \ldots, o_m)$ に対するedit costの和 $s = cost(o_1) + \ldots + cost(o_m)$ がとる最小の値

性質(二分木TとT'の木編集距離)

- 1. $d_{ED}(T, T') \leqq |T| + |T'|$
- 2. T'がTの精密化 (refinement) と仮定する.
  任意のサンプルSに対して, 次が成立する:



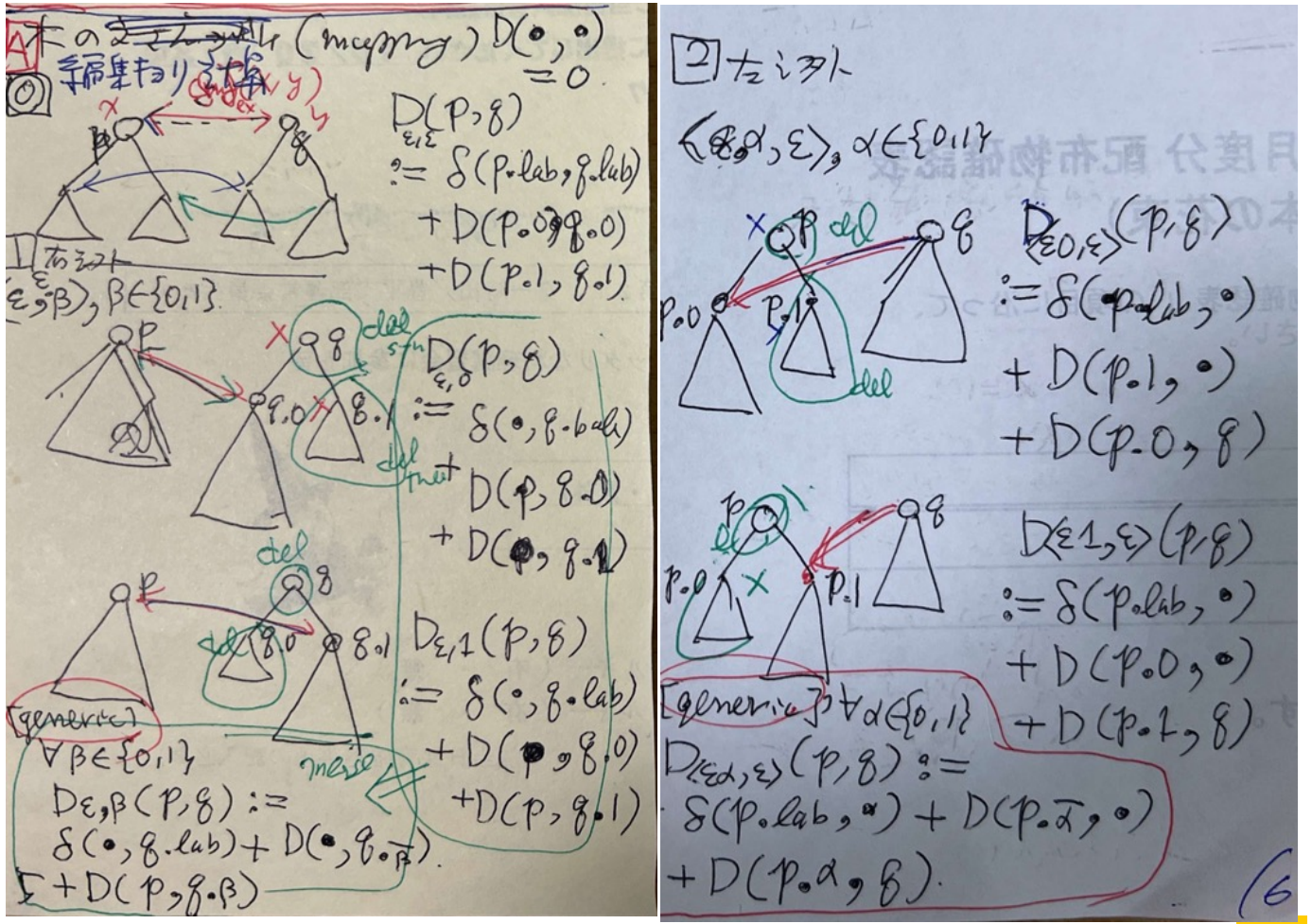
a. T'が生成する入力領域分割は, Tが生成する領域分割の細分化になっている.
（入力領域$\{0,1\}^n$の分割を生成するモデルとして, より詳細なモデルを得られる）

b. T'の葉のラベル付けで, $error(T, S) \geqq error(T', S)$が成立するようなものがある. ただし, 内部ノードのラベル付は任意である. (エラーが悪くならない)

二つの決定木の編集距離を求めるアルゴリズム

Thm. 動的計画法に基づいて，$O(n^3)$時間と$O(n^2)$領域で編集距離$d_{ED}(T, T')$を求めることができる

- 入力：決定木 TとT'
- 出力：$D(T_{root}, T'_{root})$ = TとT'の編集距離
- $D(p, q)$：次の二つの部分木の編集距離
  - ノードpを根とするTの部分木T_p
  - ノードqを根とするT'の部分木T'_q



多様な最適決定木発見（機械学習）

木の編集距離に基づく最大多様性

Max-Min

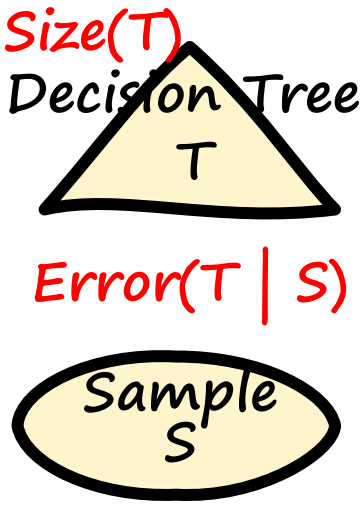
$$Dist(X_1, \ldots, X_k) = \min_{1 \le i < j \le k} d_{ED}(X_i, X_j)$$

Max-Sum

$$Dist(X_1, \ldots, X_k) = \sum_{1 \le i < j \le k} d_{ED}(X_i, X_j)$$

入力：

入力例 $I = (X, Y, S, s, e)$; $k \geqq 1$,
where 変数集合 X, ラベル集合 Y,
分類例集合 S, 整数 $s \geqq 1$, $e \geqq 0$;
正整数 k



タスク

- サイズ（葉数 or ノード数）が高々 s で，経験誤差が高々 e の決定木からなる解集合 $Sol_{OptDT}(I)$ 中の k 個の決定木の組合せ $T_1, \ldots, T_k$ で相互距離 $Dist(X_1, \ldots, X_k)$ を最大化するものを見つけよ
- とりあえず，決定木の対（$k=2$）を考える

今後の方針：多様な最適決定木発見

- 当面の目標：多様な最適決定木の対（$k=2$）を見つける問題の $n^{O(s)}$ 時間と $O(poly(k,e))$ アルゴリズム
  - ブール変数の数n, 最大サイズs, 最大誤差数e.

- 別のやさしい問題（文字列解析）
  - 編集距離に基づく極大部分列問題（MCS）のk-多様解
  - とりあえず解けた問題：ハミング距離の最長部分列の多様解（志田・小林・有村, COMP研2023.12予定）

- 最適決定木の作業目標：最初に次の部分問題を考える
  - 入力として, サイズが高々sで, 経験誤差が高々 e の(s, e)-最適決定木の一つT（参照モデル*reference model*）が与えられると仮定する.
    その上で参照モデルTに対して, 編集距離 $d_{ED}(T, T')$ を最大化する(s, e)-最適決定木T'を一つ見つける問題.
  - $d_{ED}(T, T')$ を最小化する問題は, 機械学習システム運用で有名な問題

多様な解の発見問題のAIと大規模データ解析への展開

最近の機械学習と大規模データ解析分野では…

- 複数の最適解を求めることに関心が高まっている (2010s～)
- 従来は, ただ一つの最適解を求める技術に興味(1990～2010)

文字列解析（ゲノム解析）

- 極大部分列問題（MCS）の多様解

機械学習

- 木編集距離(TED)に基づく最適決定木の多様解

ご清聴ありがとうございました