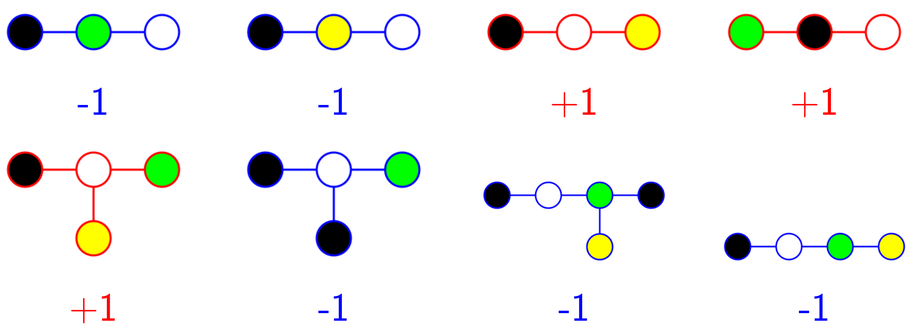
全部分グラフ指示子に基づく決定木勾配ブースティング

北海道大学 修士2年 横山侑政

グラフの教師あり分類問題

入力　クラス $(y = +1 or -1)$のわかるグラフ N 個



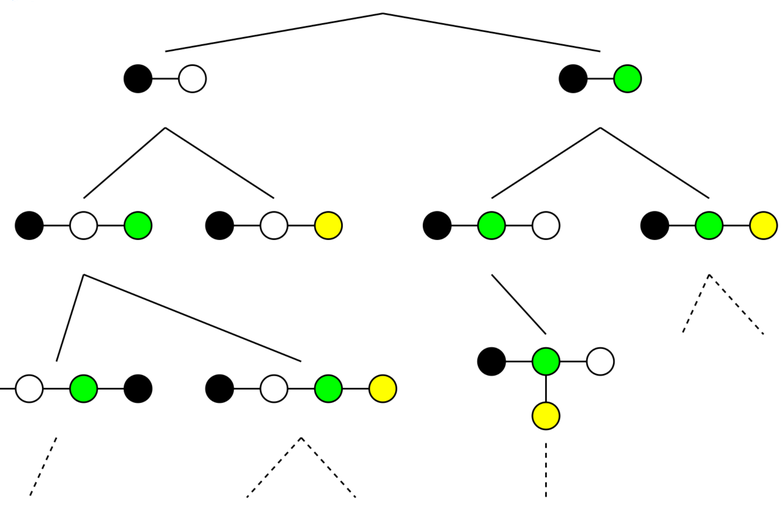
出力　グラフ G が +1か -1を予測する予測モデル F

特徴量　部分グラフ有無

	x_1	x_2	x_3	x_4	$\cdots$
G_1	1	0	1	0	$\cdots$
G_2	0	0	0	1	$\cdots$
G_3	1	1	0	0	$\cdots$
G_4	1	1	0	1	$\cdots$

部分グラフの有無の計算

探索木を考えて
深さ優先探索



親ノードは
子ノードの
部分グラフ

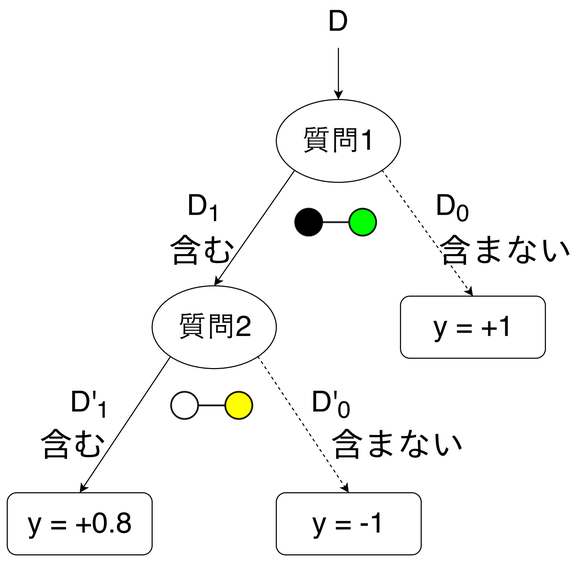
問題点　部分グラフの数が膨大
全ての部分グラフの有無を計算するのは困難

対処法　陽には全ての部分グラフを計算せず
決定木の勾配ブースティングモデル F を学習する

決定木の勾配ブースティング

決定木

入力データに
内部ノードで質問をして
葉ノードで予測値を返す



勾配ブースティング

決定木に加法的に学習する

$$F(G) = T_0(G) + sT_1(G) + sT_2(G) + sT_3(G) + \cdots \tag{1}$$

T_k は残差 r_i を予測する回帰木

$$r_i = \frac{\partial \mathrm{L}(y_i, F_{k-1}(G_i))}{\partial F} \tag{2}$$

s はシュリンケージ係数，L は損失関数

内部ノードの学習

内部ノードは二乗誤差和を最小にする部分グラフを学習する

$$\arg \min_{x_j \in X} \left[\mathrm{SSE}(D_1(x_j)) + \mathrm{SSE}(D_0(x_j))\right] \tag{3}$$

X：全ての部分グラフの集合 (数が膨大で解けない)
$D_1(x_j)$：$\{G_i : x_j$を含む$\}$，　$D_0(x_j)$：$\{G_i : x_j$を含まない$\}$
$\mathrm{SSE}(D)$：sum of square error，　$\min_m \sum_{G_i \in D}(r_i - m)^2$
データ集合 D の残差の二乗誤差の合計

部分グラフの有無の性質

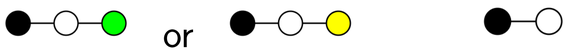
部分グラフの有無の性質

部分グラフ x_j と その拡大グラフ x'_j を考えたとき
あるグラフG_iに 部分グラフ x_j が無いならば 部分グラフ x'_j も無い

数式　$x_j \subset x'_j$ として　$G_i \not\supseteq x_j \Rightarrow G_i \not\supseteq x'_j$

逆に言えば

あるグラフ G_i が
部分グラフ x'_j を持つならば 部分グラフ x_j も持つ



部分グラフの有無の計算の枝刈り

部分グラフの有無の計算の枝刈り

式(3)を解きたいが全て部分グラフの集合 X が大きすぎて計算困難

すでに有無を計算した部分グラフ x_j の拡大グラフ x'_j での
二乗誤差の下限値が分かれば枝刈りできる

二乗誤差和の下限値

部分グラフの有無の性質を使うと以下がいえる

$$\min_{x'_j \in X_{\supset} x_j} \left[\mathrm{SSE}(D_1(x'_j)) + \mathrm{SSE}(D_0(x'_j))\right] \tag{4}$$

$$= \min_{S \subset D_1(x_j)} \left[\mathrm{SSE}(D_1(x_j) \setminus S) + \mathrm{SSE}(D_0(x_j)) \cup S\right] \tag{5}$$

$X_{\supset} x_j : \{x'_j \mid x'_j \in X, x'_j \supset x_j\}$　　x_j の拡大グラフの集合

二乗誤差和の下限値の計算

式(5)は組み合わせを考える必要があるが
以下の等式を示すことで計算が簡単になる
(前からの部分集合 と後ろからの部分集合 を考えればよい)

$$\min_{C \subset A} \left[\mathrm{SSE}(A \setminus C) + \mathrm{SSE}(B \cup C)\right] \tag{6}$$

$$= \min \begin{pmatrix} \min_{k=2,\ldots,|A|-1} \left[\mathrm{SSE}(A_{\leq\, k}) + \mathrm{SSE}(B \cup A_{>\, k})\right], \\ \min_{k=2,\ldots,|A|-1} \left[\mathrm{SSE}(A_{>\, k}) + \mathrm{SSE}(B \cup A_{\leq\, k})\right] \end{pmatrix} \tag{7}$$

$A = \{a_1, \ldots, a_n\}$,　$A_{\leq\, k} = \{a_1, \ldots, a_k\}$,　$A_{>\, k} = \{a_{k+1}, \ldots, a_n\}$
$B = \{b_1, \ldots, b_m\}$
$\mathrm{SSE}(A) = \min_m \sum_{a_i \in A}(a_i - m)^2$　　集合 A の二乗誤差の合計

証明

$$\mathrm{SSE}(A \setminus C) + \mathrm{SSE}(B \cup C) \tag{8}$$

$$= \sum_{i=0}^{n}(a_i - \bar{a}')^2 - \sum_{i=0}^{k}(c_i - \bar{a}')^2 + \sum_{i=0}^{m}(b_i - \bar{b}')^2 + \sum_{i=0}^{k}(c_i - \bar{b}')^2 \tag{9}$$

$$= -\sum_{i=0}^{k}(c_i - \bar{a})^2 - \frac{\left(\sum_{i=0}^{k}(c_i - \bar{a})\right)^2}{n-k} + \sum_{i=0}^{n}(a_i - \bar{a})^2$$
$$+ \sum_{i=0}^{k}(c_i - \bar{b})^2 - \frac{\left(\sum_{i=0}^{k}(c_i - \bar{b})\right)^2}{m+k} + \sum_{i=0}^{m}(b_i - \bar{b})^2 \tag{10}$$

$$= -\left(\frac{1}{n-k} + \frac{1}{m+k}\right)\left(\sum_{i=0}^{k} c_i\right)^2 + \left(2\bar{a}\frac{n}{n-k} - 2\bar{b}\frac{m}{m+k}\right)\sum_{i=0}^{k} c_i \tag{11}$$

$$-\frac{nk}{n-k}\bar{a}^2 + \frac{mk}{m+k}\bar{b}^2 + \sum_{i=0}^{n}(a_i - \bar{a})^2 + \sum_{i=0}^{m}(b_i - \bar{b})^2 \tag{12}$$

k を固定したとき上に凸な $\sum_{i=0}^{k} c_i$ の二次式

$\bar{a} : A$の平均，　$\bar{a}' : A \setminus C$の平均，　$\bar{b} : B$の平均，　$\bar{b}' : B \cup C$の平均，　$k = |C|$