Confidential

直感的に分かる勾配ブースティング入門

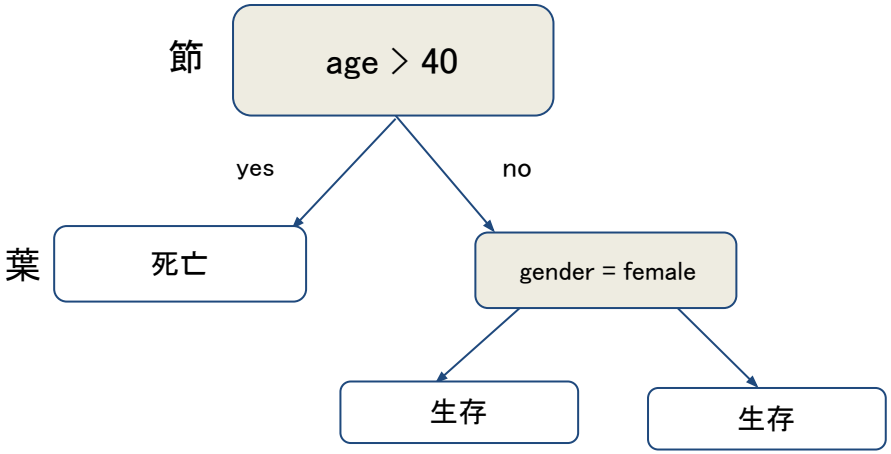
奥本翔
株式会社GRI

データで新たな事業を開発していくカンパニー。

決定木とは？

- 特徴量の値に応じて、それぞれのサンプルを適切な葉に割り当てる

- 割り当てられた葉に応じてサンプルを推定・分類する
  - 主に、分類木・回帰木の2種が存在

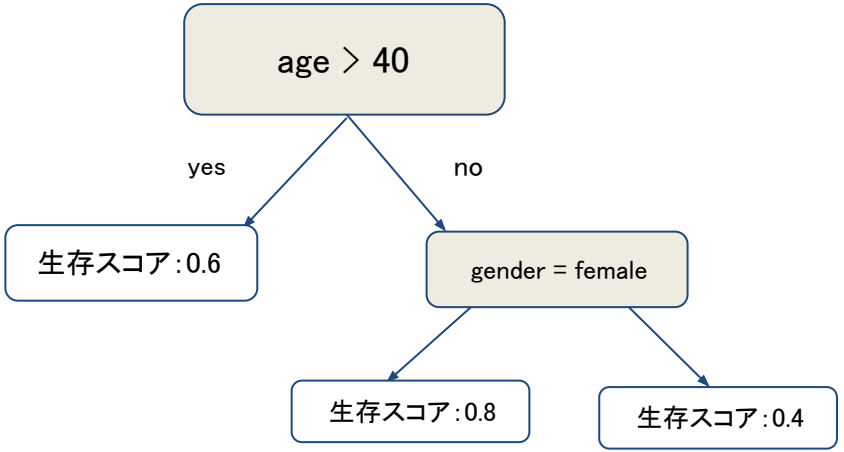


決定木を用いた分類の例

- タイタニック乗客の生存・死亡を分類するタスク
  - kaggle等、コンペの入門編で有名

- 決定木は0~1に値を取る生存スコア(生存確率に近いもの)を出力する
  - ここでは、一般的に用いられる閾値0.5の場合で説明する

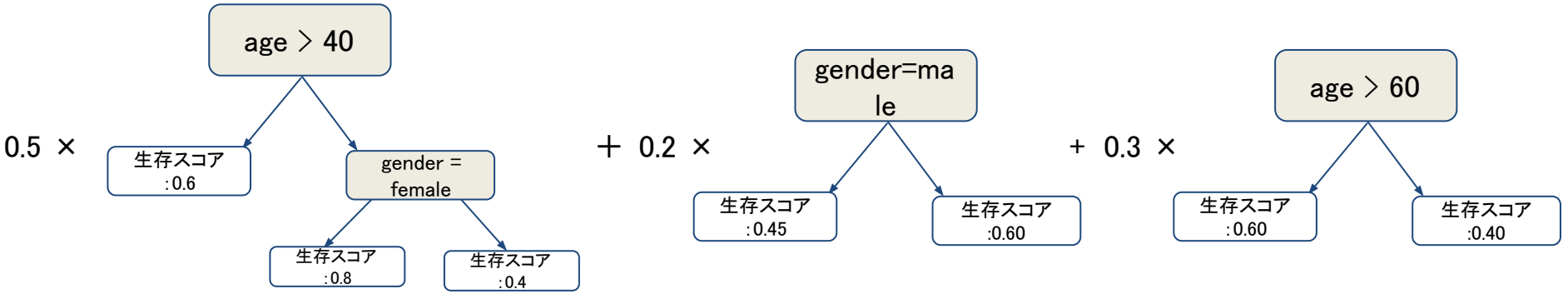
Id	age	gender	score	pred
1	30	female	0.8	生存
2	20	male	0.4	死亡
3	45	male	0.6	生存



決定木のアンサンブル

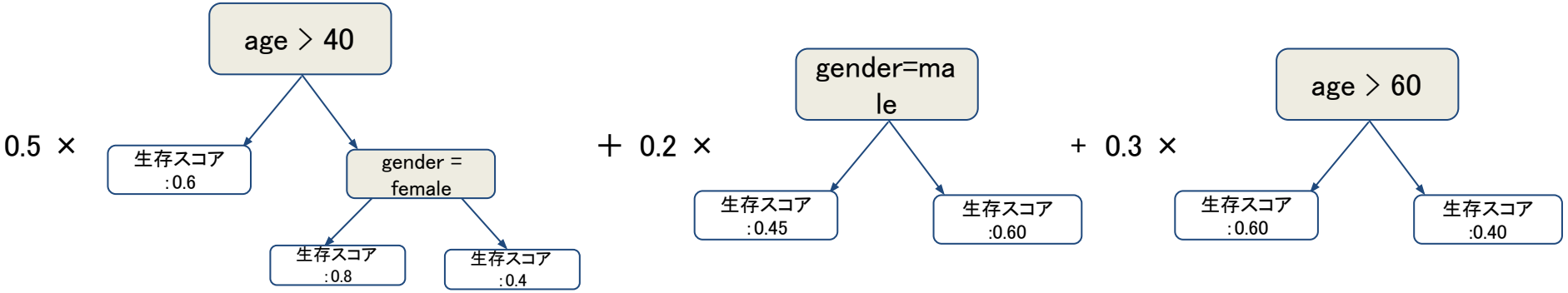
- 単一の決定木だけでは、十分な精度が出せないケースが多い
  - そこで、一般的には、アンサンブルと呼ばれるテクニックが用いられる

- 直感的には。決定木の多数決で推定・分類を行う方法
  - 単一の決定木では高い精度を出せないが、多数決で決定すれば強くなる



数値例を用いた説明

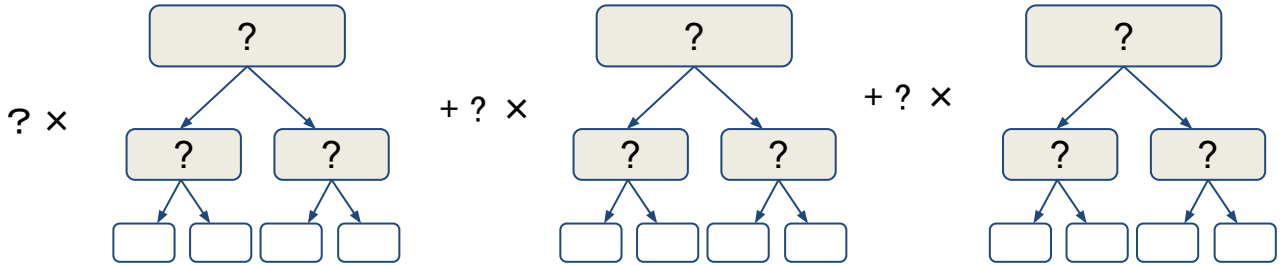
- ageが30、genderがmaleのサンプルの場合
  - スコアは、右の木から順にそれぞれ、0.4、0.45、0.4となる
  - 下図のようにアンサンブルする場合、最終的なスコアは、
    (0.5 × 0.4 + 0.2 × 0.45 + 0.3 × 0.4)で、0.41となる



決定木の学習方法

- アンサンブルをどのように学習するか？
  - 決定木・アンサンブルのための比率の２点を決めれば良い
  - ここを決めるアルゴリズムの一つが勾配ブースティング

- 節で用いる特徴量・閾値、アンサンブルの比率は教師データから学習される

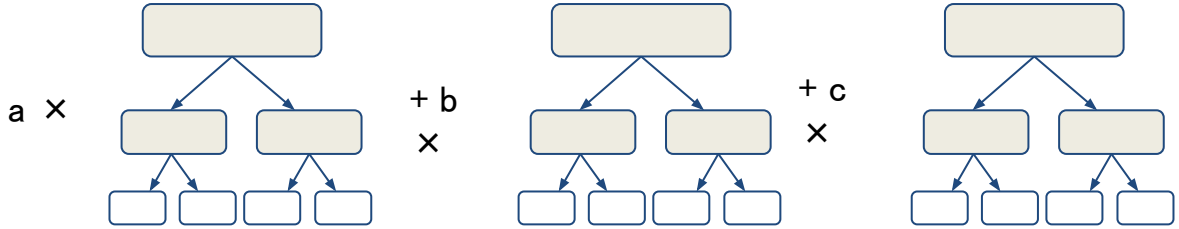


勾配ブースティングによる学習

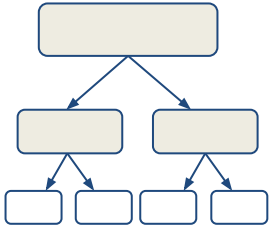
- 勾配ブースティングアルゴリズムでは、次のステップを繰り返す
  1. 追加すべき決定木・その比率を計算
  2. 前のステップまでに出来上がったアンサンブル決定木にそれを足す
  3. Step.1に戻る

- 直感的には、Step2で生成される決定木を勾配=進むべき方向と言える

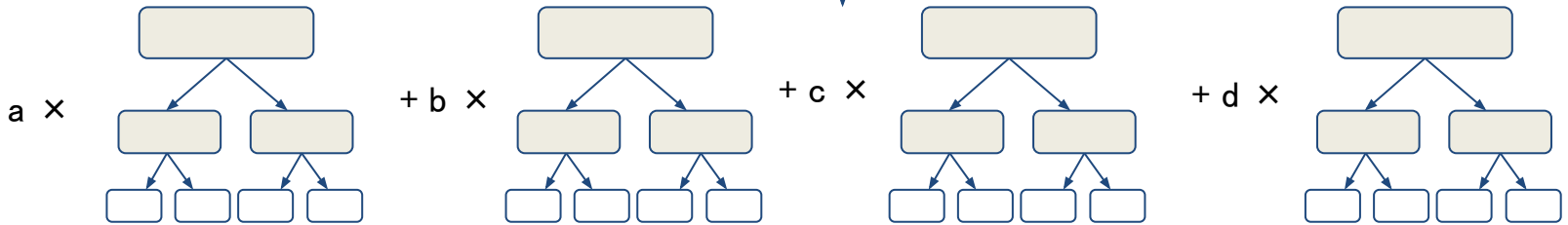
学習プロセスの図解



重みdと新たな決定木:



を作成



表データの学習で勾配ブースティングが多用される理由

- 表形式のデータを伴う予測タスクにおいて、勾配ブースティングはファーストチョイスとなることが多い
  - 画像認識や音声認識などでは深層学習系のモデルが強い

- 複雑過ぎず、単純過ぎず
  - 複雑なモデルを使うと計算量・過学習の危険性といったコストが発生する
  - 一方で、単純すぎるモデルを用いれば精度を確保することができない

- 画像認識・音声認識等と比較して、表を用いたタスクでは、解釈のしやすさが重要
  - 深層学習の解釈は難しい
  - 決定機ベースのアルゴリズムでは、重要特徴量の導出が比較的簡単にできる