

## 3.4 節 「機械学習による予測・判断」

### 補足資料

(2024 年 8 月 30 日版)

#### A.3.4.1 ランダムフォレスト

汎化性能の高くない機械学習モデルを多数集めることで、性能の高い機械学習モデルを実現しようとするアンサンブル学習というアプローチがある。本文で解説した決定木を、アンサンブル学習の一種であるバギングという手法で多数生成するのが**ランダムフォレスト** (random forest) である。

まず、データからランダムに元のデータと同じ個数のデータを復元抽出する。復元抽出したデータに対して統計的推測をおこなう手法をブートストラップという。バギングでは、統計的推測をおこなうかわりに、異なるデータで決定木のような機械学習モデルの学習をおこなって複数のモデルを作成し、それらの平均を予測とすることで汎化性能の向上を図っている<sup>2)</sup>。

ランダムフォレストの学習においても、復元抽出したデータに対して決定木の学習をおこなう。ただし、各分割をおこなうたびに分割に使う変数の候補もランダムに選択する。ランダムに復元抽出したデータで学習するとはいえ、同じデータから抽出されたデータで学習するため、得られる決定木は似たものとなってしまう。多数の決定木を作ることによる性能の向上は限定的である。そこで、分割に使う変数の候補を毎回ランダムに選ぶことで、より多様な決定木を生成しようとしている。どの程度の数の変数を候補として選択するかは性能を左右するパラメータではあるが、説明変数が  $p$  個ある場合、 $\sqrt{p}$  個あるいは  $p/3$

---

<sup>2)</sup> 多数のモデルの平均をとることで、補足資料 A.3.3.1 項で解説したモデルの分散を低減することを意図している。

個の変数を選択することが多い。

ランダムフォレストを適用した場合の模式図を図 A.3.2 に示す。通常、ランダムフォレストでは大量の決定木を生成するが、ここでは簡単のため、決定木を3つ生成した場合を考える。また、復元抽出によって同一のデータが複数抽出された場合はわずかにずらして表示している。図中の緑の点のカテゴリを予測してみよう。ランダムフォレストによる予測は、復元抽出と学習を繰り返して得られた多数の決定木のうち、あるカテゴリを出力した決定木の割合を、そのカテゴリに属する確率の予測として出力する。一番左の決定木では変数  $x$  が  $t$  より大きいのでカテゴリ A と予測される。真ん中の決定木では変数  $y$  が  $t$  より大きいのでカテゴリ A が予測される。一番右の決定木では変数  $x$  が  $t$  より小さいのでカテゴリ B が予測される。3つの決定木のうち、2つがカテゴリ A と予測しているので、カテゴリ A である確率は  $2/3$  となる。

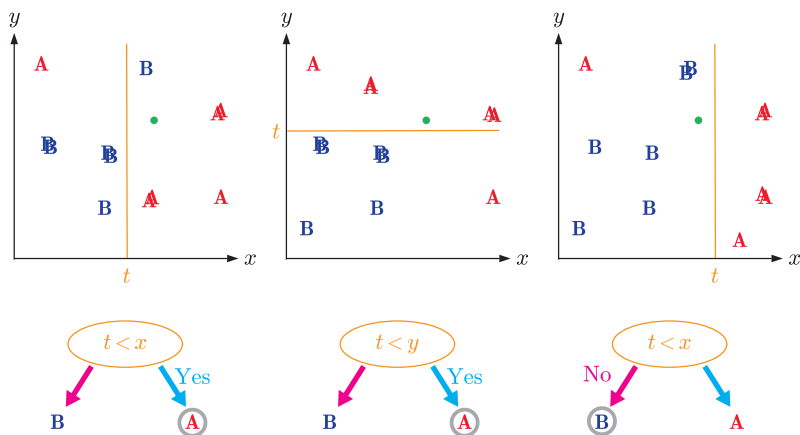


図 A.3.2 ランダムフォレストによる予測

決定木では分割を繰り返すほど複雑なモデルとなることを先に述べた。したがって、決定木の分割を終了する基準はモデルの複雑さ、そして汎化性能を決める重要なハイパーパラメータである。このようなハイパーパラメータはホールドアウト法や交差検証法によるモデルの評価に基づいて調整されることが多いが、ランダムフォレストでは、学習データと検証データを分割することなく同様の評価をおこなうことができる。各決定木は復元抽出したデータで学習するた

め、それぞれの決定木に対して学習に使われていないデータが存在する。こうしたデータを検証データとして利用することができる。各データについて、学習に使われていない決定木の出力の平均値によって、カテゴリの分類確率や量的データの予測値を計算する。すべてのデータに対して予測値を計算して、誤分類率や RMSE を計算すれば、学習に使っていないデータにモデルを適用したときの性能評価をおこなうことができる。