クロスバリデーション法を用いた決定木モデル作成

BV19063　菅原 有一

令和元年 11 月 1 日

1　研究背景

近年, 情報の価値はとてつも高くなってきている. またそれに伴い情報の量も増え, その情報がどのような事柄を表しているか, どのように処理をすれば求めている情報を手に入れられるかなどを研究する職業である” データサイエンティスト” と呼ばれる職業までできた. 以前からこの分野に興味があり, 先日データ分析に関する内容を勉強する機会があったので, そこで学んだことを踏まえて精度の高い予測モデルの作成について考察しようと思う.

2　データ分析

データ分析とは, 過去の蓄積データから未知のデータに含まれていない任意の事柄についての結果を予測するモデルを作り, その事柄について推測することである. またその精度のことを「モデルの精度」という.

3　モデルの精度

蓄積データにおける予測モデルの精度は評価指標があり, それを用いることで測れるが, 未知のデータに当てはめた際, 蓄積データの予測精度との間にしばしば乖離が発生する. したがって,

$$(\text{モデルの精度}) = (\text{蓄積データの予測精度}) \cdot (\text{乖離幅})$$

として考える必要がある.

4　評価指標

評価指標は, 個数, 分類, 順序などがありそれらに対していくつかの評価指標が知られている. 今回は AUC という評価指標を用い考察していく. AUC は主に医療診断や契約の有無などの 2 群分類の時に用いる. また値域は $0 \sim 1$ で 1 に近いほど精度がよいものとなっている.

5　AUC の精度評価方法

AUC と呼ばれるモデルの精度評価指標は, 真陽性率, 偽陽性率の変化をグラフで表した ROC 曲線と呼ばれる曲線の面積のことである. また, ROC 曲線とは縦軸を真陽性率, 横軸を偽陽性率とし, 閾値を $+\infty$ から下げた時の変化を線分でつないだ曲線である. この評価指標を用いて予測モデルを構築していく.

6　決定木モデル

AUC を評価指標として用い今回予測モデルとして作成する決定木モデルとは, データを全説明変数に対して閾値を虱潰しに探索して作るモデルである. メリットとしては, 図で表現できるため, 結果に対する根拠が明確になりやすい, 副次的なものではあるが交互作用が現れることがあるなど, 現在でもビジネスなどの場面で使われることの多いモデルである, しかし, 複数変数の線形結合には弱く, 精度を高めるために自分で新たな説明変数を作成するなどの工夫を必要とする場合もある. 決定木モデルを作るうえでのパラメータは, 木の深さ (maxdepth), 最小ノードサイズ (minnucket), 枝刈りの強さ (cp) の 3 つで, また分割するときの指標となる分割スコアの計算方法として, Gini と Entoropy の 2 種類があり, 以上から AUC が最大値をとる組み合わせを考えていく.

7　過学習

決定木モデルについて蓄積データをもとに AUC が最高になるように作っていくとやがて値域の最大値である 1 に限りなく近づく, または一致するが, その予測モデルを未知のデータに当てはめると AUC の値が蓄積データの時と比べ低いものとなってしまう. これは過学習といい, 構築データのみに現れるような分岐をしてしまい, 未知データとの違いにより評価指標の値が低くなってしまう. これを解決するために未知のデータに対して精度を検証する方法として, ホールドアウト法, クロスバリデーション法がある. ここではクロスバリデーション法を用いることとする.

8　クロスバリデーション法

クロスバリデーション法とは, 答えのわかっているデータを K 分割し, そのうち 1 個をモデル検証データとし, 残りをモデル構築データとして K 回繰り返してモデルを構築し, 未知のデータに対しての評価指標を推し量る方法である. またこれにより十分にデータが多くない場合でも精度を検証できる.

9　予測モデル作成結果

以上を踏まえ, R 言語を用いて予測モデルを作成した.[*1] 結果, 最大の AUC を出したパラメータは, maxdepth = 17, minnucket = 13, cp = $1e-10$, 分割スコアは Entoropy を用いて計算したものとなった. また, 提出結果としては 0.9178 という結果であった.

10　今後の課題

今回出た結果に関しては, ただパソコンで AUC の最大値を出す 3 つのパラメータ, 分割スコアの計算方法に関してプログラムを動かして探索しただけにすぎず, それぞれのパラメータや分割スコアの計算方法に関しての考察ができなかった. また, 今回は蓄積データに関しては一切触っておらず, 評価に関係のあまりない情報などの分別もできなかった. 以降はそれぞれの関係性や大きさによる影響, データの必要性などを考察しようと思う.

11　謝辞

今回の発表を行うにあたり, 2019 年 10 月に開催されたデータ分析講座の内容を引用いたしました. ご指導をいただいた数理科学研究会会長の佐野遼太郎氏に感謝の意を表します.

[*1] 蓄積データは SIGNATE(https://signate.jp) の【練習問題】銀行の顧客ターゲティングを用いた