重回帰分析と回帰樹木の比較

2005MM096 矢野　赳
指導教員：田中　豊

1　はじめに

データを式やモデルで近似して予測モデルを構築する際，回帰分析が有効なことは広く知られているが，これらとは異なる手法がある．目的変数に影響する説明変数を２股に枝分かれする形で予測あるいは判別を行う回帰木・分類木である．２つの実データや，線形・非線形の２つの人工データを用いて，回帰樹木と回帰分析の分析結果を比較し，検討することを目的とする．分析には統計解析ソフトＲを用いた．

2　樹木モデル

樹木モデル（tree-based model）は，非線形回帰分析，判別分析のひとつの方法で，回帰の問題では回帰木（regression tree），分類の問題では分類木（classification tree），あるいは決定木（decision tree）と呼ばれている．樹木モデルは，説明変数の値を分岐させ，それらを組み合わせて，判別・予測のモデルを構築する．しかし，このルールで木を成長させ続けると，一般に木が大きく成長し過ぎ複雑になる．そこで，不適切な枝を剪定し，できるだけ単純で推定誤差の小さい木に仕立てる剪定過程を経て，最適な木を完成させる．

3　回帰木

3.1　回帰木の分岐基準

ある目的変数 y に対し説明変数 x があるとする．ある値 c で x を分岐させるとすると，

$$S_W = \sum_{i=1}^{2} \sum_{j=1}^{ni} (y_{ij} - \bar{y}_{i\cdot})^2 \qquad (1)$$

この郡内平方和 S_w を最小にすることが回帰木の分岐基準となる．これを最小にする値 c を分岐点とする．

3.2　回帰木の剪定

回帰木の樹木はどんどん成長していく．木が大きく成長するとテストデータへあてはめたときの予測精度は良くない．そのため，分岐をどこか予測モデルとして適するところで終了しなければならない．このように最適な木にすることを剪定という．この剪定をする基準が必要になるが，R関数 tree には関数 cv.tree がある．この関数での剪定はｎ重交差確認法が用いられる．この関数 cv.tree ではデータをｎ分割するのはランダムで行われるため，毎回出力結果が異なる．このため，この作業を複数回行い，平均を用いることが考えられる．

3.3　交差確認法

交差確認法では，大きさ n の学習データを推測用のデータと評価用のデータに分け，推測用データから求めた推定値を用いたモデルを評価用データにあてはめることにより，モデルの良さを評価する．データの分け方としては，推測用のデータとして $n-1$ 個のデータを用い，残りの１個のデータを評価に用いることが多い．そして，n 個のデータの各々がちょうど１回ずつ評価用データとなるように n 通りの分け方を考え，その平均値によりモデルを評価する．モデル選択の場合には，各モデル毎に交差確認法でモデルの良さを評価し，それらを比較することにより，最良のモデルが決定できる．

4　実データでの比較

4.1　アメリカ４８州のガソリン消費量データ

重回帰分析と回帰木を比較するため，４８個のデータから３４個のトレーニングデータをランダムにとり，残りの１４個をテストデータとした．トレーニングデータについて重回帰分析，および回帰木の分析を統計解析ソフトＲで行った．回帰木は関数 tree を用いて，重回帰分析は関数 lm を用いて解析を行った．その結果にテストデータをあてはめる．これを２回行った．その結果の平均二乗誤差を比較し，予測性能を検討した．結果は回帰木・重回帰それぞれ，１回目：60.26・48.61，２回目：79.38・62.49 となった．１回目・２回目ともに重回帰分析の結果の方が平均二乗誤差が小さく，予測モデルとして回帰木より優れいてるという結果になった．

4.2　ボストンの５０６の地域の住宅価格データ

ガソリンデータと同様にレーニングデータについて重回帰分析，および回帰木の分析行い，テストデータにあてはめを行った．５０６個のデータから４００個のトレーニングデータをランダムにとり，残りの１０６個をテストデータとした．以下に回帰木の樹木構造で図１に示す．

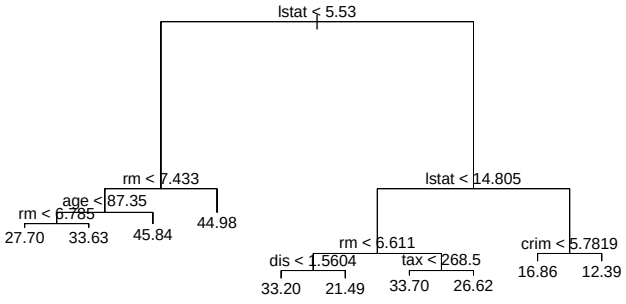

図 1　tree 回帰木

平均二乗誤差の結果は回帰木・重回帰それぞれ，１回

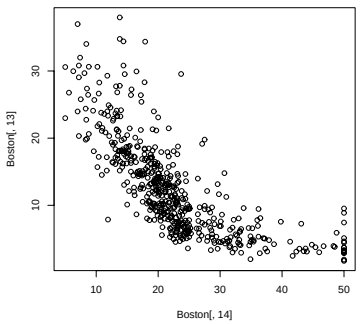

図 2　medv と lstat のプロット図

目：4.556・4.724，２回目：3.794・4.008 となった．１回目・２回目ともに回帰木の結果の方が平均二乗誤差が小さく，予測モデルとして重回帰分析より優れいてるという結果になった. 次に，目的変数を log(medv) として対数をとり解析を行った．結果は回帰木・重回帰それぞれ，0.216・0.193 となった．図３からわかるように非線形性がなくなり，重回帰の方が優れていた．

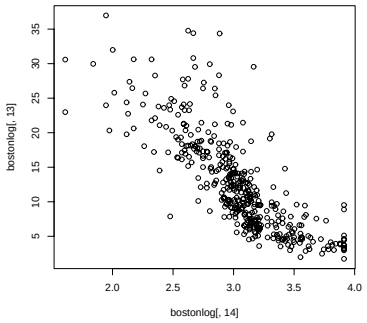

図 3　log(medv) と lstat のプロット図

5　人工データによるシュミレーション

線形モデル・非線形モデルの２通りのある関数を定義し，目的変数には誤差 e を加え，説明変数同士に相関がある説明変数を正規乱数によってデータ数５００のデータを生成した．データ数５００のうち，４００をトレーニングデータとし，残りの１００をテストデータとした．実データの解析と同様に，トレーニングデータについて重回帰分析，および回帰木の分析を統計解析ソフトRで行った．回帰木は関数 tree を用いて，重回帰分析は関数 lm を用いて解析を行った．結果から平均二乗誤差を比較し，予測性能を検討した．これを７回行った．

5.1　線形モデルの関数

線形モデルを以下の関数により生成した．
正規乱数より生成する５変数 x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i} 及び誤差 e_i それぞれの (μ, σ^2) は，(10,16)，(30,25)，(20,49)，(100,64)，(10,9)，(100,100) である．

$$y = X_{1i} + 2X_{2i} + 3X_{3i} + 4X_{4i} + 5X_{5i} + e_i \qquad (2)$$

$X_{1i} = x_{1i}$，$X_{2i} = x_{2i} + X_{1i}$，$X_{3i} = x_{3i} + X_{2i}$，$X_{4i} = x_{4i} + X_{3i}$，$X_{5i} = x_{5i} + X_{4i}$

5.2　非線形モデルの関数

線形モデルと同様に非線形モデルを以下の関数により生成した．正規乱数より生成する５変数 x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i} 及び誤差 e_i それぞれの (μ, σ^2) は，(10,16)，(10,9)，(5,16)，(8,4)，(10,9)，(30,49) である．

$$y = X_{1i}^3 + 5X_{2i}^2 + 3X_{3i}X_{4i} + 2X_{5i}^2 + e_i \qquad (3)$$

回帰木と重回帰それぞれから得られた結果にテストデータをあてはめ，予測値・残差から平均二乗誤差を求めた．回帰木・重回帰の線形モデル，非線形モデル７回分の結果の平均はそれぞれ，275.78・104.20，11704・18725 となった．この結果から，線形モデルに対しては重回帰が予測モデルとして優れており，非線形モデルに対しては回帰木が予測モデルとして優れていることがわかった．

6　まとめ

実データのアメリカ４８州のガソリン消費量データでは，重回帰分析の結果が予測モデルとして優れており，ボストン住宅価格データでは回帰木の結果が予測モデルとして優れていた．ガソリン消費量データとボストン住宅価格データでは重回帰分析と回帰木との予測精度が異なる結果となった．この違いをそれぞれのデータのプロット図から考えてみると，目的変数に影響が大きい変数と目的変数の関係が線形的と非線形的であることが読み取れる．つまり，そこが線形的であるガソリン消費量データ（FUEL と DLIC）は重回帰分析の方があてはまりがよい．逆に，非線形的であるボストン住宅価格データ（medv と lstat,nox）は回帰木の方があてはまりがよい結果となっている．したがって，目的変数に影響する説明変数と目的変数との関係が線形の場合，予測精度として重回帰分析が優れており，その逆の非線形の場合，回帰木が優れていることがわかった．また，人工データからの分析により，線形モデルに対しては重回帰分析が回帰木に比べ大いに優れていることや，非線形モデルに対しては回帰木が重回帰分析に比べて優れていることがわかった．この人工データのシュミレーションによって，２つの実データから得られた結果が明確になった．

参考文献

[1] 金　明哲：『Rによるデータサイエンス』，
森北出版株式会社，2007.

[2] 大滝　厚，堀江　宥治，Dan steinbers：
『応用２進木解析法』，
日科技連，1998.

[3] W.N. ヴェナブルズ/B.D. リプリー:
伊藤幹夫・大津泰介・戸瀬信之・中東雅樹 [訳]
『S-PLUS による統計解析』，
シュプリンガー・フェアラーク東京株式会社，2001.