機械学習特論

レポート課題　1

2019 年 11 月 7 日

創成科学研究科 基盤科学系 情報科学コース
学籍番号　19-8801-009-1

北田　和

1

1.1 目的

最小二乗法の精度がどの程度のものか検討する.

1.2 手法

M 個のデータをランダムに生成し, 特徴量 $x_i(0, 1, \ldots, M)$ のデータとした (ランダムさは, 正規分布に従い, これ以降使用するランダムな数字も正規分布に従う). これに対し, 観測データ $y_i(0, 1, \ldots, M)$ を式 (1) によって作成する.

$$y_i = \beta \times x_i + \beta_i \times 0.5 \tag{1}$$

ここで, β はランダムな定数であり, i に関係なく一つに決める. 一方, $\beta_i(0, 1, \ldots, M)$ は, i によって異なるランダムな数である. また, x_i に関係ないこの項はノイズに該当する. なお, ランダムな数は, 正規分布に従うため, 平均は 0 となる. よって, 真の関数は, 式 (2) で表すことができる.

$$y = \beta \times x \tag{2}$$

作成した特徴量 x_i と観測データ y_i から, 最小二乗法によって式 (2) の β を, 予測する. ここで, 予測した傾きを β^* とすると, 最小二乗法によって求めた, 式 (2) は式 (3) のように表せる.

$$y = \beta^* \times x \tag{3}$$

以上を計算するプログラムを以下に示す. なお, jupyter notebook を用いて解析を行った.

1.3 結果

生成した β は , 最小二乗法によって求めた $\beta*$ は , であった.
ここで, 観測データと式 (2), 式 (3) を図**??**にまとめた. 横軸は特徴量 x , 縦軸は観測データ y である.

1.4 考察

2

2.1 目的

リッジ回帰における正則化パラメータ α と予測誤差の関係を調べ, 最適な α について検討する.

2.2 手法

python のライブラリである scikit-learn 内に存在する boston housing データ使用した. boston housing データには, 13 種類の特徴量が存在する. それぞれの特徴の構成を表**??**に示す. これらの特徴量を用いて, 住宅価格を予測する.

CRIM	人口 1 人当たりの犯罪発生数
ZN	25,000 平方フィート以上の住居区画の占める割合
INDUS	小売業以外の商業が占める面積の割合
CHAS	チャールズ川によるダミー変数 (1: 川の周辺, 0: それ以外)
NOX	NOx の濃度
RM	住居の平均部屋数
AGE	1940 年より前に建てられた物件の割合
DIS	5 つのボストン市の雇用施設からの距離 (重み付け済)
RAD	環状高速道路へのアクセスしやすさ
TAX	$10,000 ドルあたりの不動産税率の総計
PTRATIO	町毎の児童と教師の比率
B	町毎の黒人 (Bk) の比率を次の式で表したもの。 $1000(\mathrm{Bk} - 0.63)^2$
LSTAT	給与の低い職業に従事する人口の割合 (%)

全体データのうち, ランダムな 8 割のデータを学習データに, 残りの 2 割のデータをテストデータとした. 予測誤差は, 予測の平均絶対誤差 (MAE) と, 平方根平均二乗誤差 (RMSE) によって導出することにした. ここで, 予測誤差が小さくなる. つまり, MAE と RMSE が最小をとるときの, 正則化パラメータ α を調べることにした. 具体的には, α を 0 から 4.99 まで, 0.01 刻みで大きくした時の, MAE と RMSE を計算し, それぞれが, 最小をとるときの α がどのくらいであったか調べる. また, 以上の手順をホールドアウト法によって, 学習データとテストデータを 100 回変更した. これにより, 導出された 100 個の α に対し, 平均と標準偏差を計算することで, どのようなデータでも, 予測誤差が小さくなるような α を検討する.

上記のプログラムを以下に示す.

2.3 結果

ホールドアウト法を用いた時の, ある学習データ, テストデータに対して, MAE と RMSE は, α の変化によって, どう変化したかを, それぞれ, 図**??**, **??** に示す. どちらの図をみてもわかるように, α が大きすぎても, 小さすぎても MAE と RMSE は最小にはならない.

ホールドアウト法によって, 導出された MSE が最小となる α の平均は 1.18 , 標準偏差は 1.67 となった. 一方, RMSE が最小となる α の平均は 0.57 , 標準偏差は 1.29 となった.

2.4 考察