データサイエンス 第一回レポート課題

工学系研究科 社会基盤学専攻 修士1年 37-176011 梶原裕希(Name: grazhdanin)

1. 予測性能を向上させるための自分なりの工夫点と結果、考察

●工夫点

工夫したところは以下の3つである。

1. 汎化性能を評価するために、テストデータを訓練データと同数とし、Validation errorを算出

予測性能を向上させるためには、過学習を防ぐことが不可欠である。よって、回帰した結果を用いて汎化性能を測れるように、訓練データと同数のテストデータを設定した。

1. 複数の手法で回帰を行い、最もValidation errorの少なかったものを選択

具体的には、線形回帰、リッジ回帰、交差検証付きリッジ回帰、Lasso、交差検証付きLasso、Elastic Net、交差検証付きElastic Net、一般線形回帰(ポアソン分布)の8つで回帰を行い、Validation errorを比較した。

(3) 回帰の結果を当てはめた時の目的変数が0以上の値となるように調整

今回予測したいのはPopularityであり、負の値を取ることはない。よって、回帰の結果得られた係数を用いて予測を行った際に、目的変数<0 ならば 目的変数=0になるように調整した。

●経過

(1) 各手法の Training errorとValidation error

各手法の結果は上の表の通りになった。この結果から、以下のことが言える。

* どの手法でも、Training ErrorよりValidation Errorが小さくなっており、過学習を防げているといえる。
* 今回の場合、交差検証付きの回帰とそうでない回帰とでは(例: RidgeとRidgeCV)後者の方がErrorは少ない。
* 目的変数がポアソン分布であることを仮定した一般線形回帰よりも、目的変数が正規分布であることを仮定したそれ以外の手法の方が高性能である

以下、Validation Errorの最も小さかったLassoを用いることで更なる性能の向上を図る。

(2) Lassoのハイパーパラメータの手動調整

ハイパーパラメータを0から0.1刻みで変化させ、Validation Errorの変化を見た。

上の表の通り、Validation Errorはハイパーパラメータが9.5の時に最小となることが分かった。

よって、ハイパーパラメータを9.5に定めることとした。

●結果

(1) 最終的に選択された手法

回帰手法: Lasso

訓練データ: 10000

テストデータ: 10000

ハイパーパラメータ: 9.5

その他: 回帰の結果を当てはめた時の目的変数が負となった場合、値を0に置換

(2) Test errorとValidation error

Test error : 6.36955668987, Validation error : 5.33103940384

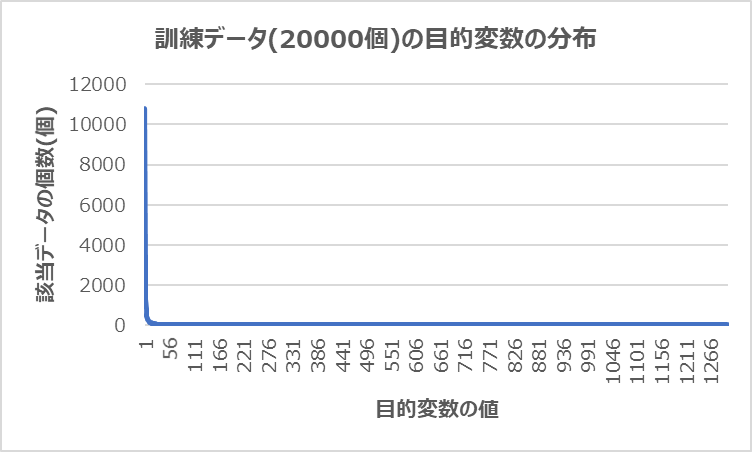
1. スコアリングサーバに提出した結果

MAE : 7.5883010733175

●考察

(1) 目的変数の分布

経過(1)で、目的変数の分布をポアソン分布と仮定するよりも正規分布と仮定した方が当てはまりが良かったと述べたが、ここで、訓練データから、目的変数の分布を調査することにする。



このように、目的変数の半分以上(20000個中10791個)が0であり、それ以外はほとんど0である。そのため、目的変数の分布が正規分布であるとは言い難い。それにも関わらず、ポアソン分布の一般線形回帰よりもlassoの方が高性能であったのは、後者はスパース性を持ち、前者はスパース性を持っていなかったためではないかと考えている。

1. その他、自由にデータを分析した結果

●Elastic NetでハイパーパラメータとL1\_ratioを変化させた時のValidation Errorの比較

ハイパーパラメータは0から1.0刻みで10.0まで変化させ、L1\_ratioは0から0.1刻みで1.0まで変化させた。その結果が以下の表である。



赤いほどValidation Errorが大きく、緑色が強いほどValidation Errorが小さい。

これによると、今回の回帰では、L1\_ratioが大きいほど、そしてハイパーパラメータが大きいほど、Validation Errorが小さいという傾向があったことがわかる。そして、L1\_ratio = 1(つまり、Lasso)、ハイパーパラメータ = 9 の時、Validation Errorが最小となることが分かる。これは、1.で最終的に導き出したLassoのハイパーパラメータと整合性のある結果である。

●各特徴の寄与の分析

各特徴と、Lasso(ハイパーパラメータ=9.5)で得られた係数の対応表を左の表に示した。正の値は赤、負の値は緑で表した。また、白抜きの値は、目的変数に比較的大きな影響を及ぼしていると思われる(絶対値が0.1以上)ものである。

この表から、各特徴の目的変数に与える影響の様子をまとめると以下のようになる。

正の影響(大): 24時間以内のコメント数

正の影響(小):基準時間以前のコメントの数、24時間以内のコメ

ント数と24~48時間以内のコメント数の差、ポストのシェア数

負の影響(大): 基準時間

負の影響(小): 訪問者数、ページをまた見に来る人の数、ページ

のカテゴリ、基準時間以前かつ最初の24時間以内のコメント数、

ポストの長さ

(5-29は明言されていない特徴のため省略)

1. ここまでの講義の感想、要望など

・Pythonを1から学ぶという講義は他にないので、初学者(自分含め)に優しい講義だと思っています。

・後ろの方だと声が聞こえにくいので、マイクを使う等して頂けると大変嬉しいです。