データサイエンス第一回課題

増井 建斗 48156621

2015 年 12 月 11 日

1 概要

多層パーセプトロンを用いてオンラインニュースの人気予測を行った．多層パーセプトロンの他，線形回帰等のモデルの構造とデータの前処理がニュースの人気予測精度に及ぼす影響について考察する．

2 提出した予測モデルの構成

提出した予測モデルは，データセットに対して前処理を加えずに学習した多層パーセプトロンである．モデルの構造は表 1 に示す．モデルの学習には Adam[2] を用いた．

表 1　多層パーセプトロンモデルの構造

Layer	1	2	3	4
Number of weights	1000	1000	1000	1
Activation	ELU	ELU	ELU	Linear
Dropout prob	0.3	0.3	0.3	-

Activation Function には，Exponential Linear Units(ELU)[1] を用いた．ELU は以下の式で定義される．

$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha(\exp x - 1) & \text{if } x < 0 \end{cases} \tag{1}$$

$$f'(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ f(x) + \alpha & \text{if } x < 0 \end{cases} \tag{2}$$

3　予測精度を向上させるための工夫点と結果，考察

表 2　前処理とモデルによる予測精度の関係

Model	LR(MSE)	LR(MAE)	MLP(MAE)
前処理なし	3209±176	2509±155	2523±405
曜日情報なし	3204±177	2512±152	-
PCA	3209±176	3510±157	-
Log	3269±159	3190±156	-
曜日情報なし,Log	3265±159	3192±156	-
曜日情報無し,PCA	3204±177	3509±157	-
Log,PCA	3269±159	3509±156	-

予測精度を向上させるため，前処理として特徴選択，Log scaling, 主成分分析を行い，それぞれが予測精度に与える影響を確認した．線形回帰モデル (LR) と多層パーセプトロン (MLP) の 2 つのモデルを用い，予測モデルと前処理それぞれの予測精度に対する影響も確認した．3 種類の前処理それぞれと予測精度の関係を表 2 に示す．値はそれぞれ validation セットに対する Mean Squared Error(MSE), Mean Absolute Error(MAE) を表し，4 fold cross validation の平均と標準偏差である．ただし，MLP については 30 fold cross validation の結果を載せている．以下，それぞれの効果について考察する．

3.1　特徴選択

特徴選択では，入力データから曜日の情報を除去する操作を行った．図 1 に，データセットの特徴量ごとの分布を示す．曜日を示す `weekday_is_monday` から `weekday_is_sunday` までの特徴は，特徴と目的変数の相関が低かったため消去することで予測精度を向上させることが出来るのではないかと考えた．表 2 における曜日情報なしの項を確認すると，LR(MSE) については若干の精度向上が確認できるが，LR(MAE) においてはむしろ精度が悪化している．従って，期待した精度向上の効果は得られなかった．

3.2　Log scaling

図 1 を確認すると，`n_tokens_content` 等，幾つかの特徴量のスケールが指数的であることが確認できる．線形モデルでは非線形な指数的スケールを持った特徴から回帰を行うことが難しいだろうと考え，それらの特徴の log をとる処理を行った．この処理を行った後の特徴量の分布が，図 2

である．この処理は期待に反して予測精度を下げる結果となった．

3.3 主成分分析 (PCA)

主成分分析を用いて入力変数を変換することで，入力変数の線形和で目的変数を表現しやすくなるのではないかと考えた．しかし，この前処理も期待に反して予測精度を下げる結果となった．この理由は，PCA は入力変数を表現しやすい射影を計算するのみで，目的変数の回帰に有益な射影を得ることが出来るとは限らないためだと考察する．

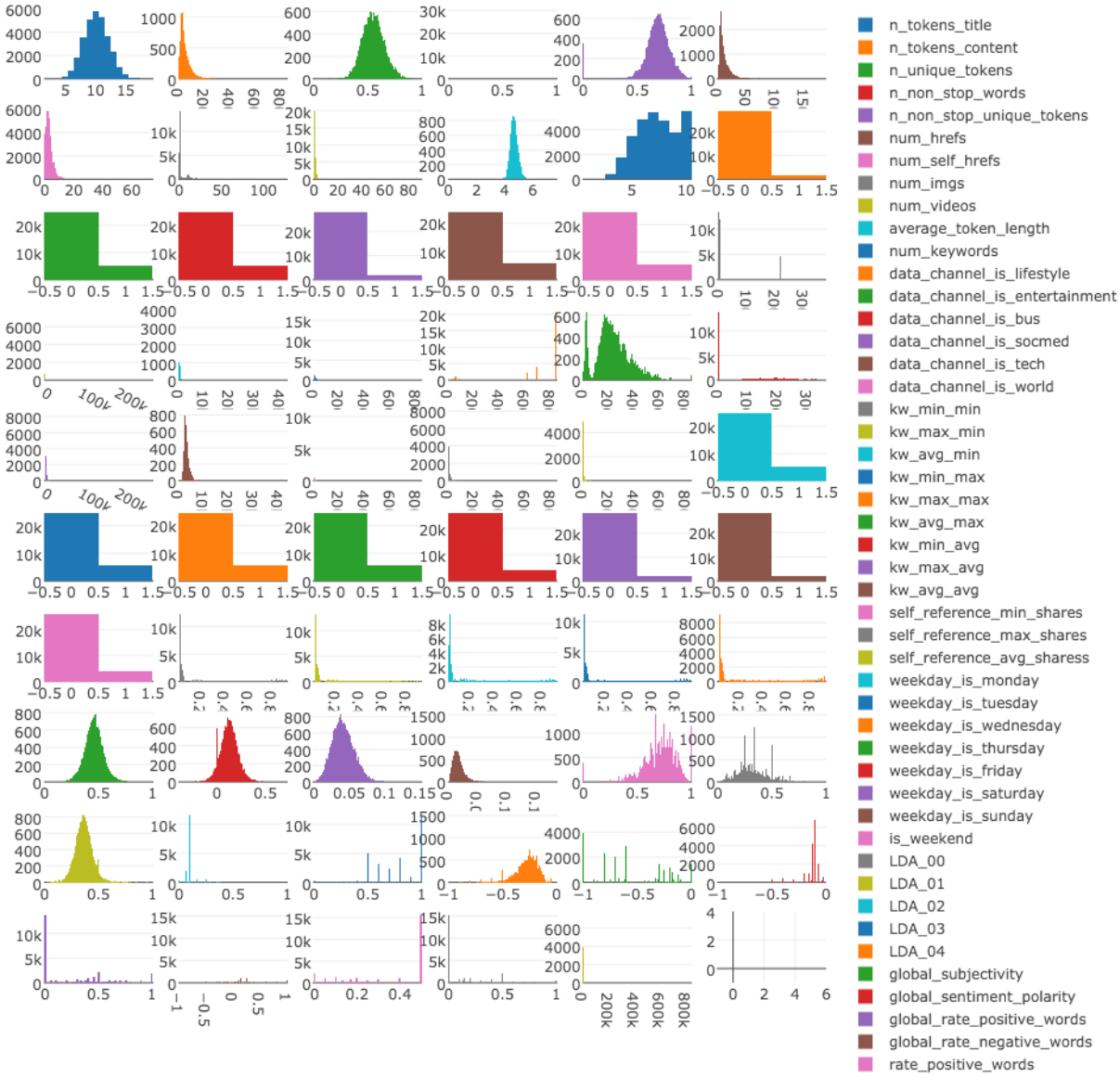

図１　学習データ・セットの各特徴ヒストグラム

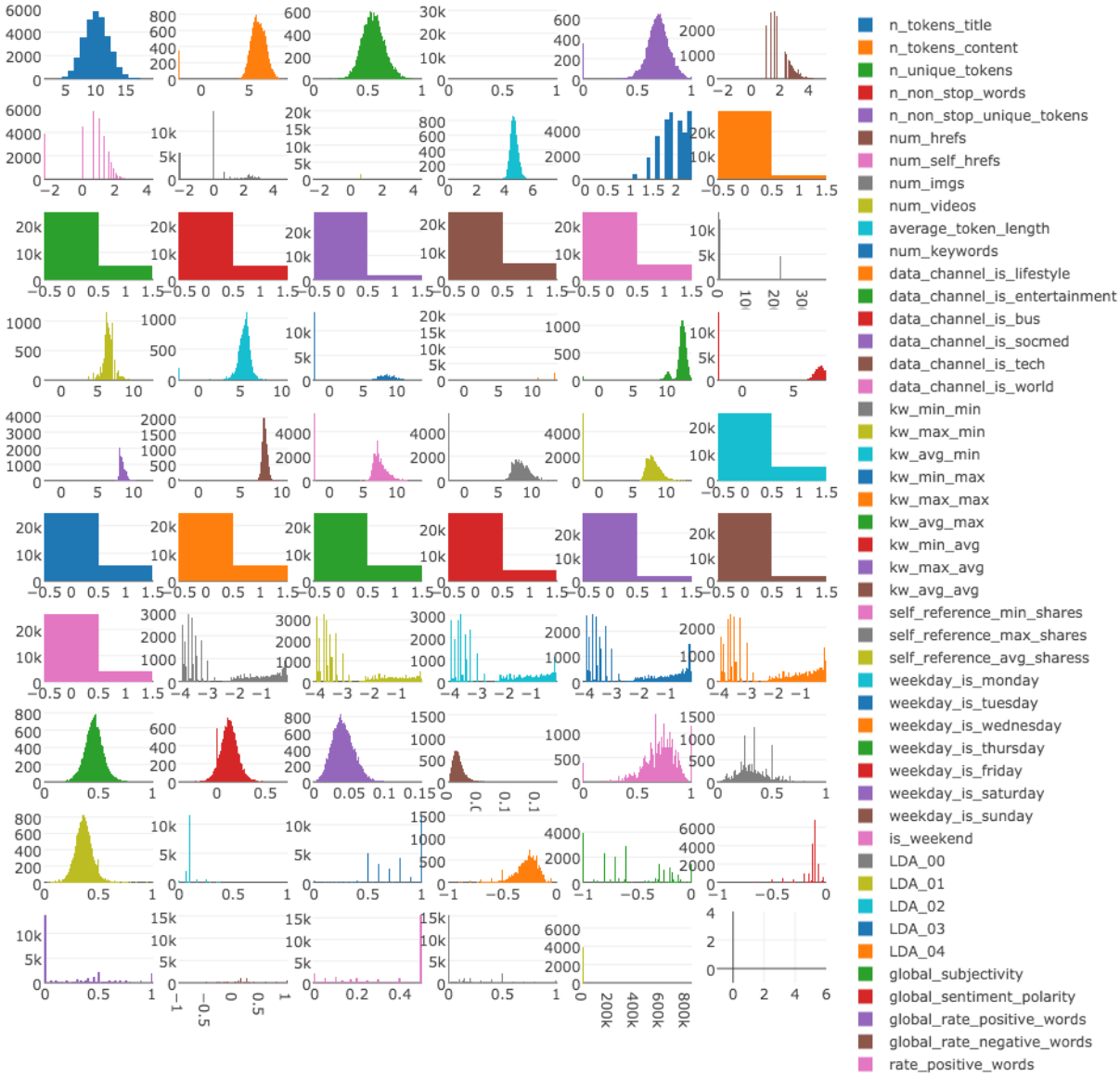

図 2　学習データセットの一部について，log ををとったもの

4　データ分析結果

4.1　目的変数分布の一様化

回帰対象となっている，shares の自然対数を取った結果のヒストグラムを図 3 に示す．図 3 から，目的変数の生起確率は一様となっていないことがわかる．従って，学習モデルは生起確率が高い値を出力するように収束すると考えられる．目的変数のスケールが指数的であるため正規確率の低い部分に 80 万などの非常に大きな値が含まれているが，この値の予測に失敗すると，テストデータセットの標本サイズによっては非常に大きな MAE が現れる．今回の場合標本サイズは 10000 であるため，80 万に対して 1 万と 1 回予測するだけで MAE に 79 の影響を与えてしまう．これを回避するためには，正規確率が低い値を正確に予測する必要があり，目的変数の分布を一様にすることで正規確率が低い値の予測精度を上げることが出来ると考えた．実際に目的変数の分布が一様になるようにデータ・セットをサンプルしなおし，回帰学習を行った結果正規確率の低い値の予測精度を上げることが出来た．しかし実際に最適化したい，元の分布の目的変数に対しては，より高い MAE を出力することになった．これはもちろん，実際の目的変数の分布と最適化に用いた目的変数の分布が異なるためである．従って，単純に分布を一様にするだけでは正規確率の低い値に対する予測精度と MAE の最小化を両立することが出来ないとわかる．

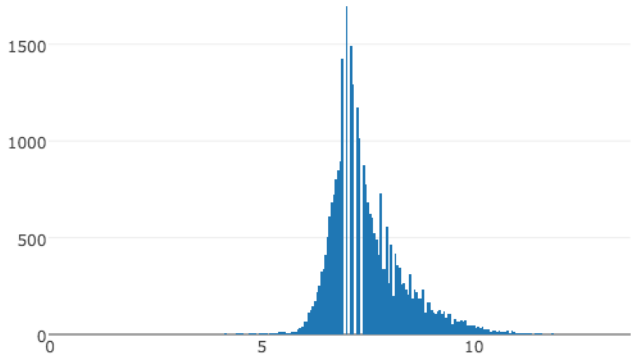

図 3　目的変数の分布

4.2 正則化パラメータの関係

LR による回帰では，正則化パラメータを用いなかった．MLP については，l1,l2,l1l2,dropout の効果を確認した．MLP では正則化を用いない場合，ある程度学習が進んだ後に学習データに対する MAE を減少させつつ Validation データに対する MAE を増大し続ける，過学習をする傾向が確認できた．これに対してそれぞれの正則化項を追加し再度学習を行った場合，全ての正則化について学習データと Validation データが逆の方向に増減する傾向が抑制され，過学習が抑制されることが確認できた．それぞれの正則化については収束の速度に差異が見られることが確認でき，特に l1 正則化のみを用いた場合には学習が比較的遅くなることが確認出来た．

4.3 Stacked Autoencoder

Stacked Autoencoder による pretraining によって予測精度が向上するか検討したが，Autoencoder を用いない場合と差異が見られなかった．

4.4 Order Classification

目的変数が指数的な値をとっていることに着目し，まず目的変数の指数を識別し，その値を追加の特徴量として目的変数の回帰を行うことで予測精度が向上出来るのでは無いかと考えた．目的変数の指数識別には MLP を用いたが，予測された指数から実際の値に戻すのみでも MAE は 3000 台が得られた．この識別された値を追加の特徴量として新たな MLP を用いて回帰を行ったが，結果は指数を用いない場合と全く同じとなった．

5 ここまでの講義の感想，要望

現在自分が研究している分野の授業だが，様々な見逃していた情報を学習することが出来た．また，実際に python でのコーディングを行うときに気をつけるべき点が示されており，大変参考になっている．

参考文献

[1] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[2] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.