ニューラルネットワークによるTwitterエンゲージメントの予測

石川県立金沢泉丘高等学校

研究の流れ

①TwitterAPIを用いてツイートを収集する
②ツイートを絵文字の消去などの前処理にかける
③前処理済みのツイートを分かち書きする
④Doc2Vecを用いてツイートをベクトル（数値データ）に変換する
⑤ベクトルとフォロー数・フォロワー数・曜日・時刻のデータを使って
モデルを作成し、検証・考察する

モデル作成の下準備

予測対象とするエンゲージメントは
いいね数とリツイート数の和
として計算しておく

利用システム

TwitterAPI

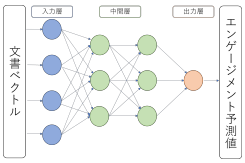
Twitter社から提供されている
特定のワードを含んだツイート
を収集することができるもの

Doc2Vec

ライブラリ「Gensim」で使用できる
文章をベクトルに変換することの
できる機能[1]

Neural Network

入力した値に、重みをかけて線形変換した
値を次の層へ送る、という計算操作がネット
ワーク状に連なったもの。ネットワーク全体
の重みを調整する操作1回を1epochと呼ぶ。
調整方法はAdamというものを用いた。



イメージ図

評価関数

平均二乗誤差(MSE)
モデルの学習時にMSEの値が小さくなることを
目的とする。

$$MSE = \frac{1}{n}\sum_{k=i}^{n}(y_i - \hat{y}_i)^2$$

y_i：i番目のデータの真の値
$\hat{y}_i$：i番目のデータの予測値

パターン

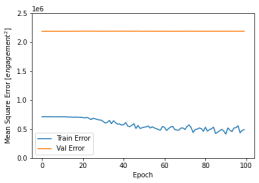
ニューラルネットワークに入力するデータの組み合わせ

パターン①：テキスト
パターン②：テキスト・フォロー数・フォロワー数・曜日・時刻
パターン③：フォロー数・フォロワー数・曜日・時刻

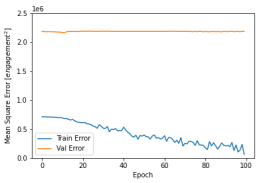
モデルの作成と考察

「オススメ」という言葉の含んだツイートを22029件収集し、モデルを学習した。ベクトルは400次元とした。

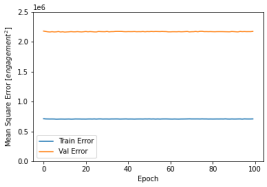
パターン①



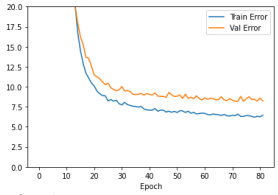
パターン②



パターン③



・縦軸：MSE　横軸：epoch
・オレンジ：検証データ
青色:訓練データ
・正しく学習できているときの
グラフは右上のような推移になる[2]

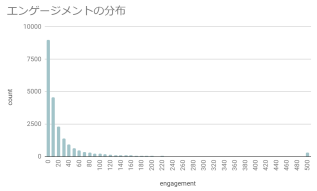


・学習しても未知のデータである検証データに対して予測の誤差が減少していない（収集データの調査へ）
・ベクトルの次元数が大きすぎて入力が多くなり予測がうまくいっていない可能性がある（モデルの改善へ）
・パターン①とパターン②より、テキスト以外のデータが予測に影響を及ぼしていると考えられる
・パターン③では訓練データでも学習が進まないためテキストは確実に予測に必要であるとわかる
➡ただし、一般的な感覚ではテキスト以外のデータとエンゲージメントの間だけでも相関性がありそう（モデルの改善へ）

収集データの調査

今回使用した「オススメ」という言葉を含んだツイート
22029件のエンゲージメントのヒストグラムを作成した。

エンゲージメント250以下の
ものが大多数



22029ツイート中
エンゲージメントが500以上
のものは**313ツイート**のみ

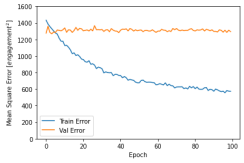
外れ値と捉えられる

エンゲージメントが極端に大きいツイートが大きな**外れ値**として予測に影響を及ぼしている可能性がある

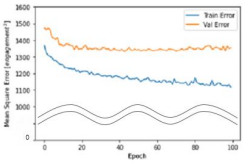
モデルの改善

エンゲージメントの範囲を250以下に、ベクトルの次元を100次元にして再度パターン②、③で学習を行った。

パターン②



パターン③



・パターン②は正しく予測できていない
➡ベクトルの次元がまだ大きすぎるかベクトル変換がよくない
・パターン③はグラフが下がった
➡エンゲージメントが250以下の小さいツイートではフォロー数・
フォロワー数・曜日・時刻とエンゲージメントには少なくとも
相関がある

今後の展望

・より高精度なモデルの作成
・どのような言葉の含んだツイートでも予測することのできる一般的なモデルの作成
・多くのエンゲージメントを得る方法の理論化と実社会で広告への応用

参考文献

[1]Tomas Mikolov. Distributed Representations of Sentences and Documents. https://cs.stanford.edu/~quoc/paragraph_vector.pdf(参照2021-11-10).
[2]Google.”回帰：燃費を予測する” https://www.tensorflow.org/tutorials/keras/regression?hl=ja(参照2021-11-10)
他多数

石川県立金沢泉丘高等学校　5

ニューラルネットワークによるTwitterエンゲージメントの予測

抄録

総務省の調査では、2021年1月時点で日本人口の42.3%がTwitterを利用しているため、ツイートに対して多くの人々から反応を得られる可能性がある。そこでニューラルネットワークを用いて反応を表すエンゲージメント（いいね・リツイート数）を予測するモデルを作成することにした。

１．研究の背景と目的

多くの人がTwitterを利用しているため、ツイートに対して様々な反応が存在する。そこでツイートがどれだけのエンゲージメントを得られるかを予測するモデルの作成を目指す。

２．方法

1. TwitterAPIを用いてツイートデータを収集する。
2. テキスト内の重ね表現や、数、絵文字に関して処理を行う。
3. Doc2Vecを用いて、テキストを数値化(ベクトル化)する。
4. そのベクトルにツイートの情報を付加したデータを、予測の基とする訓練用データと予測の誤差を測るテスト用データに分割し、前者を用いていいね・リツイート数を予測するニューラルネットワークモデルを作る。
5. テスト用データを用いて予測の誤差を測定する。
6. ツイートの情報であるテキスト・フォロー・フォロワー・時刻の組み合わせを変更し、より予測の誤差が小さいモデルを作成する。
7.

３．結果

様々なパターンを試したが、すべてにおいていいね・リツイート数の誤差の平均がおよそ80となった。また、訓練用データのみ誤差の減少が見られた。

４．考察

訓練用データのみに特化したモデルができてしまったことや、付加する情報を変化させても誤差の変化がほぼなかったことより、モデルの訓練の方法または使用するデータに問題があると考えられる。

５．結論

作成したモデルは誤差およそ80であり、訓練方法や使用データによってより誤差の小さいモデルの作成ができる可能性がある。

６．参考文献

Tomas Mikolov. Distributed Representations of Sentences and Documents.
https://cs.stanford.edu/~quoc1e/paragraph_vector.pdf(参照2021-11-10).
Google.”回帰：燃費を予測する”
https://www.tensorflow.org/tutorials/keras/regression?hl=ja(参照2021-11-10).

７．キーワード

TwitterAPI Doc2Vec ニューラルネットワーク 自然言語処理