

Self-Attentive-LSTM への順序ロジットモデル適用

Applying Ordinal logit model to Self-Attentive-LSTM

参木 裕之^{*1}
Hiroyuki Mitsugi

北野 道春^{*1}
Michiharu Kitano

渡部 裕晃^{*1}
Hiroaki Watabe

^{*1} 株式会社 大和総研
Daiwa Institute of Research Ltd.

In this research, we propose a method incorporating structure of ordered logit model into output layer of Self-Attentive-LSTM. This helps to predict ordinal variables. Model with using this method learns relationships among classes since the order of classes are assumed in advance. We show this model is useful when the objective variable is ordinal by applying to "The economic watcher", data surveyed by the Japanese Cabinet Office. When predicting the Economic assessment index, this proposed method improved F-measures of both sides of the ordinal indices compared to Self-Attentive-LSTM.

1. はじめに

機械学習において、教師あり学習でも目的変数の種類によって扱う問題が異なる。目的変数が有限離散集合であれば分類問題となるが、目的変数が実数集合であれば回帰問題となる。また、目的変数が順序変数の場合、ランキング学習や順序回帰問題となる。特に、順序ロジットモデル[McCullagh 1980]は、判別関数の係数を各クラスで共有し、切片のみが各クラスで異なるため、クラス分類と回帰の両方に位置する手法である。機械学習の応用の観点としては、近年、深層学習の研究が日々発展し、テキストマイニングにおいて、翻訳や文書生成などのタスクで精度を向上させている。しかし、目的変数が順序変数を扱った問題は筆者の知るところでは無い。例えば、説明変数がテキストを分割した形態素の系列、目的変数が順序変数の場合、LSTM(Long-Short-Term-Memory) [Hochreiter 1997] で目的変数を予測すると、精度が良くならない場合がある。

本研究では、Self-Attentive-LSTM(Long-Short-Term-Memory) [Zhouhan 2017]の出力層に順序ロジットの構造を取り入れることで、順序変数の予測に役立つ手法を提案する。LSTMはRNN(Recurrent Neural-Network)の拡張であり、テキスト等の長文の学習に優れている。LSTMにSelf-Attentiveの構造を取り入れることで、予測に寄与した形態素の寄与度を算出することが可能になる。通常の判別モデルであれば、各クラスの関係性は独立であるが、順序ロジットを取り入れる利点として、各クラスの順序性が仮定されるため、クラス間の関係性をモデルが学習できるところである。目的変数が順序変数の場合に本手法の有用性を、内閣府が調査・公表している景気ウォッチャー調査[wat]を使用して示す。景気ウォッチャー調査は景気判断理由のテキストと5段階の景気判断値が対応しているデータセットであり、2000年に調査を開始してから、約20万件のアンケート情報が確保されている良質なデータである。提案モデルでテキストから景気判断値の予測をしたところ、Self-Attentive-LSTMに比べて景気判断値の両側(景気が良い、悪い)のF-measureの精度が向上した。

2. 関連研究

本研究に関連がある順序ロジットモデルとLSTMを用いた研究を紹介する。

Jianlin らは、ニューラルネットワークの出力層に順序ロジット

の構造を適用することで、多くのデータセットに対して、通常のニューラルネットワークより精度が良いことを示している[Jianlin 2008]。目的変数の与え方をone-hot表現から工夫することで、順序ロジットの構造を取り入れている。しかし、数値データの検証のみでテキストデータの検証には触れられていない。

深層学習を景気ウォッチャー調査に適用した研究として、[山本 2016]では、5段階の景気判断値(1~5)の中央の値を除いて、1と2を「ポジティブ」、4と5を「ネガティブ」と扱い、2値判別問題としてLSTMで景気ウォッチャーデータを学習し、景気判断が「ポジティブ」か「ネガティブ」を予測している。しかし、目的変数に順序性があることはモデルに考慮されていない。

上記の問題点を踏まえて、本研究では、目的変数の順序性を考慮し、Self-Attentive-LSTMに順序ロジットモデルを適用することで、景気ウォッチャー調査の景気判断値の両側(景気が良い、悪い)のF-measureを既存モデルより精度良く予測した点に新規性があると考えている。

3. OL-ALSTMモデルの提案

提案モデルをOL-ALSTM(Ordinal Logit Self-Attentive LSTM)と呼ぶことにする。ベースモデルとなるSelf-Attentive-LSTMを紹介した後に、OL-ALSTMモデルを説明する。

3.1 Self-Attentive-LSTM

予測に寄与した単語を算出する方法としてAttention構造がある。Attentionには、翻訳等のseq2seqで使われるGlobal Attention[Bahdanau 2015]やLocal Attention[Luong 2015]があるが、本節では、Self-Attention [Zhouhan 2017]のみを解説する。

Self-Attentive-LSTMとは、LSTMにAttention構造を入れたモデルである。 $X = (x_1, x_2, \dots, x_n)$ を入力系列(系列の長さは n とする)とし、 x_t を t 時点の入力ベクトル、 $\overleftarrow{h_{t-1}}$ を $t-1$ 時点の順方向の隠れ層のベクトル、 $\overrightarrow{h_{t+1}}$ は $t+1$ 時点の逆方向の隠れ層のベクトルとおくと、Bidirectional LSTMの出力は次のように表せる。

$$\begin{aligned} \overleftarrow{h_t} &= \overleftarrow{LSTM}(x_t, \overleftarrow{h_{t-1}}) \\ \overrightarrow{h_t} &= \overrightarrow{LSTM}(x_t, \overrightarrow{h_{t+1}}) \end{aligned}$$

$\overleftarrow{h_t}$ と $\overrightarrow{h_t}$ を結合した h_t を隠れ層の t 時点のベクトルとする。unidirectional LSTMの隠れ層ベクトルの次元を u 、系列の長さを n とすると、隠れ層の系列行列 H は次のように表せ、 $n \times 2u$ の次元となる。

$$H = (h_1, h_2, \dots, h_n)^T$$

連絡先: 参木 裕之, 大和総研 フロンティアテクノロジー部,
hiroyuki.mitsugi@dir.co.jp

Attention 行列は次のように求めることができる。

$$A = \text{softmax}(W_2 \tanh(W_1 H^T))$$

W_1 は $d_a \times 2u$ の重み行列であり、 d_a はハイパーパラメータである。また、 W_2 は $r \times d_a$ の重み行列であり、 r はハイパーパラメータである。 A は $r \times n$ の次元になり、 softmax は行方向に計算し、行方向に和をとると1になることを意味する。入力 H と H から計算される A を用いると、文章の埋め込み行列 M は次のように表せる。

$$M = AH$$

埋め込み行列 M を用いて、最終的に出力層から目的変数を予測する。また、各形態素の寄与度の算出方法の1つとして、 A に対して r の平均(列方向に平均)をとることで寄与度を求めることができる。

3.2 OL-ALSTM モデル

上記の Self-Attentive-LSTM の出力層に順序ロジットの構造を適用することを考える。目的変数ベクトルを y とする(K 個のカテゴリをもつ)と、順序性を考慮しなければ、one-hot 表現とするが、順序性を考慮する場合は、 $y \leq k$ であれば、ベクトルの先頭から k 個を1とし、 $y = (1, 1, \dots, 1, 0, 0)$ のように表す[Jianlin 2008]。出力層の k 番目の出力が、目的変数が k 以下になる確率を意味する。出力層のベクトルを

$$o = (o_1, o_2, \dots, o_k, \dots, o_K)$$

とすると、 $o_i (i \leq k)$ が1、 $o_i (i \geq k)$ が0に近づくように学習し、重み行列を逆誤差伝播法で更新する。活性化関数はシグモイド関数とし、損失関数はクロスエントロピー関数とする。

$P(y_i \leq i) - P(y_i \leq i+1) = P(y_i = i)$ であるため、クラス分類を行う場合は、 $o_i - o_{i+1} := \tilde{o}_i$ と定義すると、各クラスの予測確率は、以下のように表せ、

$$\tilde{o} = (\tilde{o}_1, \tilde{o}_2, \dots, \tilde{o}_k, \dots, \tilde{o}_K)$$

最終的な予測クラスは次のように求めることができる。

$$\hat{y}_i := \arg\max_i \tilde{o}_i$$

4. 実験

OL-ALSTM モデルの精度を検証するため、Self-Attentive-LSTM と比較する。

(1) 前処理

形態素は wikipedia のダンプ[wik]から学習した Word2Vec モデルによって分散表現ベクトル化した (Word2Vec の次元は 150 次元とした) [Mikolov 2013]。学習済分散表現ベクトルを重み行列として、OL-ALSTM モデルの Embedding 層に初期値として入力し、End-to-End で学習を行った。また、未知語と padding に対応するため、それぞれ未知語用のベクトルと padding 用のベクトルを用意し、Embedding 層の重み行列に結合した。

(2) モデルのパラメータ

$u = 300$, $d_a = 350$, $r = 30$ とし、 M で特徴抽出が完了した後に 2000 次元の層と 200 次元の層を追加し、出力層と結合する。系列の長さは最長の系列にあわせ、バッチサイズは 32、ドロップアウトは 0.2、学習率は 0.001、最適化は Adam で行った。

(3) 景気ウォッチャーデータ

景気ウォッチャー調査はタクシーの運転手やスーパーの店主等の景気に敏感な方々を対象としたアンケート調査であり、景気判断理由のテキストと 5 段階の景気判断値からなる。データは PDF および CSV 形式で毎月公開され、景気の現状判断と先行き判断の 2 種類存在するが、今回は現状判断にて検証を行う。

表 1: 現状判断の全回答数割合

悪い:1	0.093
やや悪い:2	0.253
変わらない:3	0.435
やや良い:4	0.201
良い:5	0.018

(4) 学習方法

2000 年から 2016 年までのデータを 1:1 の割合で学習データと検証データに分け、検証データの Accuracy で評価する。なお、バイズ最適化等のハイパーパラメータの最適化は行っていない。

(5) 実験結果

混同行列から集計すると、OL-ALSTM は景気判断値の「1」や「5」の F-measure が大きく、回答数割合が少ない景気判断値の両側(景気が良い、悪い)を当ててるのが得意と考えられる。不均衡データであるため、「2」と「3」と「4」の F-measure が高くなる傾向がある中で、モデルとしては、「1」や「5」の景気の良い局面や悪い局面の F-measure の精度が向上したことに意味がある。また、Precision の観点から、「1」や「2」と予測したデータのうち、正解が「5」であった予測誤りの割合が減少し、Recall の観点からは、正解が「4」や「5」のデータのうち、「2」と予測された再現誤りの割合が減少していたことが確認できた。このことから、順序ロジットの構造をモデルに取り入れることで、「1」を「5」と誤判別する誤差と「1」を「2」と誤判別する誤差が異なるので、クラス間の順序性を考慮した学習と予測ができていと考えられる。

表 2: OL-ALSTM の予測結果

OL-ALSTM		Precision	Recall	F-measure
景気判断値	1	0.458	0.221	0.298
	2	0.504	0.571	0.536
	3	0.658	0.694	0.676
	4	0.701	0.746	0.726
	5	0.440	0.027	0.051

表 3: Self-Attentive-LSTM の予測結果

Self-Attentive-LSTM		Precision	Recall	F-measure
景気判断値	1	0.526	0.166	0.252
	2	0.508	0.596	0.549
	3	0.654	0.713	0.682
	4	0.720	0.735	0.728
	5	0.800	0.002	0.005

5. まとめ

本稿では Self-Attentive-LSTM の出力層に順序ロジットの構造を取り入れることで、OL-ALSTM を提案し、クラス間の関係性を順序関係に着目して学習を行った。また、目的変数が順序変数である景気ウォッチャー調査を用いてモデルの検証を行い、景気判断値の両側(景気が良い、悪い)の F-measure の精度が向上したことを確認できた。しかし、今回検証した景気ウォッチャーデータは景気判断値の両側(景気が良い、悪い)が反転している教師データは少ないため、OL-ALSTM モデルと Self-Attentive-LSTM モデルで予測精度においては大きな差は見受けられなかった。今後は、株価データのようにリスクとリターン

相関が高いデータ(上昇と下落の反転があり得る)で同様の検証を行い、提案モデルが優位となる局面を考察する予定である。

参考文献

- [Bahdanau 2015] Bahdanau, D. et al: Neural machine translation by jointly learning to align and translate, arXiv:1409.0473, 2014.
- [Hochreiter 1997] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, Vol. 9, No. 8, pp. 1735-1780, 1997.
- [Jianlin 2008] Jianlin, C. et al: A neural network approach to ordinal regression, IJCNN, 2008.
- [Luong 2015] Luong, M. et al: Effective approaches to Attention-based Neural machine translation, proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412-1421, 2015.
- [McCullagh 1980] McCullagh, P.: Regression models for ordinal data, Journal of the Royal Statistical Society B, 42, 109-142, 1980.
- [Mikolov 2013] Mikolov, T. et al: Distributed Representations of Phrases and their Compositionality, In Proceedings of the 26th Neural Information Processing Systems, NIPS 2013, pp.3111-3119, 2013.
- [wat] 景気ウォッチャー調査, 内閣府, http://www5.cao.go.jp/keizai3/watcher/watcher_menu.html
- [wik] ウィキペディアの日本語版のダンプ, ウィキメディア, <https://dumps.wikimedia.org/jawiki/>
- [Zhouhan 2017] Zhouhan, L. et al: A Structured Self-attentive Sentence Embedding, arXiv:1703.03130, 2017.
- [山本 2016] 山本裕樹, 松尾豊, 景気ウォッチャー調査の深層学習を用いた金融レポートの指数化, 第20回人工知能学会, 2016