令和 3 年 6 月 16 日 (水) 研究発表会　　M1 楠本 祐暉

進捗報告

1　現在取り組んでいる課題

- PyTorch を用いた多値分類タスク.
- 以前取り組んでいた 2 値分類タスクの結果を詳しく見る.

2　内容について

- 実装環境は GoogleColaboratry です.
- 使用したのは東北大学の BERT で, データセットは先週岡田先生に頂いた「楽天トラベルレビュー: アスペクト・センチメントタグ付きコーパス」データを用いました.
- プログラムは 2 値分類タスク時に実装したものとほとんど同じで, 主な変更点は出力クラス数等です.
- DataFrame で開いたデータは表 1 の様になっており, 旅行のレビュー文とその文章が朝食, サービス, 施設などといった対象に対してポジティブなことを言っているのか, ネガティブなことを言っているのかが記されている. データが入っていない NaN の部分は全て 0 に置き換えました.
- 表 2 は今回の実験で用いたパラメータです.
- 交差検証を行い, 正答率を表 3 にまとめました。

表 1　多値分類タスクにおけるデータの内訳

テキスト	朝食 po	朝食 ne	夕食 po	夕食 ne	風呂 po	風呂 ne	サービス po	サービス ne	立地 po	立地 ne	設備 po	設備 ne	部屋 po	部屋 ne
、お部屋も広くて, お料理もとても美味しく, 部屋の露天風呂からは星がプラネタリウムのように 広がっていて, とにかく最高でした.	1	0	1	0	1	0	0	0	0	0	0	0	1	0
立地と値段で決めました.	0	0	0	0	0	0	0	0	1	0	0	0	0	0
一部の方が指摘した通り, 廊下がタバコ臭いのが気になりました.	0	0	0	0	0	0	0	1	0	0	0	1	0	0
もう年齢的にも量ばかりは要らないと思っているので, 嬉しい変更でした.	1	0	1	0	0	0	0	0	0	0	0	0	0	0

表 2　多値分類タスクにおける実験時のパラメータ

出力クラス	バッチサイズ	エポック数	最適化関数
12	128	5	Adam

表 3　多値分類タスクにおける交差検証での正答率 (5 分割)

0.8943	0.8738	0.9034	0.8989	0.8954

3　今までにやったこと

- コンフュージョンマトリクスを確認したところ出力クラス数が 12 であるのに 2 クラス分しか出力されなかったのでプログラムのミスがどこにあるのか確認しています.

- その確認として, 今用いているプログラムは以前取り組んでいた 2 値分類タスクのプログラムを少し改変したものなので, そちらの方でコンフュージョンマトリクス (表 10) が正しく導出されるかの確認をし, 以前は行っていなかった交差検証を行った. 表 7 に実験時のパラメータを, 表 8 に正答率を示す.
- また, 表 9 に正答率の平均値および分散を示す.
- BertLayer の最終層と全結合層のみファインチューニングを行いました.
- 用いたデータ内容の詳細として, 表 5 に訓練データ, テストデータに含まれるポジティブラベル数とネガティブラベル数を示す.
- 具体的にどういった文章で間違えたのかを確認するために, ラベルの予測を間違えた文章データを取り出しました.
- 引き続き多値分類タスクに取り組んでおりましたが, 用いる損失関数を入力が一次元配列でなければならない CrossEntropyLoss ではなく BCELoss(Binary Cross Entropy Loss) や BCEWithLogitsLoss に変更しなければならないと思い,BCEWithLogitsLoss を用いると出力の予測ラベルが思ったようなものにはならなかった.
- 損失関数は BCEWithLogitsLoss を用いるべきだと思われる.
- 文章を単語 ID に変換し,BERT に入力する. 得られた出力結果中の [CLS] の埋め込み表現をリストにし, これを output とする.　二値分類では torch.max を用いて予測を行っていたが, マルチラベルの場合では各ラベルごとに二値分類を行うなどの工夫が必要だと考えた. 表 11 に output の例を示す.

4　今週行ったこと

- 予測方法として, output データに対してある閾値を超えれば 1, 下回れば 0 に置換する方法を試したが, この処理を加えると勾配計算に必要な変数がインプレース操作されたことを理由にエラーが出た. そのため,torch.round で浮動小数点数を最も近い整数に置換することで代用した. つまり, 閾値が 0.5 であることと同じである. 表 12 に用いたデータの内訳を示す. また, 正解率は 0.9174 であった.　表 13 に用いたパラメータを示す.

表 4　2 値分類タスクに用いたデータの内訳

総データ数	訓練データ数	テストデータ数
5638	4511	1127

表 5　ポジティブ, またはネガティブのラベルが付与されているデータの数

	ポジティブ数	ネガティブ数
訓練データ	2350	2161
テストデータ	573	554

表 6　予測を間違えたデータの具体例

テキスト	ラベル	予測ラベル
朝夕に富士山が見えてこそ、の料金だと思いました。見えなければ、高すぎます。 宿からの回答: この度は当館をご利用頂き誠に有難うございました。富士山をご覧になれず大変残念でございました。 富士山をご覧になれなくともお客様にご満足頂ける旅館を目指し 努力して参りたいと思います。どうぞまたのお越しをお待ち申し上げます.	0	1
金沢駅裏で立地は最高です。ホテル内もきれいでフロントも好感がもてます。 部屋の設備は普通レベルで価格相応だと思います。これまでに経験したことがなかったのですが、 ベッドの堅さと枕の具合が悪くたびたび目を覚まし、2泊とも熟睡できませんでした。	0	1
結婚式で利用しましたが、スタッフの方が優しく、安心して結婚式を任せる事ができました。 式を予約してから式までホテルのフィットネスクラブが無料で使えるので、 結婚式を予定している人は早めに予約して長く利用するとお得かも！？	1	0
食事等では何度も利用していましたので、使い勝手の良い立地はよく知っていました。 宿泊はすごく高いだろうと思っていたら意外と手頃なプランがあるのですね。 新宿西口方面のシティホテルは駅から遠いものが多いなか、ここは駅から徒歩でも苦にならない距離なのでいいですね。 地下街もありますし。さすがに古い感じはしますが、サービスも悪くないし、チャラチャラしていない老舗ホテルの貫禄があります。	1	0

表 7　2 値分類タスク実験時のパラメータ

. 出力クラス	バッチサイズ	エポック数	最適化関数
2	32	20	Adam

表 8　2 値分類タスクにおける交差検証での正答率 (5 分割)

0.8527	0.8554	0.8642	0.8607	0.8424

表 9　2 値分類タスクにおける交差検証での正答率 の平均と分散

平均	分散
0.8551	0.0000056262

表 10　2 値分類タスクでのコンフュージョンマトリクス

		予測の分類結果	
		ポジティブ	ネガティブ
実際の分類結果	ポジティブ	416	138
	ネガティブ	39	534

表 11　output の例

0.0921	-0.0216	-0.1953	-0.0232	0.0008	0.3894	0.0208	0.5026	0.1059	0.3584	0.1519	0.0559
0.3048	-0.0073	-0.1722	-0.1152	-0.0124	0.2630	0.0905	0.5103	0.3254	0.2793	0.0559	-0.0200
0.1676	0.1041	-0.2229	-0.0013	0.0629	0.5363	0.0739	0.5836	0.2767	0.4200	-0.0214	0.123

表 12　多値分類タスクに用いたデータの内訳

総データ数	訓練データ数	テストデータ数
60000	48000	12000

表 13　多値分類タスクに用いたパラメータ

値	パラメータ
入力層の次元数	768
出力層の次元数	12
バッチサイズ	12
最適化関数	Adam
損失関数	BCEWithLogitsLoss