令和 3 年 6 月 2 日 (水) ゼミ資料 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.

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.
- プログラムは先週まで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や損失関数を CrossEntropyLoss に切り替えたことです.
- 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 データ内容

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 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 交差検証での正答率 (5 分割)

0.8943	0.8738	0.9034	0.8989	0.8954

3 今週やったこと

- 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 9) 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 6 に実験時のパラメータと, 表 7 に正答率を示す. また, その正答率の平均値や分散を表 8 に示す.
-
- 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.

表 4　データ数

総データ数	訓練データ数	テストデータ数
5638	4511	1127

表 5　データ内容

テキスト	ラベル
立地がとてもよく, 料金も安くて満足でした. 風呂が少し小さいことが残念でした.	1
立地はまあまあだが施設が古いのが欠点, だけど安いのは良い. 朝食もそれなりでまあまあ満足. しかし朝食のレストランがいまいち.	0
悪くはないのですが, 立地と朝ご飯で減点です. お風呂も良くなかったような, だいぶ昔に泊まったのですが, ちょっとリピートは無いです.	0

表 6　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 7　交差検証での正答率 (5 分割)

0.8527	0.8554	0.8642	0.8607	0.8424

表 8　交差検証での正答率 の平均と分散

平均	分散
0.8551	5.6262e-05

表 9　2 値分類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