研究進捗発表　2018年10月

LDAを用いたAmazonのレビューデータのデータマイニング

B8EM1016　富田優(とみたゆう)

アウトライン

* ①はじめに
* ②前回のおさらい
* ③使用したデータ
* ④モデル適用の結果
* ⑤考察

はじめに

* 発表者：富田優
* 所属：経済学研究科1年
* 指導教員：石垣先生
* POSデータやレビューデータの分析をといったマーケティング・リサーチの分野に進む予定
* 興味ある分野：統計学、ベイズ統計、機械学習

○今日発表する内容

5月に紹介した「トピックモデルを用いた商品の評判要因分析に関する検討」という論文をもとに進めている研究の発表

前回のおさらい①

項目選択方式

◯メリット

* モニターの負担が軽い

◯デメリット

* 事前に項目を決める必要
* サンプル数が必要
* 人的労力と金銭的費用が大

自由記述方式

◯メリット

* 事前に想定できなかった評判要因を知られる

◯デメリット

* モニターの負担が重い
* 解析に労力が必要
* 多変量解析などの統計的解析手法が使いにくい

前回のおさらい②

ECサイト上のユーザーレビュー　　楽天トラベルのサイト

○メリット

* 容易に多くのデータを収集可能
* 統計処理しやすい評点情報
* 自由記述であるレビュー情報

前回のおさらい③
～トピックモデル～

パラメータΦが与えられたときの文書集合Wの確率は以下の通り

$$p(\boldsymbol{w}|\boldsymbol{\theta_d}, \boldsymbol{\Phi})$$

$$= \prod_{n=1}^{N_d} \prod_{k=1}^{K} p(z_{dn} = k|\theta_d)\, \boldsymbol{p(w_{dn}|\Phi_k)}$$

$$= \prod_{n=1}^{N_d} \sum_{k=1}^{K} \theta_{dk}\, \varphi_{kw_{dn}}$$

あとはこのθ_{dk}, $\varphi_{kw_{dn}}$,Kをデータから推定する

W:文書集合

Φ:φ_vのベクトル表示

w_d:文書dの単語集合

Φ_v:単語vが出現する確率

N_d;文書dに含まれる単語数

w_{dn}:文書dのn番目の単語

$\Phi_{w_{dn}}$:文書dのn番目の単語が出る確率

前回のおさらい④

○やりたいこと

　レビューデータをもっと有効活用して、消費者の商品に対する判断基準を理解し、さらなる購買につなげたい

○具体的な手法

　LDA(トピックモデル)を用いてトピック分布と単語分布を推定

○回帰分析

　商品のレーティングを、トピック分布、単語分布、価格、消費者の属性に回帰する

○どの評判要因がレーティングに影響しているのか分析

実験データ

Amazon.comのレビューデータ	
分類	Musical instruments
レビュー数	10,213
総語彙数	22756

	asin	helpful.0	helpful.1	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>
1	B00000JBLH	3	4	5	I bought m~	09 3, 2004	A32T2H815~	ARH	A soli~	1094169600
2	B00000JBLH	7	9	5	"WHY THIS ~	12 15, 20~	A3MAFS04Z~	"Let it Be ~	Price ~	1197676800
3	B00000JBLH	3	3	2	I have an ~	01 1, 2011	A1F1A0QQP~	Mark B	Good f~	1293840000
4	B00000JBLH	7	8	5	I've start~	04 19, 20~	A49R5DBXX~	R. D Johnson	One of~	1145404800
5	B00000JBLH	0	0	5	For simple~	08 4, 2013	A2XRMQA6P~	Roger J. Bu~	Still ~	1375574400
6	B00000JBLH	10	12	5	While I do~	01 23, 20~	A2JFOHC9W~	scott_from_~	Every ~	1011744000
7	B00000JBLH	3	4	5	I've had a~	01 17, 20~	A38NELQT9~	W. B. Halper	A work~	1168992000
8	B00000JBLH	0	0	5	Bought thi~	11 14, 20~	AA8M6331N~	ZombieMom	Fast s~	1384387200
9	B00000JBLU	3	3	5	This is a ~	12 7, 2010	A25C2M3QF~	Comdet	Nice d~	1291680000
10	B00000JBLU	0	0	5	I love thi~	12 2, 2013	A1RTVWTWZ~	"Hb \"Black~	Love I~	1385942400

Perplexity

* *Perplexityは分岐数または選択肢の数を表している*
* *モデルによって単語の候補をどれだけ絞り込めるか*
* *より絞り込める方がよい→値が低い方がよい*
* $L\left(\boldsymbol{w^{test}}\middle|\boldsymbol{M}\right)=\sum_{d=1}^{M}\sum_{w_{d,i}\in w_d^{test}}\log p\left(w_{d,i}\middle|M\right)$
* $PPL\left(\boldsymbol{w^{test}}\middle|\boldsymbol{M}\right)=exp\left\{-\frac{L\left(\boldsymbol{w^{test}}\middle|\boldsymbol{M}\right)}{\sum_{d=1}^{M}n_d^{test}}\right\}$
* $\boldsymbol{w^{test}}$**:テストデータの単語集合**
* $\boldsymbol{M}$ **:学習されたモデル**
* $w_{d,i}$**:ドキュメント**d**の**i**番目の単語**

トピック数とパープレキシティ

Topic	Number of vocabulary	Perplexity (train)	Perplexity (test)
10	22756	1352.709	1547.395
15	22756	1305.474	1493.553
20	22756	1273.079	1461.516
25	22756	1247.578	1436.943

回帰①

- 多重回帰
- モデル
- $R_i = \alpha_0 + \sum_{k=1}^{14} \beta_k \theta_{ik}$
- R_i　i番目のレビューの評点
- α_0　定数項
- β_k　トピックKの回帰係数
- θ_{ik}　i番目のレビューのトピックKの確率

回帰① 結果

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.8734 -0.4196  0.3671  0.5237  1.7290

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6675     0.3548   4.699 2.64e-06 ***
topic1        4.2448     0.4505   9.422  < 2e-16 ***
topic2        3.3025     0.4633   7.128 1.09e-12 ***
topic3        2.4109     0.4459   5.407 6.55e-08 ***
topic4        3.5780     0.4373   8.183 3.11e-16 ***
topic5        3.8634     0.4384   8.813  < 2e-16 ***
topic6        2.7120     0.4296   6.312 2.86e-10 ***
topic7       -0.6614     0.5414  -1.222  0.22183
topic8        3.2352     0.4317   7.494 7.24e-14 ***
topic9       -1.9210     0.5107  -3.761  0.00017 ***
topic10       3.8186     0.4994   7.646 2.25e-14 ***
topic11       4.7641     0.4917   9.689  < 2e-16 ***
topic12       4.8655     0.5126   9.492  < 2e-16 ***
topic13       3.6773     0.4555   8.074 7.60e-16 ***
topic15       4.3723     0.4665   9.372  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8738 on 10198 degrees of freedom
Multiple R-squared:  0.04974,   Adjusted R-squared:  0.04844
F-statistic: 38.13 on 14 and 10198 DF,  p-value: < 2.2e-16
```

回帰②

* 順序ロジスティック回帰
* モデル
* $y_i = \frac{1}{1+e^{-z_i}}$
* $z_i = \alpha_0 + \sum_{k=1}^{14} \beta_k \theta_{ik}$
* y_i　i番目のレビューの評点(ただし、$y_i \begin{cases} 1\ if\ R_i \geq 5 \\ 0\ otherwise \end{cases}$)
* $R_i \geq 5,4,3,2$ と4パターンで推定

4パターンでβ_kは同じと仮定

α_0は4種類

回帰② 結果

```
Coefficients:
         Value Std. Error t value
topic1   8.752      1.115   7.848
topic2   5.720      1.104   5.182
topic3   3.813      1.053   3.623
topic4   6.750      1.052   6.414
topic5   7.420      1.063   6.982
topic6   4.801      1.013   4.738
topic7  -5.312      1.254  -4.235
topic8   6.009      1.035   5.806
topic9  -3.344      1.203  -2.780
topic10  8.286      1.227   6.753
topic11  8.157      1.197   6.817
topic12 10.239      1.255   8.156
topic13  6.839      1.083   6.314
topic15  9.416      1.148   8.202

Intercepts:
    Value   Std. Error t value
1|2  1.2477  0.8484     1.4706
2|3  2.0487  0.8472     2.4182
3|4  3.1256  0.8469     3.6908
4|5  4.4186  0.8472     5.2154

Residual Deviance: 19076.18
AIC: 19112.18
```

回帰結果②

intercept	beta	T-value
1|2	-5.2930	-9.1209
2|3	-4.4895	-7.7663
3|4	-3.4063	-5.9063
4|5	-2.1032	-3.6518
AIC	19025.25	

回帰③

* レーティングをワード分布に回帰
* Lasso推定
* モデル
* $R_i = \alpha_0 + \sum_{v=1}^{22756} \gamma_v \varphi_{iv} I(v)_i$
* R_i　i番目のレビューの評点
* α_0　定数項
* γ_v　単語vの回帰係数
* φ_{ik}　i番目のレビューの単語vの出現確率
* $I(v)_i$ i番目のレビューの単語vの出現回数

回帰③　結果

term	beta	term	beta	term	beta
even	-1.667771162	back	-4.855232252	star	-13.65773519
better	-1.345167395	year	1.594289593	cheap	-9.782175282
need	2.29803766	perfect	11.12399192	bad	-8.164416925
high	1.770711918	problem	-1.191243457	unit	-2.494619464
nice	1.500921608	becaus	-4.15477247	pig	-1.350565567
onli	-5.926596794	best	7.323327732	ubass	-1.156704462
great	1.494628887	seem	-4.279183092	ordering	1.429908253
get	-1.938838758	can	1.568680456	squeezing	-1.986226015
love	4.079614215	alway	1.895679879	apt	-4.644972477
tri	-9.813784969	someth	-8.999300582	onstagegranted	-21.30583837
end	-2.179060762	easi	5.88910488	bonamassa	1.101577968

回帰④

* レーティングをトピック分布とワード分布に回帰
* Lasso推定
* モデル
* $R_i = \alpha_0 + \sum_{k=1}^{13} \beta_k \theta_{ik} + \sum_{v=1}^{22756} \gamma_v \varphi_{iv} I(v)_i$
* R_i i番目のレビューの評点
* α_0 定数項
* β_k トピックKの回帰係数
* γ_v 単語vの回帰係数
* θ_{ik} i番目のレビューのトピックKの確率
* φ_{ik} i番目のレビューの単語vの出現確率
* $I(v)_i$ i番目のレビューの単語vの出現回数

考察

* ①15のトピックを得られたが、トピックの解釈が難しい
* ②トピック抽出にランダム性があるので、ロバストじゃない
* ③モデルの推定と回帰は一緒にやった方がいいのではないか
* ④レーティングを上げる、あるいは下げる傾向があるトピック・単語の回帰係数を得られたが、商品ごとに用いた方が係数の解釈しやすいand他のデータを使いやすい(価格データ等)のではないか

課題

1. トピック数の決め方　階層ディリクレモデル
2. 回帰と一緒にトピックモデルを推定
3. 回帰の仕方　0過剰ポアソン分布、順序ロジットでスパース推定
4. 商品ごとにデータを分けるべきなのか　(商品に対する単語のイメージを量的に抽出と捉えることができる)
5. 単語を制限するべきか(名詞と形容詞のみにする等)
6. ノイズ(やらせ)の混入
7. 階層モデルへの拡張
8. 他の変数をどう入れるべきか

参考

1. Ldaのモデル選択におけるperplexityの評価(東京農工大学工学部情報工学科2年 森尾 学 2016/02/01)
2. トピックモデルによる統計的潜在意味解析(奥村2013)
3. トピックモデルを用いた商品の評判要因分析に関する検討(月岡2013et al)