発展: テキスト分析の紹介

川田恵介

Table of contents

1 **テキストデータ** 2
- 1.1 実例 . . . 2
- 1.2 テキスト分析 . . . 2
- 1.3 例：Open-end question . . . 2
- 1.4 例: 武蔵大学の印象 . . . 2
- 1.5 本講義での応用 . . . 2

2 **前処理** 3
- 2.1 復習: 母平均の推定問題 . . . 3
- 2.2 VS カテゴリー変数 . . . 3
- 2.3 VS 連続変数 . . . 3
- 2.4 テキスト変数の難しさ . . . 4
- 2.5 テキストデータを使った予測モデル構築 . . . 4
- 2.6 前処理: Token . . . 4
- 2.7 例: 武蔵大学の印象 . . . 4
- 2.8 前処理：Bag of words . . . 5
- 2.9 例: 武蔵大学の印象 . . . 5

3 **頻度分析** 5
- 3.1 全体 . . . 5
- 3.2 グループの特徴づけ . . . 6
- 3.3 一般 . . . 6
- 3.4 専門 . . . 6

4 **線型予測モデル** 7
- 4.1 OLS の前提条件 . . . 7
- 4.2 LASSO . . . 7
- 4.3 実例 . . . 7
- 4.4 実例: LASSO . . . 7

Type	Title
一般	新型コロナ対策としてのマスク着用義務化——アメリカの政策評価と日本への示唆
一般	現状放置なら５月下旬に緊急事態宣言の恐れ
専門	緊急事態宣言、ステージ II まで継続を
専門	日本の新型コロナウイルス感染症拡大の現状と感染リスク

4.5 まとめ 7

1 テキストデータ

1.1 実例

- 日本 & 経済学におけるコロナ研究リスト

1.2 テキスト分析

- 事例の多くは、” テキスト” で記録されてきた
  - 電子化された膨大なテキストデータが容易に入手可能（SNS, オンラインアンケート、口コミ評価）
- テキスト分析への需要拡大
  - 統計・機械学習的手法の応用が急速に進む

1.3 例：Open-end question

- 多くの調査には、自由記述欄が含まれる
  - 例：商品への感想を書いてください
- 回答結果を制約しない
  - Closed-end: 回答できる範囲を研究者が指定 (例: 年齢)
  - いろんな回答結果が記録できる
  - 分析者が想定していない情報を得られる可能性

1.4 例: 武蔵大学の印象

1.5 本講義での応用

- 特定のグループが使いがちな単語とは？

ID	Text
1	キャンパスが綺麗
2	池袋から近い
3	キャンパスがおしゃれ

- 予測モデル $Y \sim$ テキスト の構築
  - “変数の数” が爆発的に増加するので、機械学習に比較優位

2 前処理

2.1 復習: 母平均の推定問題

- 基本: “似ている” 複数の事例を集計して、母平均を推定する
  - 例: 練馬区の物件は” 似ている”
- 予測変数 X の役割 = 似ている事例かどうかの判断基準
  - 統計・機械学習 = データから似ている度を判定する
- 伝統的な X : カテゴリー/連続変数
  - テキスト変数は大きく異なる特徴を持つ

2.2 VS カテゴリー変数

- 性別、国籍、学部など
- カテゴリー変数 = “少数” の値しかない
  - 同じ値をとるサンプルが複数存在
  - “値が同じであれば似ている”
- テキスト変数 = 無限大の種類がある

2.3 VS 連続変数

- 年齢、身長など
- 連続変数 = 同じ値をとるサンプルは極めて少数だが、
  - 値が近いかどうかは自明
  - 例: 160cm は、190cm よりも、161cm と似ている

- テキスト変数
  - テキストが近いかどうかは自明ではない

2.4 テキスト変数の難しさ

- テキストは情報が” 豊富” すぎるため、変数の値間の距離が定義できない
  - 似ている文章とは？
- 何らかの単純化が必要

2.5 テキストデータを使った予測モデル構築

- 合わせ技でモデル構築
  - 事前に $Text$ を単純化する (前処理)
  - 大量の X に対応した手法で推計 (機械学習)

2.6 前処理: Token

- テキストを単語の羅列として単純化（Token 化）
- 日本語は単語化が難しい
  - 分かち書きをしない
  - quanteda パッケージを用いれば解決可能

2.7 例: 武蔵大学の印象

```
Tokens consisting of 3 documents and 1 docvar.
text1 :
[1] "キャンパス" "が"          "綺麗"

text2 :
[1] "池袋" "から" "近い"

text3 :
[1] "キャンパス" "が"          "おしゃれ"
```

2.8 前処理：Bag of words

- Token 化しただけでは、依然として、全ての事例が異なる値を有する
  - さらなる単純化が必須
- 代表的な手法は、Bag of words
  - 単語の出現頻度を数える
- 文脈や語順は捨象
  - 発展: N-gram, embedding

2.9 例: 武蔵大学の印象

```
Document-feature matrix of: 3 documents, 7 features (57.14% sparse) and 1 docvar.
       features
docs    キャンパス が 綺麗 池袋 から 近い おしゃれ
  text1          1  1    1    0    0    0        0
  text2          0  0    0    1    1    1        0
  text3          1  1    0    0    0    0        1
```

3 頻度分析

- どのような単語が使われているか
  - グループごとに集計も可能
  - テキスト分析版、記述統計分析

3.1 全体

```
    の  コロナ    と  新型  感染    —    に    を
  167    96    64    55    39    31    31    29
 経済    症    :    ・ ウイルス  分析  禍  影響
   29    27    27    25    24    22    22    21
における    へ    で  企業
   19    16    15    15
```

- よくわからない

3.2 グループの特徴づけ

- 単純に集計すると助詞や助動詞など、“文章を特徴づける上で、そこまで重要ではない単語” が上位に来る
- グループ (例: 満足度が高い VS 低い, **一般むけ** VS **専門むけ**) を特徴づける単語の探索
- chi2 指標: 単語がグループ間で偏りなく使用される場合に比べて、分布がどの程度偏っているのか？

3.3 一般

	feature	chi2	p	n_target	n_reference
1	「	7.305724	0.006873528	8	1
2	」	7.305724	0.006873528	8	1
3	**学**	7.305724	0.006873528	8	1
4	**で**	7.285821	0.006950095	11	4
5	2	5.373110	0.020449490	5	0
6	pos	5.373110	0.020449490	5	0
7	**みる**	5.373110	0.020449490	5	0
8	**経済**	4.550576	0.032907684	17	12
9	**格差**	3.882136	0.048802424	4	0
10	**編**	3.882136	0.048802424	4	0

3.4 専門

	feature	chi2	p	n_target	n_reference
1	**における**	6.704033	0.009619509	17	2
2	**調査**	3.727837	0.053512636	12	2
3	**た**	3.699188	0.054438951	8	0
4	**データ**	3.077562	0.079379450	10	1
5	**感染**	2.092691	0.148005189	28	11
6	**い**	1.815805	0.177813428	5	0
7	**による**	1.815805	0.177813428	5	0
8	**関係**	1.815805	0.177813428	5	0
9	**行動**	1.745022	0.186503736	10	2
10	**ウィルス**	1.435926	0.230799676	7	1

4 線型予測モデル

- $Y \sim$ テキスト を推定する
- 変数の数が多すぎて (事例数を超える)、伝統的な手法は一般に機能しない
  - テキスト分析が難しかった理由
  - LASSO などの大量の変数に対応した手法を活用

4.1 OLS の前提条件

- $Y_i \sim \beta_0 + .. + \beta_L X_L$
  - ざっくり、事例数 $> 3\times$ 変数数であれば、ある程度の推定精度を期待できる
- 変数数 $>$ 事例数であれば、原理的に推定できない
  - テキスト分析ではしばしば発生する

4.2 LASSO

- 変数数 $>$ 事例数でも推定できる

4.3 実例

- $Y = 1$ 一般むけの文章, $Y = 0$ 専門家向けの文章
- $X =$ 論文のタイトル

4.4 実例: LASSO

- 一般むけ/専門的を予測する単語

```
          pos          集中            い            た          関係
 5.340909e-01  5.340909e-01  6.869854e-15 -4.659091e-01 -4.659091e-01
        家計          による          組織
-4.659091e-01 -4.659091e-01 -4.659091e-01
```

4.5 まとめ

- 前処理 + 機械学習が最有力な選択肢

- 大量の発展的議論
  - 経済学における量的テキスト分析入門
  - Text as Data
  - Text Algorithms in Economics