テキスト分析
機械学習

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-11-04

1 テキスト変数

1.1 非伝統的データ

- データ分析 = 収集された事例から学ぶ方法
  - 事例は”数字”以外でも記録される
  - テキスト: 日本においては少なくとも 1500 年前から利用 (稲荷山古墳出土鉄剣)
    - 近年はインターネット/SNS の発展もあり大量のテキスト情報が利用可能に
  - 動画/音声/画像: 近年、積極的に利用されている
- 機械学習を応用した分析が主流

1.2 例: COVID 関連論文

- Y = シリアル値 (1900 年 1 月 1 日を 1 とした日付の通し番号)、$D = 1$ (一般向け)、0 (専門家向け)

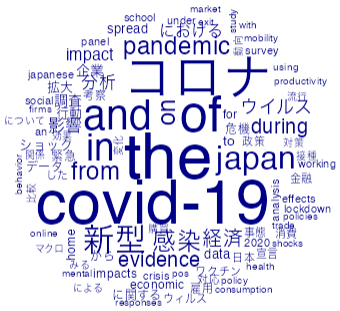
```
library(tidyverse)

read_csv("Public/text.csv")
```

```
# A tibble: 264 × 3
       Y     D
Title
   <dbl> <dbl>
<chr>
 1 44049     1 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……
 2 44273     1 現状放置なら５月下旬に緊急事態宣言の恐れ
 3 44105     1 接触確認アプリCOCOAを行動経済学で読み解く
 4 44274     1 新型コロナ対策がもたらす効果の定量的分析～緊急事態宣言解除後のシナリオとは？
```

```
〜……
 5 44028      1 新型コロナ危機と経済政策
 6 44232      1 感染症データにおける「従属性」と因果推論
 7 44018      1 「疫学」と「マクロ経済学」の視点から –最新論文に見る感染症対策と経済活動維持
の最適解とは–……
 8 43929      1 コロナ危機は需要ショックなのか供給ショックなのか？
 9 44029      1 コロナ危機の経済学 ： 提言と分析
10 44044      1 新型コロナウイルスのマクロ経済学 (1) 感染症拡大防止政策のトレードオフ……
# i 254 more rows
```

1.3 例 ワードクラウド



1.4 例

- 研究のサイクル: 研究発信は、研究者向け (論文等) と一般向け (記事等) (D) に大別できる
  - どちらが先 (Y) に発信されるか?
  - 同じような研究課題の場合は?
    - タイトルの情報 (X) から判断

1.5 実際の例

- 予測研究: 書類選考への活用
  - エントリーシートの内容 (X) から、「人物像/剽窃の有無」 (Y) などを予測

1.6 実際の例

- 物価研究: 消費者物価指数
  - “同じような消費生活をする (X)”場合に、去年と今年 (D) で必要な費用 (Y) の比較
  - 長年の課題は、商品の質の評価 (ステレス値上げ/品質の改善)
    - 例: PC の値段は、品質を考慮した場合、低下しているかもしれない
- Bajari et al. (2023) : 商品の質を画像やテキストなどから測定

1.7 個人的経験

- 実務家からの関心をもたれやすい
  - データ分析 = “数値分析”という誤ったイメージのせいで、“利用できないと思われていた情報”が活用できるのは魅力的?
- ここまで学習してきた手法が使えるので、学生時代に、チャレンジしてみるのもおすすめ

2 課題

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”
- 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
- “少数”の値しか存在しないのであれば
  - 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、異なるのであれば似ていない
- 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
- 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 = テキストが近いとは?

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、変数の値が似ているかどうかが不明確
  - 似ている文章とは？
- 何らかの単純化が必要

3 分析手順

3.1 前処理

1. $Text$を単純な変数に置き換える (前処理)
2. 大量のXに対応した手法で推計 (機械学習の活用)

3.2 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

3.3 例: 武蔵大学の印象

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

3.4 例: 武蔵大学の印象

```
library(tidyverse)

library(quanteda)

corp <- corpus(Example, text_field = "Text")

corp
```

```
Corpus consisting of 3 documents and 1 docvar.
text1 :
"キャンパスが綺麗"

text2 :
"池袋から近い"

text3 :
"キャンパスがおしゃれ"
```

3.5 例: 武蔵大学の印象

```
token <- tokens(corp)

token
```

```
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"
```

3.6 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, word-embedding

3.7 例: 武蔵大学の印象

```
dfm <- dfm(token)

convert(dfm, "matrix")
```

```
       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
```

- “単語を変数とする”データに加工

3.8 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

3.9 頻度

```
topfeatures(dfm)
```

```
キャンパス          が       綺麗       池袋       から       近い    おしゃれ
         2           2          1          1          1          1          1
```

3.10 ワードクラウド

```
library(quanteda.textplots)

textplot_wordcloud(dfm, 20,min_count = 1)
```

4 線型予測モデル

4.1 テキストによる予測

- $Y \sim$ 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

4.2 OLS の前提条件

- 複雑なモデルの推定に向いてない
  - 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

4.3 LASSO

- 複雑なモデルの推定に向く
  - 変数数 $>$ 事例数でも推定できる

4.4 実例: LASSO

- 効果年を予測する単語

```
Model = hdm::rlasso(
  y = data$Y,
  x = X, # テキストを加工した変数
  post = FALSE
)

Coef= coef(Model)[-1]

Coef[Model$index == TRUE]
```

```
     コロナ           を           は     ショック         pos        購買
-74.7745061 -12.6727723  -0.1227596  -0.8245517 -89.5733217 -95.8838353
         in     effects
  6.5622978   2.7902845
```

4.5 実例: 二重選択

- 効果年を予測する単語

```
Model = hdm::rlassoEffect(
  y = data$Y,
  d = data$D,
  x = X # テキストを加工した変数
)

confint(Model)
```

```
       2.5 %    97.5 %
d1 -261.2084 -149.1888
```

4.6 実例: 単純比較

```
estimatr::lm_robust(Y ~ D, data)
```

```
              Estimate Std. Error   t value     Pr(>|t|)   CI Lower   CI Upper
(Intercept) 44304.6651   12.48143 3549.6472 0.000000e+00 44280.0884 44329.2417
D            -228.0651   19.71466  -11.5683 2.813002e-25  -266.8844  -189.2457
             DF
(Intercept) 262
D           262
```

4.7 Takeaway

- テキストを X として用いる方法については、多くの議論が蓄積されている

- Y ないし D として用いる方法は、より難しい

- 基本的なアイディアは、(1) 前処理によりある程度単純化、 (2) 大量の変数を扱える方法で推定
  - 本スライドでは、シンプルな方法 (bug of words)を紹介

4.8 Reference

Bibliography

Bajari, P. et al. (2023) “Hedonic prices and quality adjusted price indices powered by AI,” arXiv preprint arXiv:2305.00044 [Preprint].