Text as Data as Measurement

William Lowe

Hertie School of Governance

14th September 2020

Last week

Last week we talked rather abstractly about models that connected the 'message' θ and the words W (or whatever features we decided to treat as exchangeable)

Let's be a bit more specific

Decisions, decisions

Are we modeling

- → the generation process
- → the understanding process
- → or maybe both…

DECISIONS, DECISIONS

Are we modeling

- → the generation process
- → the understanding process
- → or maybe both…

Por qué no los dos?

$P(\theta)$	*Prior expectations*
$P(\{W\} \mid \theta)$	*Generation*
$P(\theta \mid \{W\}) = \dfrac{P(\{W\} \mid \theta)P(\theta)}{\int P(\{W\} \mid \theta)P(\theta)d\theta}$	*Understanding*

Decisions, decisions

Examples:

Document classification: θ is the probability that this document is about social policy

- → Naive Bayes Classification, learn all the things
- → (Regularized) Logistic Regression, go straight for $P(\theta \mid \{W\})$

Thematic analysis: θ is the proportion of social policy mentions in the document

- → Topic Models, learn all the things
- → Content Analysis Dictionaries, assert $P(\{W\} \mid \theta)$ and go straight for $P(\theta \mid \{W\})$

We'll take a closer look at thematic analysis next week, so let's look at classification

Document classification

Naive Bayes:

Let Z be one of *two* possible document topics

$$P(\{W\} \mid Z) = \prod^{V} P(W_v \mid Z) \qquad \textit{The naive part}$$
$$P(Z = 1) = \theta$$

Now we see a

Naive Bayes

Estimating the probability that a word profile $\{W\}_j$ occurs given that the document is liberal $P(\{W\}_j \mid Z = \text{'Lib'})$ is more challenging, because any one word profile is likely to occur only once.

Assumption: words are assumed to be generated *independently* given the category Z

$$
\begin{aligned}
P(\{W\}_j \mid Z = \text{'Lib'}) &= \prod_i P(W_i \mid Z = \text{'Lib'}) \\
P(\text{'Affirmative Action'} \mid Z = \text{'Lib'}) &= P(\text{'Affirmative'} \mid Z = \text{'Lib'}) \cdot \\
&\quad P(\text{'Action'} \mid Z = \text{'Lib'})
\end{aligned}
$$

Naive Bayes

With this assumption, we can estimate the probability of observing a word i given that the document is liberal: proportion of word i in liberal training set.

The classifier then chooses the class Z (Liberal or Conservative) with the highest aggregate probability.

Note that every new word adds a bit of information that re-adjusts the conditional probabilities.

Naive Bayes

Note that with two classes (here: liberal and conservative) this has a rather neat interpretation:

$$\frac{P(Z = \text{`Lib'} \mid \{W\}_j)}{P(Z = \text{`Con'} \mid \{W\}_j)} =$$

$$\prod_i \frac{P(W_i \mid Z = \text{`}Lib\text{'})}{P(W_i \mid Z = \text{`}Con\text{'})} \times \frac{P(Z = \text{`}Lib\text{'})}{P(Z = \text{`}Con\text{'})}$$

Logging this probability ratio, every new word *adds* a bit of information that pushes the ratio above or below 0

Naive Bayes

Example: Naive Bayes with only word class 'discriminat*'.

$$P(W = \text{'discriminat*'} \mid Z = \text{'Lib'}) = (26 + 13)/(20002 + 18722) \approx 0.001$$
$$P(W = \text{'discriminat*'} \mid Z = \text{'Con'}) = (70 + 48)/(17368 + 17698) \approx 0.003$$

Assume that liberal and conservative supporting briefs are equally likely (true in the training set)

$$\frac{P(Z = \text{'Lib'})}{P(Z = \text{'Con'})} = 1$$

Last step: calculate posterior classification probabilities for a new document (based on occurrence of this word).

Naive Bayes

Amicus brief from ‘King County Bar Association’ containing 3667 words and 4 matches to disciminat*.

```
that "the state shall not [discriminate] against, or grant preferential treatment
the lingering effects of racial [discrimination] against minority groups in this
remedy the effects of societal [discrimination]. Another four Justices (Stevens
that "the state shall not [discriminate] against, or grant preferential treatment
```

Naive Bayes

A priori, the probabilities are…

Probability that we observe the word discriminat* 4 out of 3667 times if the document is liberal:

```
> dbinom(4, size=3667, prob=0.001007127)
[1] 0.1930602
```

Probability that we observe the word discriminat* 4 out of 3667 times if the document is conservative:

```
> dbinom(4, size=3667, prob=0.003365083)
[1] 0.004188261
```

Logged probability ratio = 3.83

Naive Bayes

Conclusion: Seeing 4 instances of discriminat* gives the posterior classification probabilities

- → $\theta_{\text{liberal}} = \frac{0.193}{0.193+0.004} = 0.979$
- → $\theta_{\text{conservative}} = 1\text{-}0.979\text{=}0.021$

This is *quite* confident

- → …but other words will be less loaded or push the other way