Text as Data as Measurement

William Lowe

Hertie School of Governance

15th 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?

$$P(\theta) \qquad \textit{Prior expectations}$$

$$P(\{W\} \mid \theta) \qquad \textit{Generation}$$

$$P(\theta \mid \{W\}) = \frac{P(\{W\} \mid \theta)P(\theta)}{\int P(\{W\} \mid \theta)P(\theta)d\theta} \qquad \textit{Understanding}$$

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

Example

Example application: Evans et al. (2007) attempt to

- → Distinguish the amicus briefs from each side of two affirmative action cases: Regents of the University of California v. Bakke (1978) and Grutter/Gratz v. Bollinger 2003.
- → Characterize the language used by each side

We can label the Plaintiff as 'Conservative' and the Respondents as 'Liberal'

> *All told, Bakke included 57 amicus briefs (15 for the conservative side and 42 for liberals) and Bollinger received 93 (19 conservative and 74 liberal).*
>
> *(Evans et al., 2007)*

The four briefs of Plaintiffs and Respondents formed the 'training data'

Naive Bayes Classification

The document category is $Z \in \{\text{Lib}, \text{Con}\}$

$$P(Z) = \theta \qquad \textit{Prior probability}$$

$$P(\{W\} \mid Z) = \prod_j P(W_j \mid Z) \qquad \textit{The naive part}$$

Words are assumed to be generated *independently* given the category Z

$$P(\text{`Affirmative Action'} \mid Z = \text{`Lib'}) = P(\text{`Affirmative'} \mid Z = \text{`Lib'})P(\text{`Action'} \mid Z = \text{`Lib'})$$

Naive Bayes Classification

The document category is $Z \in \{\text{Lib}, \text{Con}\}$

$$P(Z) = \theta \qquad \textit{Prior probability}$$

$$P(\{W\} \mid Z) = \prod_j P(W_j \mid Z) \qquad \textit{The naive part}$$

Words are assumed to be generated *independently* given the category Z

$$P(\text{`Affirmative Action'} \mid Z = \text{`Lib'}) = P(\text{`Affirmative'} \mid Z = \text{`Lib'})P(\text{`Action'} \mid Z = \text{`Lib'})$$

Classification here means doing something with

$$P(Z \mid \{W\})$$

the *posterior distribution*

→ Strictly, this is just probability estimation. Classification is a separate decision problem.

Naive Bayes

Estimating $\theta = P(Z = \text{Lib}) = (1 - P(Z = \text{Con}))$ is easy.

→ Count the Liberal documents and divide by the total number of documents

Similarly, estimating $P(W_v \mid X)$ is straightforward

$$P(W_j \mid Z = \text{`Lib'}) = \frac{C_j}{\sum_v^{W \in \text{`Lib'}} C_v}$$

where C_v is the count of tokens of W_v.

Actually at this point we have a modeling choice (McCallum & Nigam, 1993):

→ $P(W_j \mid Z = \text{Lib})$ is Binomial

→ $P([W_1 \ldots W_V] \mid Z = \text{Lib})$ is Multinomial

→ Some *transformation* of $P([W_1 \ldots W_V] \mid Z = \text{Lib})$ (e.g. 'tfidf') is Normal

Naive Bayes

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

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

Discrimination

Example: Naive Bayes with only word class ‘discriminat*’.

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

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

and

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

Posterior probability ratio is about 1/3 in favour of the document supporting the conservative side

Conservative vocabulary

Term[a]	*Avg. Freq. per Lib. Brief*	*Avg. Freq per Cons. Brief*	*Chi²*	*Interpretive Code Examples*[b]
Conservative Words				
PREFER*	2.83	41.79	39.18	Proceduralist; Race/Gender Neutral Justice
BENIGN	0.07	1.17	36.14	Intent vs. Consequences; Constraint
DISCRIM*	14.86	25.04	24.13	Proceduralist; Race/Gender Neutral Justice
PURPORT*	0.44	1.88	24.13	Skepticism
CLASSIF*	2.1	11.54	22.39	Proceduralist; Race/Gender Neutral Justice
NARROW-TAILORING	0.05	0.96	19.73	Proceduralist; Strict Scrutiny
REJECT*	2.75	7.79	19.15	Oppositional Posture
JUSTIF*	2.39	12.79	18.91	Proceduralist; Constraint
FORBID*	0.38	1.63	18.91	Proceduralist; Constraint; Race/Gender Neutral Justice
PROHIBITS	0.13	0.71	18.08	Proceduralist; Constraint
RATIONALE	0.66	5.92	17.58	Proceduralist; Legalistic
AMORPHOUS	0.25	1.29	14.62	Proceduralist; Skepticism
RACE-BASED	1.08	10.46	10.59	Proceduralist; Pejorative counterpart to liberal RACE-CONSCIOUS

Liberal vocabulary

Liberal Words				
LEADERS	2.70	0.13	31.03	Impact; Development
WORLD	3.00	0.42	18.74	Impact; Global
NATION*	21.0	7.04	17.90	Impact; Communitarian
IMPACT*	4.13	1.04	17.49	Impact
EFFECTIVE	2.78	0.75	16.54	Impact; Effectiveness
SOCIAL	6.84	1.71	16.05	Impact; Communitarian
COMMUNIT*	8.75	1.75	15.35	Impact; Communitarian
BUSINESS*	4.56	0.58	10.28	Impact; Efficiency; Distributive Justice
DESEGREGATION	2.34	0.17	10.24	Remedial Justice
GROW*	2.38	0.33	10.24	Change; Development
WORKFORCE	1.64	0.00	9.81	Impact; Distributive Justice; Development
RACE-CONSCIOUS	7.14	1.50	7.80	Proceduralist; Euphemistic counterpart to conservative RACE-BASED

- → There are no identifiable *uniquely* partisan words
- → but these associations are stable in cases 28 years apart

Discrimination

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
```

Every generative model

Courtesy of Bayes theorem, the posterior probability of a document being liberal is

$$P(Z = \text{‘Lib’} \mid W_j) = \frac{\prod P(W_j \mid Z = \text{‘Lib’})P(Z = \text{‘Lib’})}{\prod P(W_j \mid Z = \text{‘Lib’})P(Z = \text{‘Lib’}) + \prod P(W_j \mid Z = \text{‘Con’})P(Z = \text{‘Con’})}$$

but let’s do a little rearranging

$$P(Z = \text{Lib} \mid W_j) = \frac{1}{1 + \exp(-\eta)}$$

$$\eta = \log \frac{P(Z = \text{‘Lib’})}{P(Z = \text{‘Con’})} + \sum_j \log \frac{P(W_j \mid Z = \text{‘Lib’})}{P(W_j \mid Z = \text{‘Con’})}$$

which might remind you of a model you’ve seen before…

HAS A DISCRIMINATIVE ALTER EGO

$$P(Z = \text{`Lib'} \mid W_j) = \frac{1}{1 + \exp(-\eta)}$$

$$\eta = \beta_0 + C_1\beta_1 + C_2\beta_2 + \ldots + C_2\beta_V$$

where C_v is the count of word v.

HAS A DISCRIMINATIVE ALTER EGO

$$P(Z = \text{`Lib'} \mid W_j) = \frac{1}{1 + \exp(-\eta)}$$

$$\eta = \beta_0 + C_1\beta_1 + C_2\beta_2 + \ldots + C_2\beta_V$$

where C_v is the count of word v.

This is a logistic regression (Nigam et al., 1999). It's called 'MaxEnt' by computational linguists.

From which we conclude (Jordan, 1999)

- → These are in a certain sense the 'same model'
- → As it happens, *any* exponential family choice for $P(W_j \mid Z)$ has logistic regression as its discriminative model

Naive Bayes and Logit

Logistic regression is more focused

- → No interest in $P([W_1 \ldots W_V])$. Words can be conditionally independent, or not. It just wants the decision boundary

Slower and hungrier

- → β estimates converge at rate N, compared to $\log N$ for Naive Bayes' probability ratios
- → We fit Naive Bayes on four documents. Logistic regression will require *heavy regularization* to work with so many fewer documents than words

Usually better

- → Classification performance is usually better: lower bias, higher variance
- → (Interpretation is trickier)

The model tradeoff

This performance tradeoff is very general:

→ By adding bias (strong assumptions about the data) we can reduce variance

→ By adding flexibility we can reduce bias and have a more expressive model, but we'll need more and better data

The interpretation tradeoff is also general:

→ Better statistical performance often leads to less interpretable models (Chang et al., 2009)

→ In social science applications we usually prefer the interpretable side!

References

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. & Blei, D. M. (2009). 'Reading tea leaves: How humans interpret topic models'. *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 288–296.

Evans, M., McIntosh, W., Lin, J. & Cates, C. (2007). 'Recounting the courts? applying automated content analysis to enhance empirical legal research'. *Journal of Empirical Legal Studies*, *4*(4), 1007–1039.

Jordan, M. I. (Ed.). (1999). 'Learning in graphical models'. MIT Press.

McCallum, A. & Nigam, K. (1993). 'A comparison of event models for naive bayes text classification'. *AAAI/ICML-98 Workshop on Learning for Text Categorization*, 41–48.

Nigam, K., Lafferty, J. & McCallum, A. (1999). 'Using maximum entropy for text classification'. 1–7.