4. Supervised Techniques II (flipped)

DS-GA 1015, Text as Data
Arthur Spirling

March 9, 2021

Housekeeping

Housekeeping

HW 1 out: coming in on March 9, 2021, at 11pm (NY time).

Housekeeping

HW 1 out: coming in on March 9, 2021, at 11pm (NY time).

Naive Bayes

Set up

Set up

We're interested in the probability that an email is in a given category,

Set up

We're interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

Set up

We're interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

Q It is straightforward to calculate $\Pr(d|c)$ (under some simplifying assumptions).

Set up

We're interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

Q It is straightforward to calculate $\Pr(d|c)$ (under some simplifying assumptions). How do we do it?

Set up

We're interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

Q It is straightforward to calculate $\Pr(d|c)$ (under some simplifying assumptions). How do we do it?

A

$$\Pr(d|c) = \prod_{k=1}^{K} \Pr(t_k|c)$$

Set up

We're interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

Q It is straightforward to calculate $\Pr(d|c)$ (under some simplifying assumptions). How do we do it?

A

$$\Pr(d|c) = \prod_{k=1}^{K} \Pr(t_k|c)$$

What (literally) is the $\Pr(t_k|c)$?

Set up

We're interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

Q It is straightforward to calculate $\Pr(d|c)$ (under some simplifying assumptions). How do we do it?

A

$$\Pr(d|c) = \prod_{k=1}^{K} \Pr(t_k|c)$$

What (literally) is the $\Pr(t_k|c)$?

Q But we actually want $\Pr(c|d)$.

Set up

We're interested in the probability that an email is in a given category, given its features—i.e. frequency of terms.

Q It is straightforward to calculate $\Pr(d|c)$ (under some simplifying assumptions). How do we do it?

A

$$\Pr(d|c) = \prod_{k=1}^{K} \Pr(t_k|c)$$

What (literally) is the $\Pr(t_k|c)$?

Q But we actually want $\Pr(c|d)$. Why? What is this?

Reminder: Bayes' Theorem

Reminder: Bayes' Theorem

Reminder: Bayes' Theorem

Recall that:

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that A occurs given that B occurred $=$ the probability of both A and B occurring, divided by the probability that B occurs.

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that A occurs given that B occurred = the probability of both A and B occurring, divided by the probability that B occurs.

e.g. you know a die shows an odd number, what is the probability that this odd number is 3?

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that A occurs given that B occurred $=$ the probability of both A and B occurring, divided by the probability that B occurs.

e.g. you know a die shows an odd number, what is the probability that this odd number is 3? $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}}$

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}$$

- the probability that A occurs given that B occurred $=$ the probability of both A and B occurring, divided by the probability that B occurs.

e.g. you know a die shows an odd number, what is the probability that this odd number is 3? $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that A occurs given that B occurred $=$ the probability of both A and B occurring, divided by the probability that B occurs.

e.g. you know a die shows an odd number, what is the probability that this odd number is 3? $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

- of course, it is also true that $\Pr(B|A) = \frac{\Pr(B,A)}{\Pr(A)}$.

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}$$

- the probability that A occurs given that B occurred $=$ the probability of both A and B occurring, divided by the probability that B occurs.

e.g. you know a die shows an odd number, what is the probability that this odd number is 3? $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

- of course, it is also true that $\Pr(B|A) = \frac{\Pr(B,A)}{\Pr(A)}$.
- but then, since $\Pr(A,B) = \Pr(B,A)$, we must have $\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$, and thus...

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that A occurs given that B occurred = the probability of both A and B occurring, divided by the probability that B occurs.

e.g. you know a die shows an odd number, what is the probability that this odd number is 3? $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

- of course, it is also true that $\Pr(B|A) = \frac{\Pr(B,A)}{\Pr(A)}$.
- but then, since $\Pr(A, B) = \Pr(B, A)$, we must have $\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$, and thus... Bayes' law

Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that A occurs given that B occurred $=$ the probability of both A and B occurring, divided by the probability that B occurs.

e.g. you know a die shows an odd number, what is the probability that this odd number is 3? $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$.

- of course, it is also true that $\Pr(B|A) = \frac{\Pr(B,A)}{\Pr(A)}$.
- but then, since $\Pr(A, B) = \Pr(B, A)$, we must have $\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$, and thus... Bayes' law

$$\boxed{\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}.}$$

And. . .

And...

- interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$.

And...

- interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$.

- But we can re-write like this:

$$\Pr(A|B) \propto \Pr(A)\Pr(B|A)$$

Why?

And. . .

- interest is in $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$.
- But we can re-write like this:

$$\Pr(A|B) \propto \Pr(A)\Pr(B|A)$$

Why?

Here, $\Pr(A)$ is our prior for A, while $\Pr(B|A)$ will be the likelihood for the data we saw.

Exercise

Exercise

1. We know $\Pr(A, B) = \Pr(B, A)$. Can we conclude $\Pr(A|B) = \Pr(B|A)$?

Exercise

1 We know $\Pr(A, B) = \Pr(B, A)$. Can we conclude $\Pr(A|B) = \Pr(B|A)$?

2 If $\Pr(A|B) = \Pr(A)$,

Exercise

1. We know $\Pr(A, B) = \Pr(B, A)$. Can we conclude $\Pr(A|B) = \Pr(B|A)$?

2. If $\Pr(A|B) = \Pr(A)$, what does that tell us about events A and B?

Exercise

1 We know $\Pr(A, B) = \Pr(B, A)$. Can we conclude $\Pr(A|B) = \Pr(B|A)$?

2 If $\Pr(A|B) = \Pr(A)$, what does that tell us about events A and B?

3 A subject claims to have psychic abilities—he can tell you how a (fair) coin will come down in nine tosses. He has less than a $\frac{1}{500}$ chance of being correct by chance, but he succeeds in the task! Do you 'update' that he has psychic abilities? Why or why not?

Recall. . .

Recall...

We can express our quantity of interest as:

$$\Pr(c|d) = \frac{\Pr(c)\Pr(d|c)}{\Pr(d)}$$

and

$$\Pr(c|d) \propto \underbrace{\Pr(c)}_{\text{prior}}$$

Recall. . .

We can express our quantity of interest as:

$$\Pr(c|d) = \frac{\Pr(c)\Pr(d|c)}{\Pr(d)}$$

and

$$\Pr(c|d) \propto \underbrace{\Pr(c)}_{\text{prior}} \underbrace{\prod_{k=1}^{K} \Pr(t_k|c)}_{\text{likelihood}}$$

where $\Pr(c)$ is the prior probability of a document occurring in class c; and $\Pr(t_k|c)$ is interpreted as "measure of the how much evidence t_k contributes that c is the correct class"

BTW n-grams

BTW *n*-grams

Does adding *n*-grams (bigrams) as well as/instead of unigrams help with supervised learning?

BTW *n*-grams

Does adding *n*-grams (bigrams) as well as/instead of unigrams help with supervised learning?

NO! Bekkerman and Allan. "Using Bigrams in Text Categorization". UMass Amherst. 2003.

BTW *n*-grams

Does adding *n*-grams (bigrams) as well as/instead of unigrams help with supervised learning?

NO! Bekkerman and Allan. "Using Bigrams in Text Categorization". UMass Amherst. 2003.

NO! Pang, Lee, and Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86, 2002

BTW *n*-grams

Does adding *n*-grams (bigrams) as well as/instead of unigrams help with supervised learning?

NO! Bekkerman and Allan. "Using Bigrams in Text Categorization". UMass Amherst. 2003.

NO! Pang, Lee, and Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86, 2002

YES! Kushal, Lawrence, and Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." In Proceedings of WWW, pages 519-528, 2003

BTW *n*-grams

Does adding *n*-grams (bigrams) as well as/instead of unigrams help with supervised learning?

- NO! Bekkerman and Allan. “Using Bigrams in Text Categorization”. UMass Amherst. 2003.
- NO! Pang, Lee, and Vaithyanathan. “Thumbs up? Sentiment classification using machine learning techniques.” In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79-86, 2002

- YES! Kushal, Lawrence, and Pennock. “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews.” In Proceedings of WWW, pages 519-528, 2003

- yeah? Tan, Wang, and Lee. “The use of bigrams to enhance text categorization.” Information Processing and Management, 38(4):529-546, 2002.

Different Example

Different Example

	email	words	classification
	1	money actual prince	spam
	2	prince inherit amount	spam
training			

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{ham}) =$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$

$$\Pr(\text{inherit}|\text{ham}) =$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$
$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$
$\Pr(\text{amount}|\text{ham}) =$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$
$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$
$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$

$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$

$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$

$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

$$\Pr(\text{inherit}|\text{spam}) =$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|d) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{inherit}|\text{spam}) =$$

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|d) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|d) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{amount}|\text{spam}) =$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{amount}|\text{spam}) = \frac{1}{6}$$

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{amount}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{spam}|\text{d}) \propto \frac{2}{5}\frac{1}{6}\frac{1}{6}\frac{1}{6} = 0.0028$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{amount}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{spam}|\text{d}) \propto \frac{2}{5}\frac{1}{6}\frac{1}{6}\frac{1}{6} = 0.0028$$

$$\rightarrow \quad c_{map} =$$

Different Example

	email	words	classification
training	1	money actual prince	spam
	2	prince inherit amount	spam
	3	inherit amount money	ham
	4	cost amount amazon	ham
	5	prince william inherit	ham
test	6	inherit inherit amount	?

$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{inherit}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{amount}|\text{ham}) = \frac{2}{9}$$
$$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5}\frac{2}{9}\frac{2}{9}\frac{2}{9} = 0.0065$$

$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{inherit}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{amount}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{spam}|\text{d}) \propto \frac{2}{5}\frac{1}{6}\frac{1}{6}\frac{1}{6} = 0.0028$$

$$\rightarrow \boxed{c_{map} = \text{ham}}$$

Exercise

Exercise

A feature of NB classification is that while the estimated probabilities can be wildly wrong, the classification decisions (the classes to which the documents are assigned) are correct.

Exercise

A feature of NB classification is that while the estimated probabilities can be wildly wrong, the classification decisions (the classes to which the documents are assigned) are correct.

1. Why does this happen?

Exercise

A feature of NB classification is that while the estimated probabilities can be wildly wrong, the classification decisions (the classes to which the documents are assigned) are correct.

1. Why does this happen?

2. What does this imply about the relationship between estimation ('modeling') and accuracy?

Exercise

A feature of NB classification is that while the estimated probabilities can be wildly wrong, the classification decisions (the classes to which the documents are assigned) are correct.

1. Why does this happen?
2. What does this imply about the relationship between estimation ('modeling') and accuracy?
3. Via the *maximum a posteriori* (map) notion, we can easily extend Naive Bayes to multiple classes. Explain how.

Bayesian "Poisoning"

Bayesian "Poisoning"

Q If you wanted to 'fool' a spam filter, but still wanted to get your key words ('viagra', 'fortune' etc) across, what could you do?

Bayesian "Poisoning"

Q If you wanted to 'fool' a spam filter, but still wanted to get your key words ('viagra', 'fortune' etc) across, what could you do? What would you add to the email to make it look legitimate?

Bayesian "Poisoning"

Q If you wanted to 'fool' a spam filter, but still wanted to get your key words ('viagra', 'fortune' etc) across, what could you do? What would you add to the email to make it look legitimate?

A random text perhaps, or (better) anything not associated with spam (Perhaps in white or very small font, and put it at the end—why?).

Bayesian "Poisoning"

Q If you wanted to 'fool' a spam filter, but still wanted to get your key words ('viagra', 'fortune' etc) across, what could you do? What would you add to the email to make it look legitimate?

A random text perhaps, or (better) anything not associated with spam (Perhaps in white or very small font, and put it at the end—why?). What does this do to the likelihood, and thus c_{map}?

Bayesian "Poisoning"

Q If you wanted to 'fool' a spam filter, but still wanted to get your key words ('viagra', 'fortune' etc) across, what could you do? What would you add to the email to make it look legitimate?

A random text perhaps, or (better) anything not associated with spam (Perhaps in white or very small font, and put it at the end—why?). What does this do to the likelihood, and thus c_{map}?

Bayesian "Poisoning"

Q If you wanted to 'fool' a spam filter, but still wanted to get your key words ('viagra', 'fortune' etc) across, what could you do? What would you add to the email to make it look legitimate?

A random text perhaps, or (better) anything not associated with spam (Perhaps in white or very small font, and put it at the end—why?). What does this do to the likelihood, and thus c_{map}?

Q Ultimately, this will *also* mean the spam filter will mark more and more 'ham' emails as spam. What administrative changes would you expect users to make in this case? Why is this good for the spammer?

Bayesian "Poisoning"

Q If you wanted to 'fool' a spam filter, but still wanted to get your key words ('viagra', 'fortune' etc) across, what could you do? What would you add to the email to make it look legitimate?

A random text perhaps, or (better) anything not associated with spam (Perhaps in white or very small font, and put it at the end—why?). What does this do to the likelihood, and thus c_{map}?

Q Ultimately, this will *also* mean the spam filter will mark more and more 'ham' emails as spam. What administrative changes would you expect users to make in this case? Why is this good for the spammer?

A Generally encourages relaxing of spam filter:

Bayesian "Poisoning"

Q If you wanted to 'fool' a spam filter, but still wanted to get your key words ('viagra', 'fortune' etc) across, what could you do? What would you add to the email to make it look legitimate?

A random text perhaps, or (better) anything not associated with spam (Perhaps in white or very small font, and put it at the end—why?). What does this do to the likelihood, and thus c_{map}?

Q Ultimately, this will *also* mean the spam filter will mark more and more 'ham' emails as spam. What administrative changes would you expect users to make in this case? Why is this good for the spammer?

A Generally encourages relaxing of spam filter: more spam.

Confusion Matrix

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Recall : $\frac{\text{number of TP}}{\text{number of TP} + \text{number of FN}} = \frac{a}{a+b}$.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Recall : $\frac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Recall : $\frac{\text{number of TP}}{\text{number of TP} + \text{number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

F : $2\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}$. Harmonic mean of precision and recall.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Recall : $\frac{\text{number of TP}}{\text{number of TP} + \text{number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

F : $2\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}$. Harmonic mean of precision and recall.

FYI...

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.

FYI...

$$\text{Precision}: \frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}.$$

Fraction of the documents predicted to be J, that were in fact J.

FYI...

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.

Fraction of the documents predicted to be J, that were in fact J.

$\rightarrow$ $\boxed{\text{positive predictive value (PPV)}}$

FYI...

Precision : $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.

Fraction of the documents predicted to be J, that were in fact J.

$\rightarrow$ $\boxed{\text{positive predictive value (PPV)}}$

Recall : $\frac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.

FYI...

Precision : $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

$\rightarrow$ $\boxed{\text{positive predictive value (PPV)}}$

Recall : $\frac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

FYI. . .

Precision : $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

$\rightarrow$ $\boxed{\text{positive predictive value (PPV)}}$

Recall : $\frac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

$\rightarrow$ hit-rate,

FYI...

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

$\rightarrow$ $\boxed{\text{positive predictive value (PPV)}}$

Recall : $\frac{\text{number of TP}}{\text{number of TP} + \text{number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

$\rightarrow$ $\boxed{\text{hit-rate, sensitivity}}$

FYI...

Precision : $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

$\rightarrow$ $\boxed{\text{positive predictive value (PPV)}}$

Recall : $\frac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

$\rightarrow$ $\boxed{\text{hit-rate, sensitivity}}$

True Negative Rate: $\frac{\text{number of TN}}{\text{number of FP + number of TN}} = \frac{d}{c+d}$.

FYI...

Precision : $\dfrac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

$\rightarrow$ $\boxed{\text{positive predictive value (PPV)}}$

Recall : $\dfrac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

$\rightarrow$ $\boxed{\text{hit-rate, sensitivity}}$

True Negative Rate: $\dfrac{\text{number of TN}}{\text{number of FP + number of TN}} = \frac{d}{c+d}$.

$\rightarrow$ $\boxed{\text{probability classified as } \neg J \text{ given } \neg J \text{ is correct. Specificity.}}$

Exercise

Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks.

Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks. Fortunately, such emails are very, very rare (0.0001% of all emails).

Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks. Fortunately, such emails are very, very rare (0.0001% of all emails).

1 For such a task,

Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks. Fortunately, such emails are very, very rare (0.0001% of all emails).

1 For such a task, there's probably a trade-off between precision and recall. Explain why.

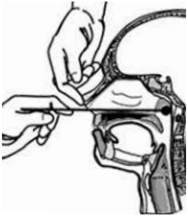
Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks. Fortunately, such emails are very, very rare (0.0001% of all emails).

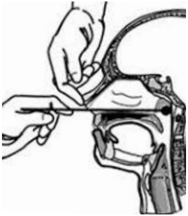
1. For such a task, there's probably a trade-off between precision and recall. Explain why.
2. We may be skeptical of using accuracy as a performance indicator in this case. Explain why.

Covid Testing: Ideal

Covid Testing: Ideal

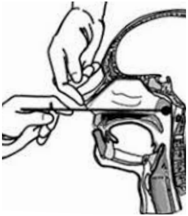


Covid Testing: Ideal



		Predicted	
		positive	negative
Actual	positive	100	0
	negative	0	100

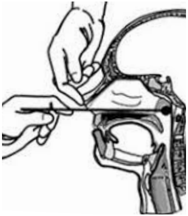
Covid Testing: Ideal



		Predicted	
		positive	negative
Actual	positive	100	0
	negative	0	100

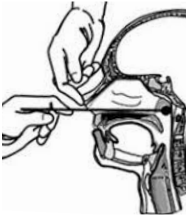
Covid Testing

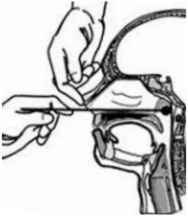
Covid Testing



Covid Testing

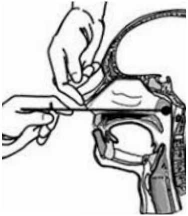
Seems to be around 90-100% sensitivity.





Seems to be around 90-100% sensitivity. So 0-10% false negatives.

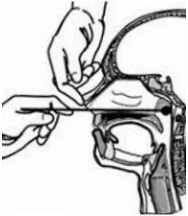
Covid Testing



Seems to be around 90-100% sensitivity. So 0-10% false negatives.

Seems to be around 80-100% specificity.

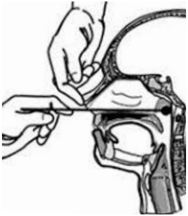
Covid Testing



Seems to be around 90-100% sensitivity. So 0-10% false negatives.

Seems to be around 80-100% specificity. So 0-20% false positives.

Covid Testing

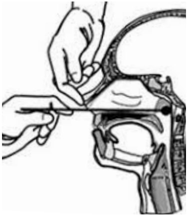


Seems to be around 90-100% sensitivity. So 0-10% false negatives.

Seems to be around 80-100% specificity. So 0-20% false positives.

Q Hard to get exact numbers here—why?

Covid Testing



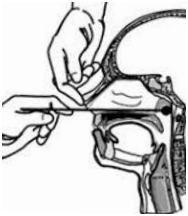
Seems to be around 90-100% sensitivity. So 0-10% false negatives.

Seems to be around 80-100% specificity. So 0-20% false positives.

Q Hard to get exact numbers here—why?

Q We would probably prefer a more sensitive test (for Covid, pregnancy, HIV),

Covid Testing



Seems to be around 90-100% sensitivity. So 0-10% false negatives.

Seems to be around 80-100% specificity. So 0-20% false positives.

Q Hard to get exact numbers here—why?

Q We would probably prefer a more sensitive test (for Covid, pregnancy, HIV), even if it came at the expense of more false positives (and thus reduced specificity). Why?

"99% accuracy"

“99% accuracy”

Culture | Machine Learning

share this

A.I. Algorithm Recognizes Terrorist Propaganda with 99 Percent Accuracy

The war on terror goes digital.

By Kevin Litman-Navarro on February 13, 2018

Filed Under A.I., Algorithms & Data

The UK-based company ASI Data Science unveiled a machine learning algorithm Wednesday that can identify terrorist propaganda videos with 99 percent accuracy.

WordScores

Example

Example

Neo-Nazi manifesto uses ‘`immigrant`’ 25 times in 1000 words, while Communists use it only 5 times.

Example

Neo-Nazi manifesto uses '`immigrant`' 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005}$

Example

Neo-Nazi manifesto uses '`immigrant`' 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

Example

Neo-Nazi manifesto uses '`immigrant`' 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005}$

Example

Neo-Nazi manifesto uses '`immigrant`' 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

Example

Neo-Nazi manifesto uses ‘`immigrant`’ 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so $S_i = 0.83 - 0.16 = 0.66$

Example

Neo-Nazi manifesto uses '`immigrant`' 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party,

Example

Neo-Nazi manifesto uses ‘`immigrant`’ 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions `immigrant` 20 times in a thousand words.

Example

Neo-Nazi manifesto uses '`immigrant`' 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions `immigrant` 20 times in a thousand words.

well the relevant calculation for that word is $0.02 \times 0.66 = 0.0132$.

Example

Neo-Nazi manifesto uses ‘`immigrant`’ 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions `immigrant` 20 times in a thousand words.

well the relevant calculation for that word is $0.02 \times 0.66 = 0.0132$.

but virgin manifesto, from Labour party,

Example

Neo-Nazi manifesto uses ‘`immigrant`’ 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions `immigrant` 20 times in a thousand words.

well the relevant calculation for that word is $0.02 \times 0.66 = 0.0132$.

but virgin manifesto, from Labour party, mentions it 10 times in a thousand words: $0.01 \times 0.66 = 0.006$

Example

Neo-Nazi manifesto uses '`immigrant`' 25 times in 1000 words, while Communists use it only 5 times.

then $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so $S_i = 0.83 - 0.16 = 0.66$

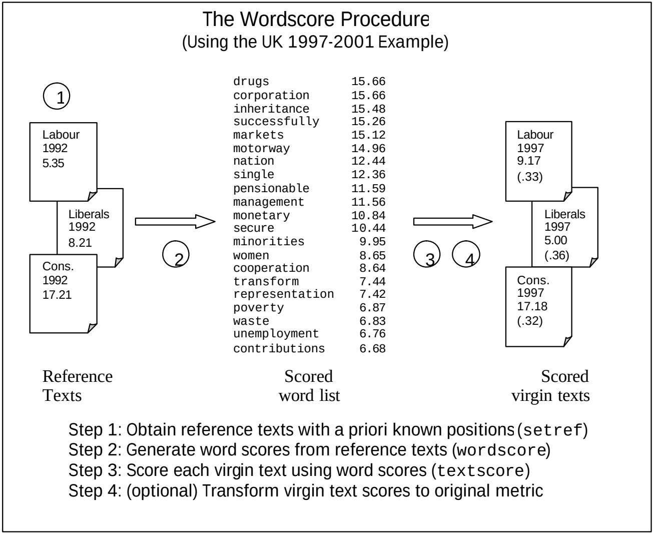
we see a virgin manifesto, from the Conservative party, and it mentions `immigrant` 20 times in a thousand words.

well the relevant calculation for that word is $0.02 \times 0.66 = 0.0132$.

but virgin manifesto, from Labour party, mentions it 10 times in a thousand words: $0.01 \times 0.66 = 0.006$

$\rightarrow$ can rescale these back to original $(-1, 1)$ dimension.

New Labour Moderates its Economic Policy



Shrinkage

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why?

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why? Consider a function word like 'the'.

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why? Consider a function word like 'the'. Do we anticipate that it will be used differently by (far) left vs (far) right?

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why? Consider a function word like 'the'. Do we anticipate that it will be used differently by (far) left vs (far) right? What score will it get?

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why? Consider a function word like 'the'. Do we anticipate that it will be used differently by (far) left vs (far) right? What score will it get? What does that imply about that word?

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why? Consider a function word like 'the'. Do we anticipate that it will be used differently by (far) left vs (far) right? What score will it get? What does that imply about that word?

$\rightarrow$ So the score we give to 'the' will be close to $\bar{\theta}_{\text{ref}}$.

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why? Consider a function word like 'the'. Do we anticipate that it will be used differently by (far) left vs (far) right? What score will it get? What does that imply about that word?

$\rightarrow$ So the score we give to 'the' will be close to $\bar{\theta}_{\text{ref}}$. But this means uninformative words—of which there are generally a very large number—get centrist word scores.

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why? Consider a function word like 'the'. Do we anticipate that it will be used differently by (far) left vs (far) right? What score will it get? What does that imply about that word?

$\rightarrow$ So the score we give to 'the' will be close to $\bar{\theta}_{\text{ref}}$. But this means uninformative words—of which there are generally a very large number—get centrist word scores.

$\rightarrow$ those manifestos get centrist scores.

Shrinkage

"In applications, estimated document scores invariably have a much smaller variance than reference document scores and are bunched around $\bar{\theta}$, the mean of the reference document scores" (Lowe, 2008)

Why? Consider a function word like 'the'. Do we anticipate that it will be used differently by (far) left vs (far) right? What score will it get? What does that imply about that word?

$\rightarrow$ So the score we give to 'the' will be close to $\bar{\theta}_{\text{ref}}$. But this means uninformative words—of which there are generally a very large number—get centrist word scores.

$\rightarrow$ those manifestos get centrist scores.

Attempts to rescale not always very convincing.

Example of WordScores

Example of WordScores

Example of WordScores

Labour manifesto as ‘longest suicide note in history’

Example of WordScores

Labour manifesto as 'longest suicide note in history'

Code as -1

Example of WordScores

Labour manifesto as 'longest suicide note in history'

Code as -1

Conservative manifesto promised trade union curbs, deflation etc.

Example of WordScores

Labour manifesto as 'longest suicide note in history'

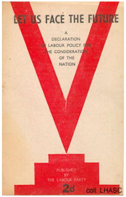
Code as -1

Conservative manifesto promised trade union curbs, deflation etc.

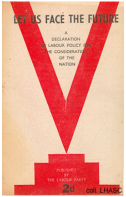
Code as $+1$

Now, the virgin texts. . .

Now, the virgin texts. . .

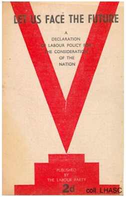


Now, the virgin texts...



Labour, 1945: "The Labour Party is a Socialist Party, and proud of it." Nationalization, welfare state.

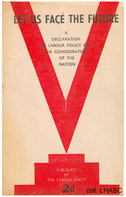
Now, the virgin texts...



Labour, 1945: "The Labour Party is a Socialist Party, and proud of it." Nationalization, welfare state.

Conservative, 1979: stop nationalization, tackling inflation, restricting unions.

Now, the virgin texts...

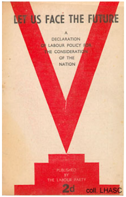


Labour, 1945: "The Labour Party is a Socialist Party, and proud of it." Nationalization, welfare state.

Conservative, 1979: stop nationalization, tackling inflation, restricting unions.

In quanteda...

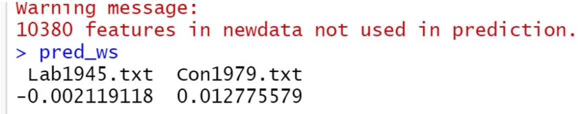
Now, the virgin texts...



Labour, 1945: "The Labour Party is a Socialist Party, and proud of it." Nationalization, welfare state.

Conservative, 1979: stop nationalization, tackling inflation, restricting unions.

In quanteda...



```
Warning message:
10380 features in newdata not used in prediction.
> pred_ws
 Lab1945.txt   Con1979.txt
-0.002119118  0.012775579
```

Crowdsourcing

Galton and the Wisdom of Crowds

Galton and the Wisdom of Crowds

Visits "West of England Fat Stock and Poultry Exhibition" (1906).

Q Guessers have to pay small fee to play, and there are prizes.

Galton and the Wisdom of Crowds

Visits "West of England Fat Stock and Poultry Exhibition" (1906).

Q Guessers have to pay small fee to play, and there are prizes. Why is this important?

Galton and the Wisdom of Crowds

Visits "West of England Fat Stock and Poultry Exhibition" (1906).

Q Guessers have to pay small fee to play, and there are prizes. Why is this important?

→ "sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best."

Galton and the Wisdom of Crowds

Visits "West of England Fat Stock and Poultry Exhibition" (1906).

Q Guessers have to pay small fee to play, and there are prizes. Why is this important?

→ "sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best."

Q There judgements were "uninfluenced by oratory and the like". What does this mean, why is this important?

Galton and the Wisdom of Crowds

Visits “West of England Fat Stock and Poultry Exhibition” (1906).

Q Guessers have to pay small fee to play, and there are prizes. Why is this important?

→ “sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best.”

Q There judgements were “uninfluenced by oratory and the like”. What does this mean, why is this important?

“Now the middlemost estimate is 1207 lb, and the weight of the dressed ox proved to be 1198 lb.”

Q What would you guess the average (mean) error to be?

Galton and the Wisdom of Crowds

Visits "West of England Fat Stock and Poultry Exhibition" (1906).

Q Guessers have to pay small fee to play, and there are prizes. Why is this important?

→ "sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best."

Q There judgements were "uninfluenced by oratory and the like". What does this mean, why is this important?

"Now the middlemost estimate is 1207 lb, and the weight of the dressed ox proved to be 1198 lb."

Q What would you guess the average (mean) error to be? (37lbs)

Crowdsourcing as Concept

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016)

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average),

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not:

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not: this is not a survey to estimate their opinions of the labels—

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not: this is not a survey to estimate their opinions of the labels—we care about the labels themselves.

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not: this is not a survey to estimate their opinions of the labels—we care about the labels themselves.

BTW crowdsourcing can certainly be used for such 'survey' tasks—

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not: this is not a survey to estimate their opinions of the labels—we care about the labels themselves.

BTW crowdsourcing can certainly be used for such 'survey' tasks—see Berinsky et al (2012) for a review of Mechanical Turk for political science use.

Benoit et al. example

Benoit et al. example

Identify Which Of Two Text Segments Contains Easier Language

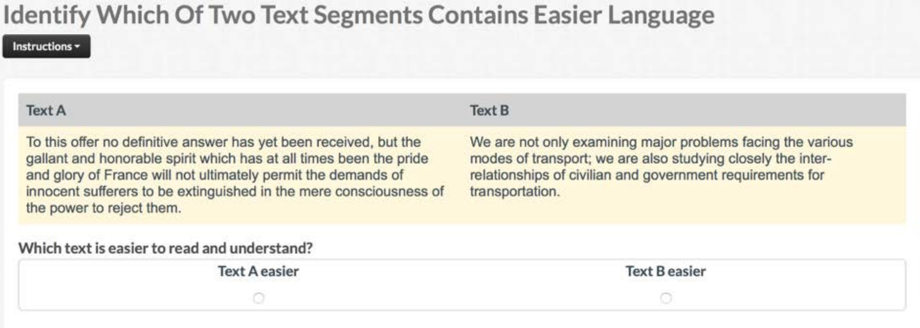
Instructions ▾

Text A	Text B
To this offer no definitive answer has yet been received, but the gallant and honorable spirit which has at all times been the pride and glory of France will not ultimately permit the demands of innocent sufferers to be extinguished in the mere consciousness of the power to reject them.	We are not only examining major problems facing the various modes of transport; we are also studying closely the inter-relationships of civilian and government requirements for transportation.

Which text is easier to read and understand?

Text A easier	Text B easier
○	○

Benoit et al. example



Identify Which Of Two Text Segments Contains Easier Language

Instructions

Text A	Text B
To this offer no definitive answer has yet been received, but the gallant and honorable spirit which has at all times been the pride and glory of France will not ultimately permit the demands of innocent sufferers to be extinguished in the mere consciousness of the power to reject them.	We are not only examining major problems facing the various modes of transport; we are also studying closely the inter-relationships of civilian and government requirements for transportation.

Which text is easier to read and understand?

Text A easier	Text B easier
○	○

Why *pairwise* comparisons?

Benoit et al. example

Identify Which Of Two Text Segments Contains Easier Language

Instructions

Text A	Text B
To this offer no definitive answer has yet been received, but the gallant and honorable spirit which has at all times been the pride and glory of France will not ultimately permit the demands of innocent sufferers to be extinguished in the mere consciousness of the power to reject them.	We are not only examining major problems facing the various modes of transport; we are also studying closely the inter-relationships of civilian and government requirements for transportation.

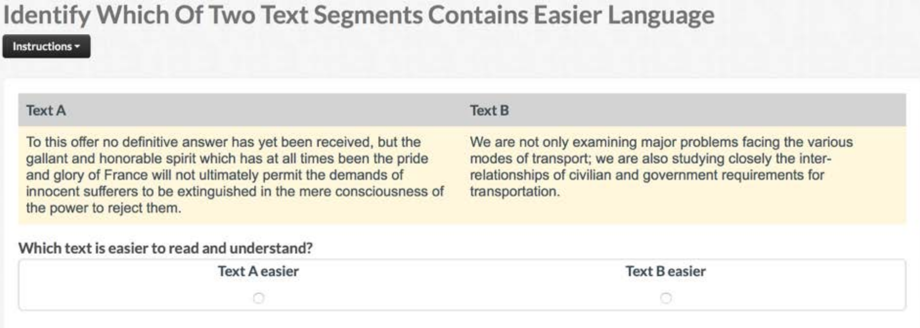
Which text is easier to read and understand?

Text A easier	Text B easier
○	○

Why *pairwise* comparisons?

We had to make sure the snippets we very similar length: why?

Benoit et al. example



Identify Which Of Two Text Segments Contains Easier Language

Instructions

Text A	Text B
To this offer no definitive answer has yet been received, but the gallant and honorable spirit which has at all times been the pride and glory of France will not ultimately permit the demands of innocent sufferers to be extinguished in the mere consciousness of the power to reject them.	We are not only examining major problems facing the various modes of transport; we are also studying closely the inter-relationships of civilian and government requirements for transportation.

Which text is easier to read and understand?

Text A easier

Text B easier

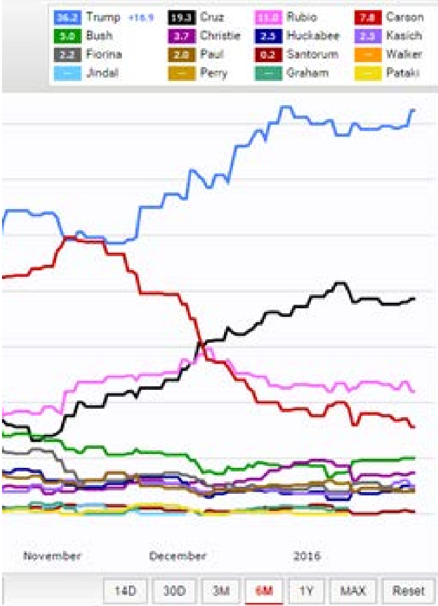
Why *pairwise* comparisons?

We had to make sure the snippets we very similar length: why?

Some universities have banned crowd-sourcing as unethical: why?

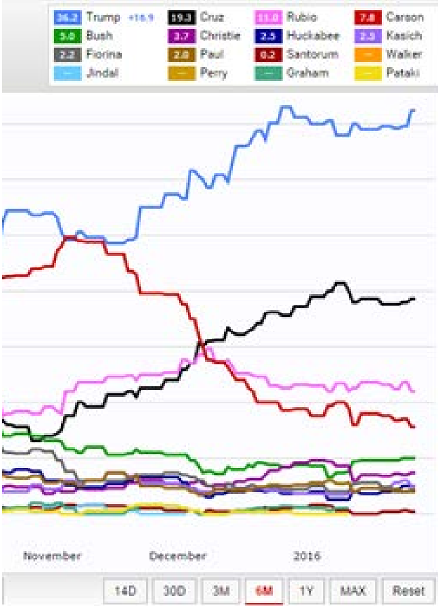
Exercise

Exercise



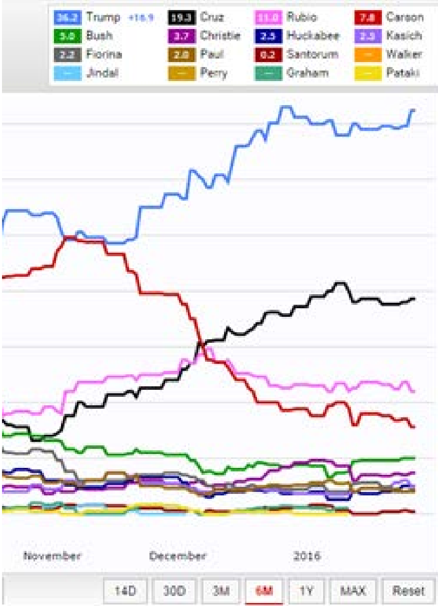
You work for a polling company and have access to a crowdsourcing service,

Exercise



You work for a polling company and have access to a crowdsourcing service, and want to know who will win the US Presidential election.

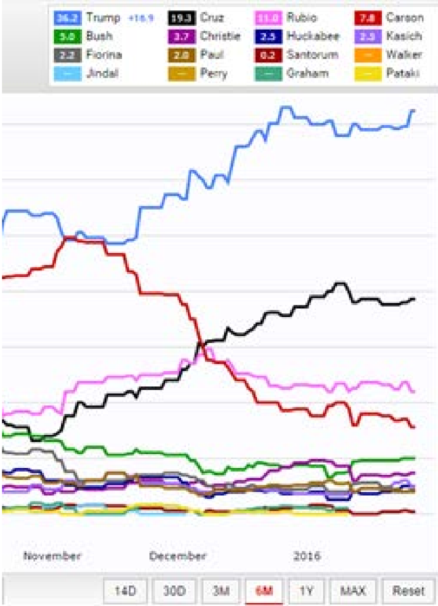
Exercise



You work for a polling company and have access to a crowdsourcing service, and want to know who will win the US Presidential election.

1 Suppose the question you *have* to implement is 'Which of these candidates do you prefer?' Can we crowdsource this? What are the threats to inference?

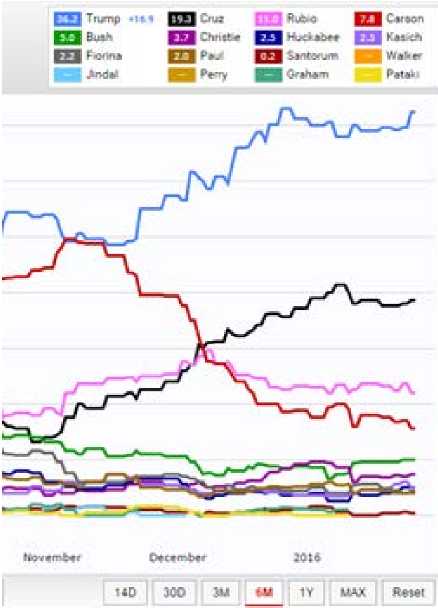
Exercise



You work for a polling company and have access to a crowdsourcing service, and want to know who will win the US Presidential election.

1. Suppose the question you *have* to implement is 'Which of these candidates do you prefer?' Can we crowdsource this? What are the threats to inference?
2. Given the Galton/'Wisdom of Crowds' idea, what would be a better question?

Exercise



You work for a polling company and have access to a crowdsourcing service, and want to know who will win the US Presidential election.

1. Suppose the question you *have* to implement is 'Which of these candidates do you prefer?' Can we crowdsource this? What are the threats to inference?
2. Given the Galton/'Wisdom of Crowds' idea, what would be a better question?

When Prediction Markets Fail

When Prediction Markets Fail

Missed Brexit and Trump. Why? (Snowberg, Wolfers & Zitzewitz)

When Prediction Markets Fail

Missed Brexit and Trump. Why? (Snowberg, Wolfers & Zitzewitz)

Needs to be actual information to be aggregated. “Does Iraq have WMDs?” would not have worked.

When Prediction Markets Fail

Missed Brexit and Trump. Why? (Snowberg, Wolfers & Zitzewitz)

Needs to be actual information to be aggregated. "Does Iraq have WMDs?" would not have worked.

Markets can also be too thin, for various reasons: e.g. insider trading drives off bettors, or question is too ambiguously posed.

When Prediction Markets Fail

Missed Brexit and Trump. Why? (Snowberg, Wolfers & Zitzewitz)

Needs to be actual information to be aggregated. “Does Iraq have WMDs?” would not have worked.

Markets can also be too thin, for various reasons: e.g. insider trading drives off bettors, or question is too ambiguously posed.

Behavioral biases: people like betting long-shots. People may also not be fully rational wrt bets (e.g. prices for England winning in England).

When Prediction Markets Fail

Missed Brexit and Trump. Why? (Snowberg, Wolfers & Zitzewitz)

Needs to be actual information to be aggregated. "Does Iraq have WMDs?" would not have worked.

Markets can also be too thin, for various reasons: e.g. insider trading drives off bettors, or question is too ambiguously posed.

Behavioral biases: people like betting long-shots. People may also not be fully rational wrt bets (e.g. prices for England winning in England).

Q What are analogies to these in crowdsourcing tasks in text coding?

Extra Material

Estimation Notes I

Estimation Notes I

e.g. If there are $K = 3$ stems of interest,

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles:

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large,

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable:

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable: try having y of length 2^{300}.

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0],[0,0,1],[0,1,0],[1,0,0],[1,1,0],[1,0,1],[0,1,1],[1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable: try having y of length 2^{300}. Plus, number of possible stem profiles (y) is much larger than number of observations,

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0],[0,0,1],[0,1,0],[1,0,0],[1,1,0],[1,0,1],[0,1,1],[1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable: try having y of length 2^{300}. Plus, number of possible stem profiles (y) is much larger than number of observations, meaning that many of the profile combinations are never observed (we have no information about them).

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0],[0,0,1],[0,1,0],[1,0,0],[1,1,0],[1,0,1],[0,1,1],[1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable: try having y of length 2^{300}. Plus, number of possible stem profiles (y) is much larger than number of observations, meaning that many of the profile combinations are never observed (we have no information about them).

Estimation Notes II

so choose subset of 5–25 stems and estimate Pr(c),

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation),

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

$\rightsquigarrow$ kernel smoothing of sparse matrices.

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

$\rightsquigarrow$ kernel smoothing of sparse matrices.

Judge *relative* performance via mean absolute proportion error.

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

$\rightsquigarrow$ kernel smoothing of sparse matrices.

Judge *relative* performance via mean absolute proportion error.

NB "*among all documents in a given category, the prevalence of particular word profiles in the labeled set should be the same in expectation as in the population set*".

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

$\rightsquigarrow$ kernel smoothing of sparse matrices.

Judge *relative* performance via mean absolute proportion error.

NB "*among all documents in a given category, the prevalence of particular word profiles in the labeled set should be the same in expectation as in the population set*". This is key assumption. btw, what happened to the danger of drift?!