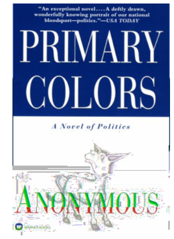
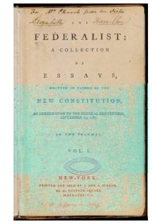
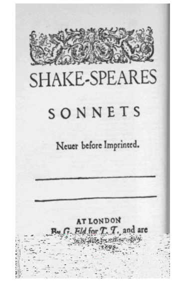
Statistical Approaches to Authorship Attribution

David Madigan
Rutgers

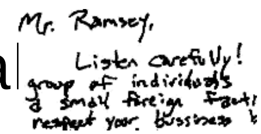
Some Background

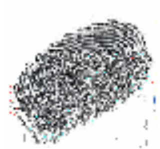




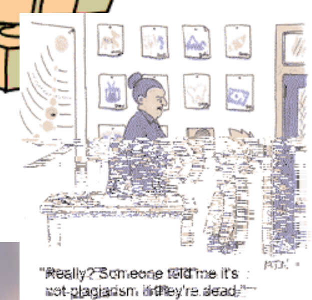


- Identification technologies important for homeland security and in the legal system





- Authorship attribution for textual artifacts using "topic independent" stylometric features has a long history



- Historical focus on small numbers of authors and low-dimensional representations via function words

Statistical Text Analysis

SCIENCE.

FRIDAY, MARCH 11, 1887.

THE CHARACTERISTIC CURVES OF COMPOSITION.

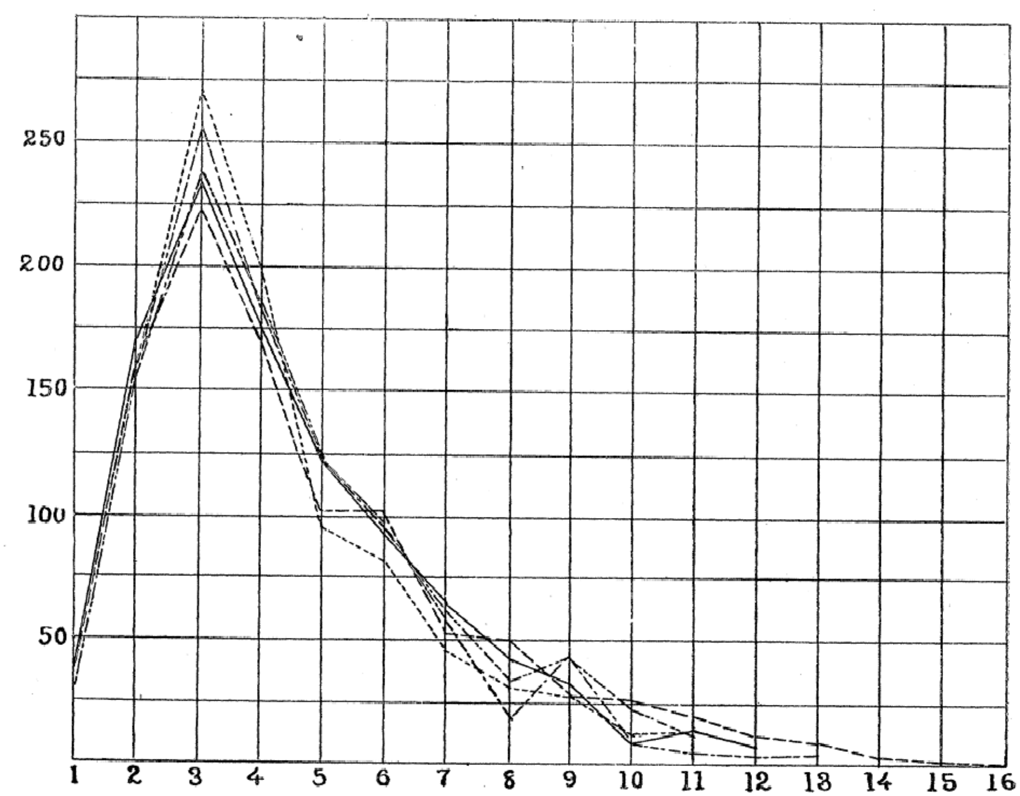


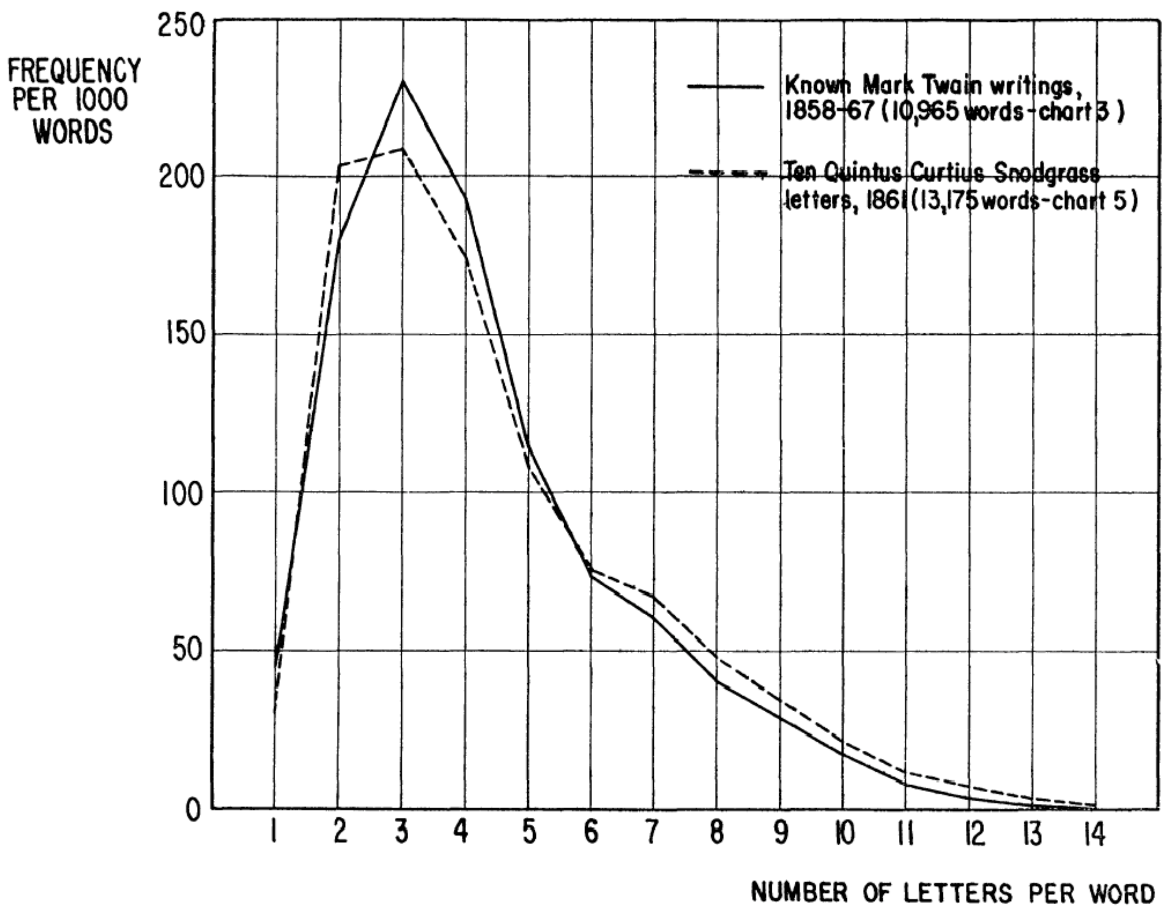
FIG. 2. — SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'

Mendenhall

AMERICAN STATISTICAL ASSOCIATION JOURNAL, MARCH 1963

MARK TWAIN AND THE QUINTUS CURTIUS SNODGRASS LETTERS: A STATISTICAL TEST OF AUTHORSHIP

CLAUDE S. BRINEGAR

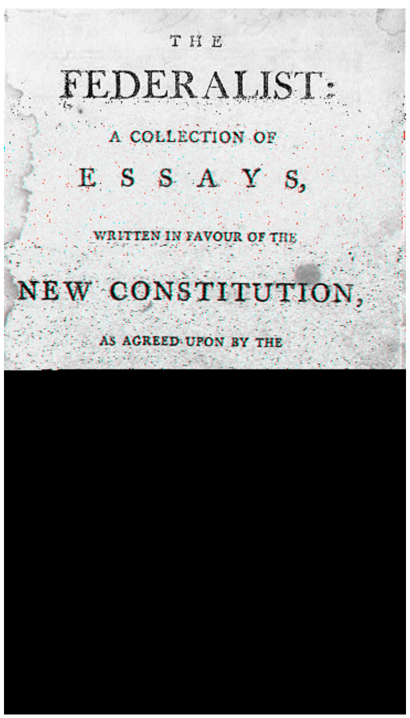


X

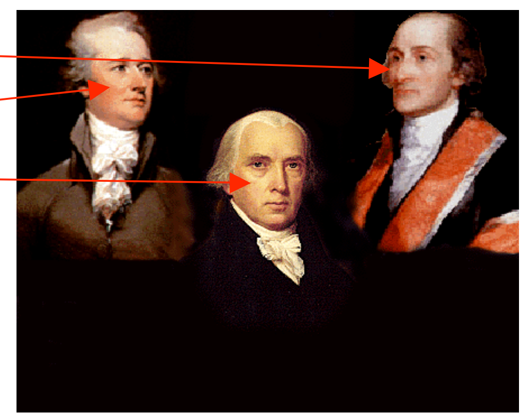
The Federalist

- “The authorship of certain numbers of the ‘Federalist’ has fairly reached the dignity of a well-established historical controversy.” (Henry Cabot Lodge, 1886)

- Historical evidence is muddled


THE
FEDERALIST:
A COLLECTION OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE

Paper Number	Author
1	Hamilton
2-5	Jay
6-9	Hamilton
10	Madison
11-13	Hamilton
14	Madison
15-17	Hamilton
18-20	Joint: Hamilton and Madison
21-36	Hamilton
37-48	Madison
49-58	**Disputed**
59-61	Hamilton
62-63	**Disputed**
64	Jay
65-85	Hamilton



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 Volume 58

ENCE IN AN AUTHORSHIP PROBLEM[1,2]

rative study of discrimination methods applied authorship of the disputed *Federalist* papers

INFERI

A compa to the

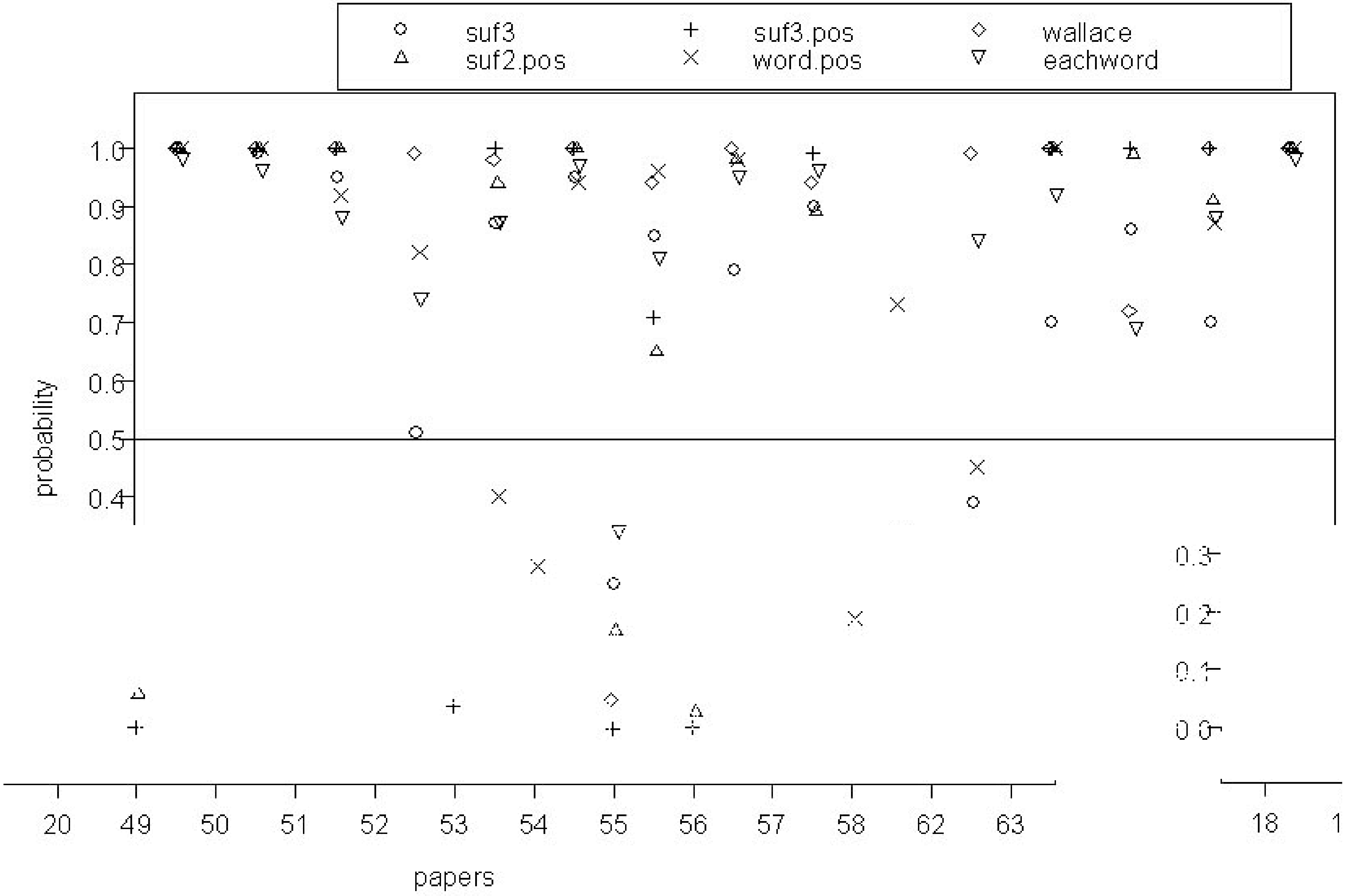
FREDERICK MOSTELLER

- Used function words with Naïve Bayes with Poisson and Negative Binomial model
- Out-of-sample predictive performance

F. Summing up

In summary, the following points are clear:

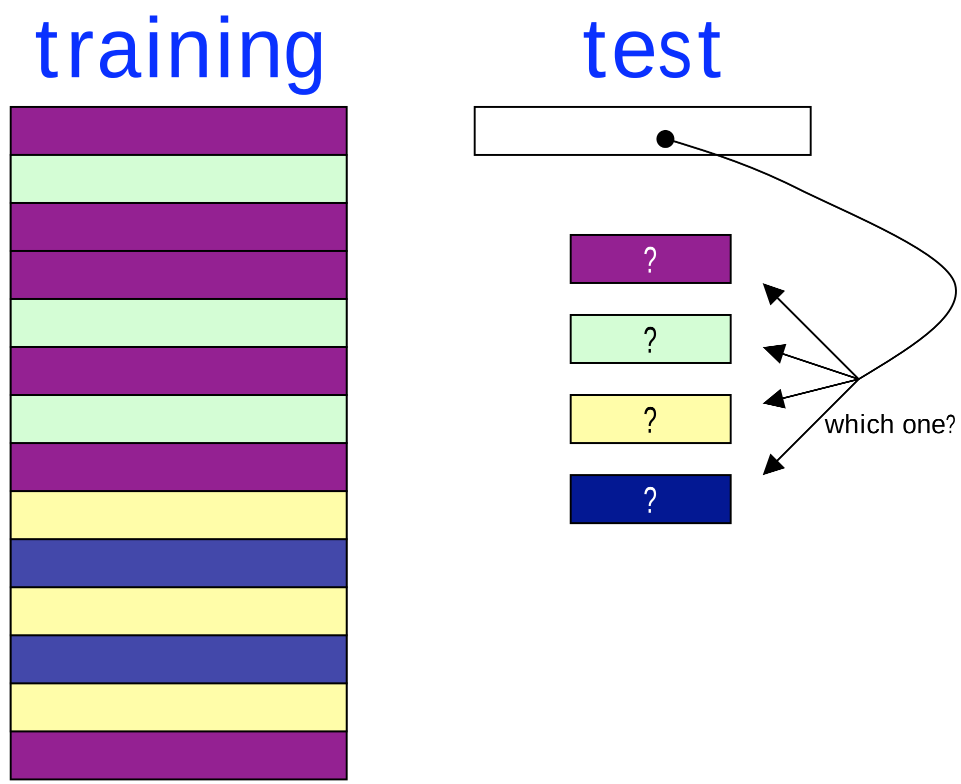
1) Madison is the principal author. These data make it possible to say far more than ever before that the odds are enormously high that Madison wrote the 12 disputed papers. Weakest support is given for No. 55. Support for Nos. 62 and 63, most in doubt by current historians, is tremendous.



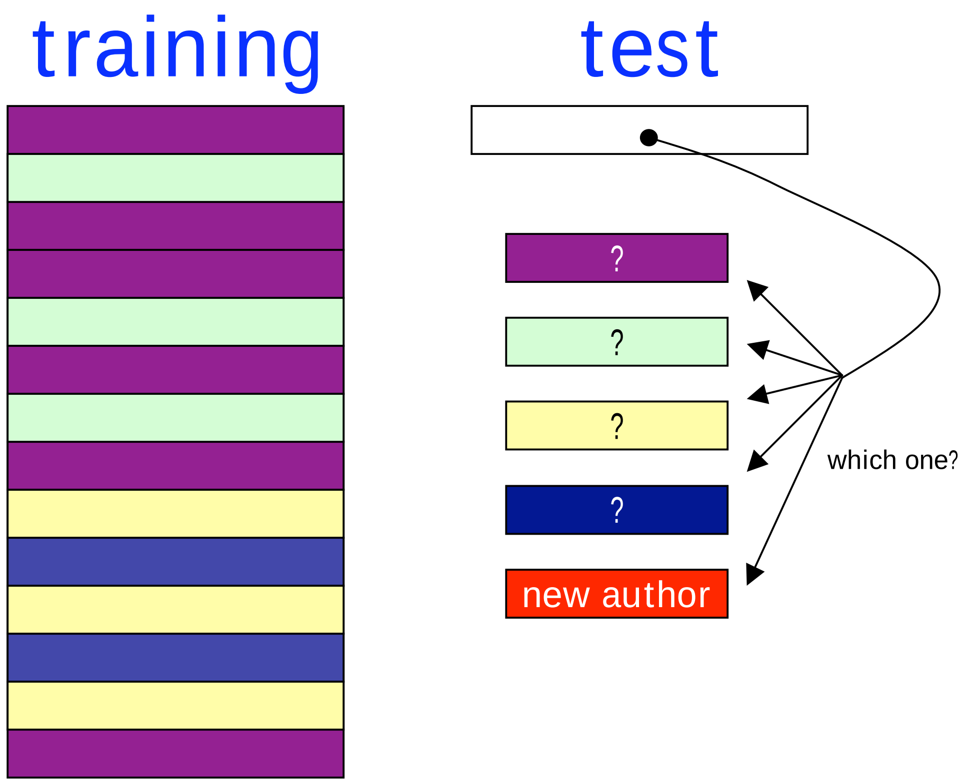
suf3
suf3.pos
wallace
suf2.pos
word.pos
eachword
probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
20
49
50
51
52
53
54
55
56
57
58
62
63
18
papers

Different Attribution Problems

1 of K

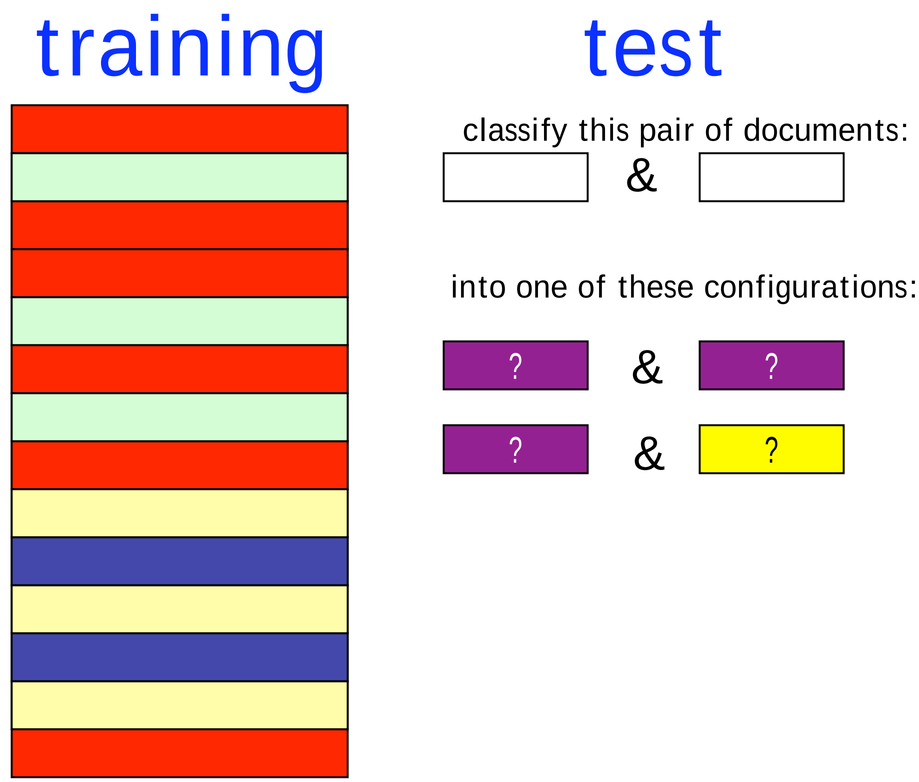


“odd man out”

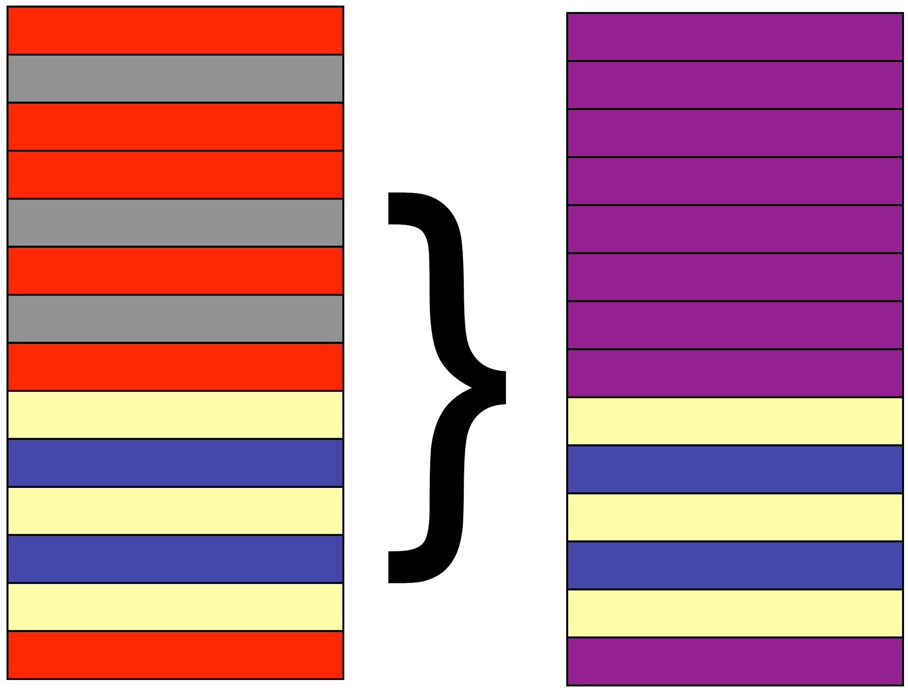


aka “novelty detection”

document pairs



anti-aliasing



in this example, the red author and the grey author are the same real person

Other Related Problems

- Author gender
- Author nationality
- Sentiment (positive/negative feeling)
- Rhetorical style
- Multi-authored documents

1-of-K Sample Results: brittany-l

Feature Set	% errors	Number of Features
“Argamon” function words, raw tf	74.8	380
POS	75.1	44
1suff	64.2	121
1suff*POS	50.9	554
2suff	40.6	1849
2suff*POS	34.9	3655
3suff	28.7	8676
3suff*POS	27.9	12976
3suff+POS+3suff*POS+Argamon	27.6	22057
All words	23.9	52492

4.6 million parameters

89 authors with at least 50 postings. 10,076 training documents, 3,322 test documents.

BMR-Laplace classification, default hyperparameter

Predictive Modeling

$$y \quad f \boldsymbol{x} \theta$$

$y \in \quad c_1 \qquad c_m$

y

$c_1 \qquad c_m$

Brief Aside to Study Probability…

Discrete Random Variables

- A is a Boolean-valued random variable if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
- Examples
- A = The US president in 2023 will be male
- A = You wake up tomorrow with a headache
- A = You have Ebola

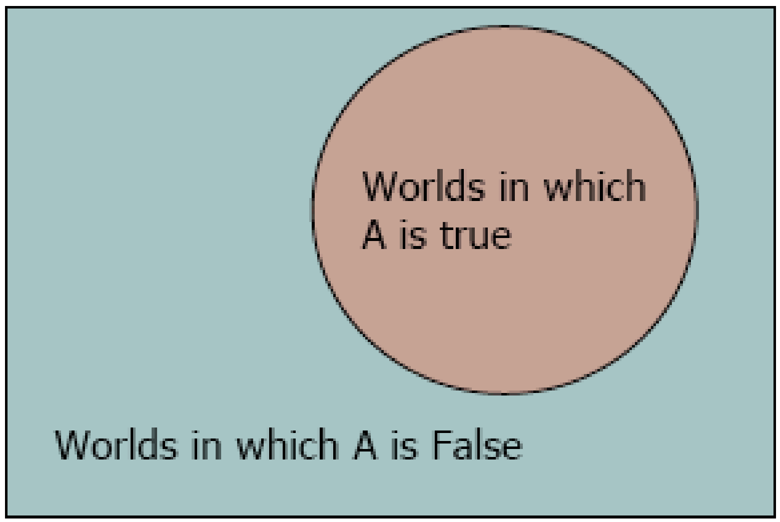
Probabilities

- We write P(A) as "the fraction of possible worlds in which A is true"
- We could at this point spend 2 hours on the philosophy of this.
- But we won't.

Visualizing A

Event space of all possible worlds

Its area is 1



P(A) = Area of reddish oval

The Axioms of Probability

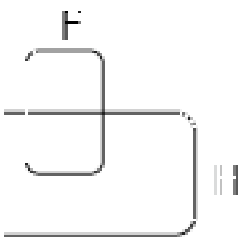
- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

Conditional Probability

- P(A|B) = Fraction of worlds in which B is true that also have A true

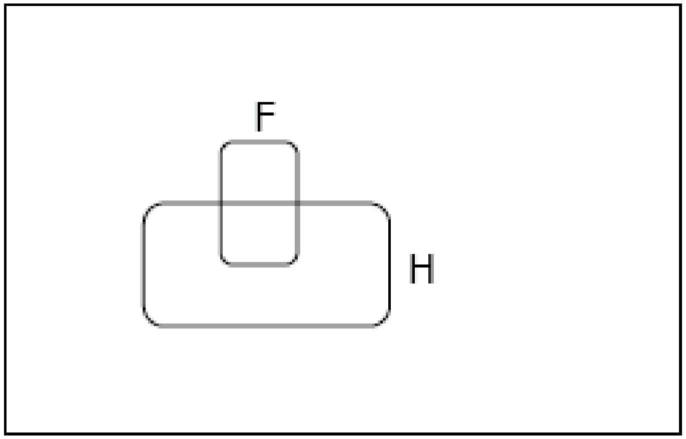
H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2



"Headaches are rare and flu is rarer, but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

Conditional Probability



P(H|F) = Fraction of flu-inflicted worlds in which you have a headache

$$= \frac{\text{\#worlds with flu and headache}}{\text{\#worlds with flu}}$$

$$= \frac{\text{Area of “H and F” region}}{\text{Area of “F” region}}$$

$$= \frac{P(H \wedge F)}{P(F)}$$

H = “Have a headache”
F = “Coming down with Flu”

$P(H) = 1/10$
$P(F) = 1/40$
$P(H|F) = 1/2$

Back to the Main Story…

Probabilistic Classification

$p\ c_k$ c_k

c_k $p\ \mathbf{x}\ \ c_k$

$p\ c_k\ \ \mathbf{x}\quad p\ \mathbf{x}\ \ c_k\ \ p\ c_k\ \ p\ \mathbf{x}$

$$p\ \mathbf{A}\ \ \mathbf{B}\quad p\ \mathbf{A}\ \mathit{and}\ \mathbf{B}\ /\ p\ \mathbf{B}$$

$$p\ \mathbf{B}\ \ \mathbf{A}\quad p\ \mathbf{A}\ \mathit{and}\ \mathbf{B}\ /\ p\ \mathbf{A}$$

$$p\ \mathbf{A}\ \ \mathbf{B}\quad \frac{p\ \mathbf{B}\ \ \mathbf{A}\times p\ \mathbf{A}}{p\ \mathbf{B}}$$

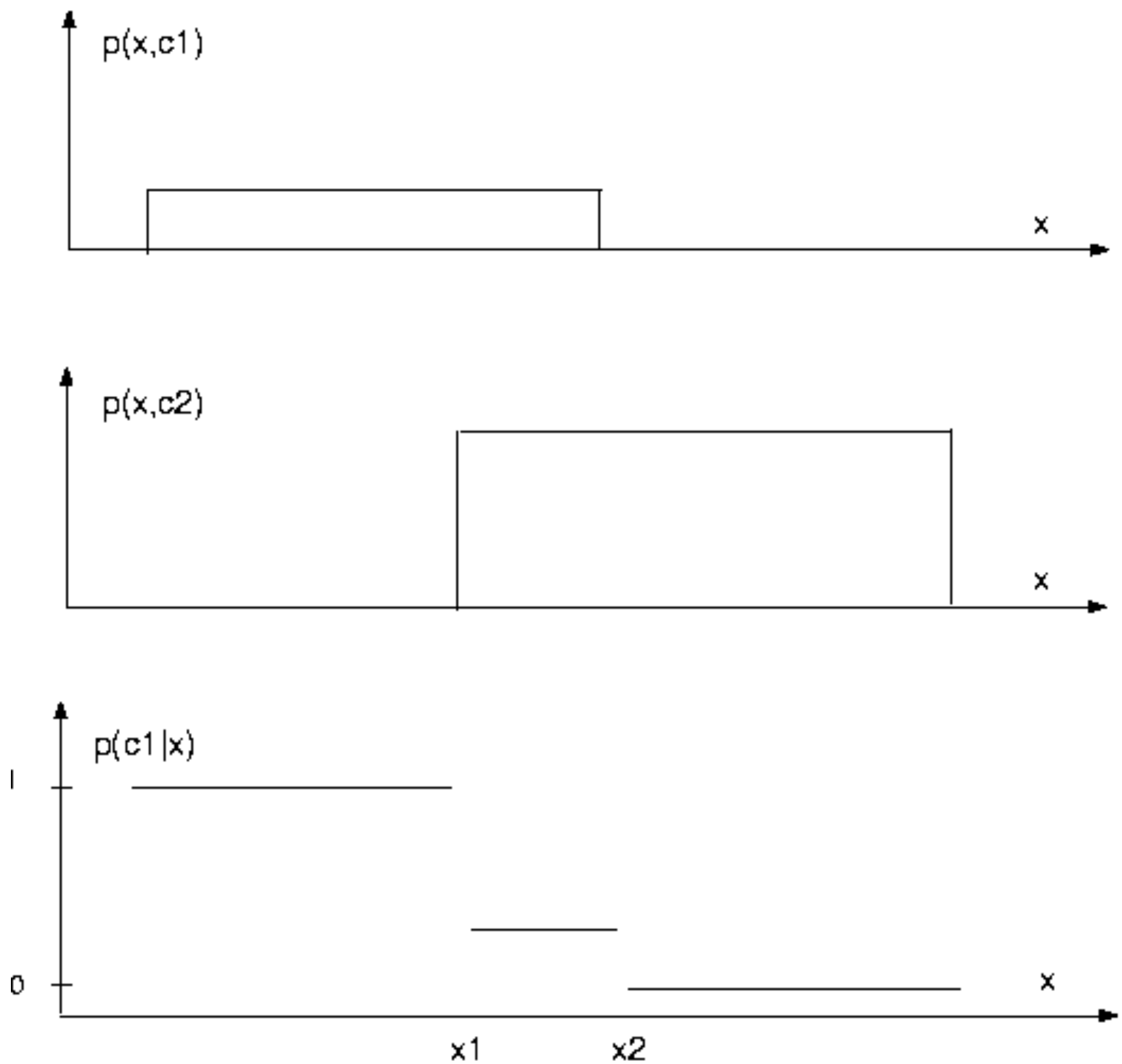
Stochastic Independence

$$p\ \mathbf{A}\quad \mathbf{B}\qquad p\ \mathbf{A}$$

$$p\ \mathbf{B}\quad \mathbf{A}\qquad p\ \mathbf{B}$$

$$=> p\ \mathbf{A}\quad \mathbf{B}\qquad p\ \mathbf{A}\qquad p\ \mathbf{A}\ \boldsymbol{and}\ \mathbf{B}\ \ /\ p\ \mathbf{B}$$

$$p\ \mathbf{A}\ \boldsymbol{and}\ \mathbf{B}\ = p\ \mathbf{A}\qquad p\ \mathbf{B}$$



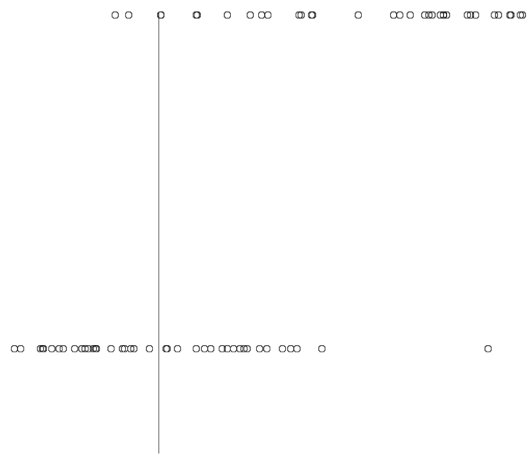
p(x,c1)
x
p(x,c2)
x
p(c1|x)
1
0
x
x1
x2

Classifier Types

Discriminative $p \ c_k \quad \boldsymbol{x}$

Generative $p \ \boldsymbol{x} \quad c_k$

Regression for Binary Classification



Naïve Bayes via a Toy Spam Filter Example

#	Message	Spam
1	the quick brown fox	no
2	the quick rabbit ran and ran	yes
3	rabbit run run run	no
4	rabbit at rest	yes

Training data comprising four labeled e-mail messages.

#	and	at	brown	fox	quick	rabbit	ran	rest	run	the
1	0	0	1	1	1	0	0	0	0	1
2	1	0	0	0	1	1	2	0	0	1
3	0	0	0	0	0	1	0	0	3	0
4	0	1	0	0	0	1	0	1	0	0

Term vectors corresponding to the training data.

	X_1	X_2	X_3	X_4	X_5	X_6	Y
#	**brown**	**fox**	**quick**	**rabbit**	**rest**	**run**	**Spam**
1	1	1	1	0	0	0	0
2	0	0	1	1	0	2	1
3	0	0	0	1	0	3	0
4	0	0	1	1	1	0	1

Term vectors after stemming and stopword removal with the Spam label, coded as 0=no, 1=yes.

Naïve Bayes Machinery

$$Pr(Y=1|X_1=x_1,\ldots,X_d=x_d)$$

$$=\frac{Pr(Y=1)\times Pr(X_1=x_1,\ldots,X_d=x_d|Y=1)}{Pr(X_1=x_1,\ldots,X_d=x_d)}$$

$$\log\frac{Pr(Y=1|X_1=x_1,\ldots,X_d=x_d)}{Pr(Y=0|X_1=x_1,\ldots,X_d=x_d)}$$

$$=\log\frac{Pr(Y=1)}{Pr(Y=0)}+\log\frac{Pr(X_1=x_1,\ldots,X_d=x_d|Y=1)}{Pr(X_1=x_1,\ldots,X_d=x_d|Y=0)}$$

Naïve Bayes Machinery

$$Pr(X_1 = x_1, \ldots, X_d = x_d | Y = 1) = \prod_{i=1}^{d} Pr(X_i = x_i | Y = 1)$$

$$Pr(X_1 = x_1, \ldots, X_d = x_d | Y = 0) = \prod_{i=1}^{d} Pr(X_i = x_i | Y = 0)$$

$$\log \frac{Pr(Y = 1 | X_1 = x_1, \ldots, X_d = x_d)}{Pr(Y = 0 | X_1 = x_1, \ldots, X_d = x_d)}$$

$$= \log \frac{Pr(Y = 1)}{Pr(Y = 0)} + \sum_{i=1}^{d} \log \frac{Pr(X_i = x_i | Y = 1)}{Pr(X_i = x_i | Y = 0)}$$

Maximum Likelihood Estimation

$$\log \frac{\widehat{Pr}(Y=1)}{\widehat{Pr}(Y=0)} = \log \frac{2/4}{2/4} = 0$$

weights of evidence

$$\log \frac{\widehat{Pr}(X_3=1|Y=1)}{\widehat{Pr}(X_3=1|Y=0)} = \log \frac{2/2}{1/2} = \log 2$$

	X_1	X_2	X_3	X_4	X_5	X_6	Y
#	brown	fox	quick	rabbit	rest	run	Spam
1	1	1	1	0	0	0	0
2	0	0	1	1	0	2	1
3	0	0	0	1	0	3	0
4	0	0	1	1	1	0	1

Naïve Bayes Prediction

	X_1	X_2	X_3	X_4	X_5	X_6
	brown	**fox**	**quick**	**rabbit**	**rest**	**run**
Term Present	-1.10	-1.10	0.51	0.51	1.10	0
Term Absent	0.51	0.51	-1.10	-1.10	-0.51	0

Estimated Weights of evidence for the example.

	X_1	X_2	X_3	X_4	X_5	X_6
	brown	**fox**	**quick**	**rabbit**	**rest**	**run**
Term Vector	0	0	1	1	1	0
Weight of Evidence	0.51	0.51	0.51	0.51	1.10	0

Your turn!

Essay #	“upon”	“whilst”	Author
1	yes	yes	Hamilton
2	yes	no	Hamilton
3	yes	no	Hamilton
4	no	yes	Madison
5	yes	no	Madison

6	yes	no	???