Bayesian Spam Detection

Jeremy Eberhardt

University of Minnesota, Morris

eberh060@morris.umn.edu

December 6, 2014

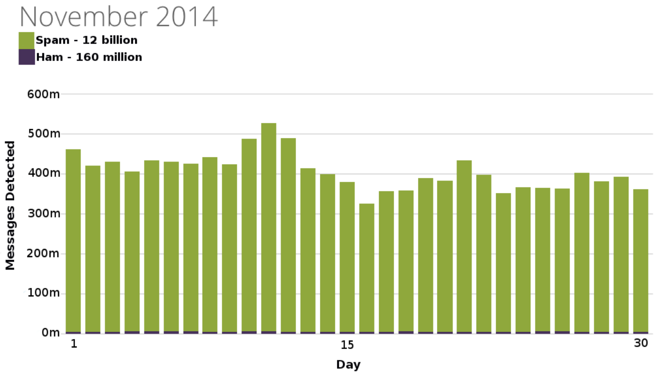
What is it?

- Spam
  - Anything that is undesired by the user
  - Email spam
  - Comment spam
- Ham
  - Non-spam
- Bayesian Approach
  - Statistics based document classification

"Wow! I've got one from someone I know!"

Why do We Care?

- 70-90% of all emails are spam
- Global issue
- Security
  - Advertising
  - Scams
  - Identity theft
- Quality of life



Overview

1. Intro
   - Setup
2. Naive Bayes
   - Explanation and example
3. Multinomial Bayes
   - Explanation and example
4. Multivariate Bayes
   - Explanation and example
5. Feature Selection
6. Tests
   - Multinomial Bayes
   - Multivariate Bayes
7. Conclusion

Setup

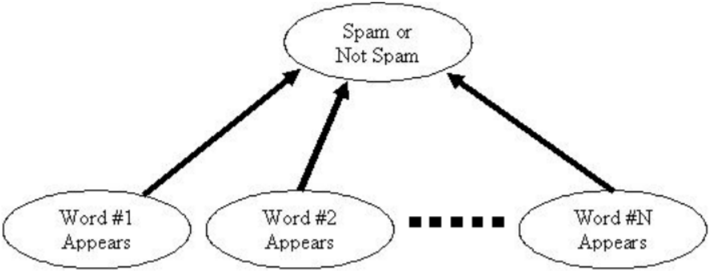
- Training data
  - Prepare the filter before use
  - Pre-classified documents that the user specifies
- *Prior* probability
  - Probability that an event occurs
- *Conditional* probability
  - Probability of an event given that another event has occurred

Setup

- Training data
  - Prepare the filter before use
  - Pre-classified documents that the user specifies
- *Prior* probability
  - Probability that an event occurs
- *Conditional* probability
  - Probability of an event given that another event has occurred

$$P(snowing) \text{ VS } P(snowing|summer)$$

Naive Bayes Classifier



Naive Bayes Classifier

$$P(S|W) = Spamicity(W)$$

The probability that a document is spam given that word W occurs in the document.

$$\frac{Count(S,W) \cdot P(S)}{Count(S,W) \cdot P(S) + Count(H,W) \cdot P(H)}$$

Classify the Document

$$P(S|All\ \ words) = \frac{Spamicity(All\ \ words)}{Spamicity(All\ \ words)+Hamicity(All\ \ words)}$$

Compare to threshold or

$$P(H|All\ \ words) = \frac{Hamicity(All\ \ words)}{Hamicity(All\ \ words)+Spamicity(All\ \ words)}$$

Naive Bayes Example

Spamicity

$$\frac{Count(S,W) \cdot P(S)}{Count(S,W) \cdot P(S) + Count(H,W) \cdot P(H)}$$

Classification

$$\frac{Spamicity(All\ \ words)}{Spamicity(All\ \ words) + Hamicity(All\ \ words)}$$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	*?*

Naive Bayes Example

Spamicity

$$\frac{Count(S,W) \cdot P(S)}{Count(S,W) \cdot P(S) + Count(H,W) \cdot P(H)}$$

Classification

$$\frac{Spamicity(All\ \ words)}{Spamicity(All\ \ words) + Hamicity(All\ \ words)}$$

S(Purple): $\frac{1 \cdot 2/3}{(1 \cdot 2/3)+(1 \cdot 1/3)} = 2/3$

S(Circle) $= 4/5$

H(Purple): $\frac{1 \cdot 1/3}{(1 \cdot 1/3)+(1 \cdot 2/3)} = 1/3$

H(Circle) $= 1/5$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

Naive Bayes Example

Spamicity

$$\frac{Count(S,W) \cdot P(S)}{Count(S,W) \cdot P(S) + Count(H,W) \cdot P(H)}$$

Classification

$$\frac{Spamicity(All\ \ words)}{Spamicity(All\ \ words) + Hamicity(All\ \ words)}$$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

S(Purple): $\frac{1 \cdot 2/3}{(1 \cdot 2/3) + (1 \cdot 1/3)} = 2/3$

S(Circle) $= 4/5$

H(Purple): $\frac{1 \cdot 1/3}{(1 \cdot 1/3) + (1 \cdot 2/3)} = 1/3$

H(Circle) $= 1/5$

$$P(S) = \frac{0.667 * 0.8}{(0.667 * 0.8) + (0.333 * 0.2)} \approx 0.89$$

$$P(H) = \frac{0.333 * 0.2}{(0.333 * 0.2) + (0.667 * 0.8)} \approx 0.11$$

Naive Bayes Example

Spamicity

$$\frac{Count(S,W)\cdot P(S)}{Count(S,W)\cdot P(S)+Count(H,W)\cdot P(H)}$$

Classification

$$\frac{Spamicity(All\ \ words)}{Spamicity(All\ \ words)+Hamicity(All\ \ words)}$$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

S(Purple): $\frac{1\cdot 2/3}{(1\cdot 2/3)+(1\cdot 1/3)} = 2/3$

S(Circle) $= 4/5$

H(Purple): $\frac{1\cdot 1/3}{(1\cdot 1/3)+(1\cdot 2/3)} = 1/3$

H(Circle) $= 1/5$

$$P(S) = \frac{0.667*0.8}{(0.667*0.8)+(0.333*0.2)} \approx 0.89$$

$$P(H) = \frac{0.333*0.2}{(0.333*0.2)+(0.667*0.8)} \approx 0.11$$

Spam

Multinomial Bayes

- Optimization of Naive Bayes classifier
- Multinomial distribution of words
- Words are independent
- Instead of counting documents, count words
- Instead of calculating $P(S|W)$ calculate $P(W|S)$

$$P(S|All\ \ words) = \underset{\substack{\uparrow \\ Prior}}{P(S)} \cdot \underset{\substack{\uparrow \\ Conditional}}{P(W_1|S)^{f_1}} \cdot \ldots \cdot P(W_n|S)^{f_n}$$

Multinomial Bayes Example

Priors

$P(H) = \frac{1}{3} \quad P(S) = \frac{2}{3}$

Conditional

$P(W|S) = \frac{Count(W,S)+1}{Count(S)+Vocabulary}$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	*?*

Multinomial Bayes Example

Priors

$P(H) = \frac{1}{3} \quad P(S) = \frac{2}{3}$

Conditional

$P(W|S) = \frac{Count(W,S)+1}{Count(S)+Vocabulary}$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

$P(Purple|S) = (1+1)/(6+5) = 2/11$

$P(Circle|S) = 2/11$

$P(Purple|H) = (2+1)/(4+5) = 3/9$

$P(Circle|H) = 2/9$

Multinomial Bayes Example

Priors

$P(H) = \frac{1}{3} \quad P(S) = \frac{2}{3}$

Conditional

$P(W|S) = \frac{Count(W,S)+1}{Count(S)+Vocabulary}$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

$P(Purple|S) = (1+1)/(6+5) = 2/11$

$P(Circle|S) = 2/11$

$P(Purple|H) = (2+1)/(4+5) = 3/9$

$P(Circle|H) = 2/9$

$P(S) = 2/3 * (2/11)^2 * 2/11$
≈ 0.004

$P(H) = 1/3 * (3/9)^2 * 2/9$
≈ 0.008

Multinomial Bayes Example

Priors

$P(H) = \frac{1}{3} \quad P(S) = \frac{2}{3}$

Conditional

$P(W|S) = \frac{Count(W,S)+1}{Count(S)+Vocabulary}$

Doc	**Content**	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

$P(Purple|S) = (1+1)/(6+5) = 2/11$

$P(Circle|S) = 2/11$

$P(Purple|H) = (2+1)/(4+5) = 3/9$

$P(Circle|H) = 2/9$

$P(S) = 2/3 * (2/11)^2 * 2/11$
≈ 0.004

$P(H) = 1/3 * (3/9)^2 * 2/9$
≈ 0.008

Ham

Multivariate Bayes

- Another optimization of Naive Bayes
- Similar to Multinomial Bayes, simpler
- Combines ideas from Naive Bayes and Multinomial Bayes
- Calculate probabilities like Multinomial Bayes
- Counts documents like Naive Bayes

Multivariate Bayes

$$P(S|All\ words) = P(S) \cdot P(W_1|S)^{f_1} \cdot \ldots \cdot P(W_n|S)^{f_n}$$

Multivariate Bayes

$$P(S|All\ \ words) = \underset{\substack{\uparrow \\ Prior}}{P(S)} \cdot P(W_1\underset{\substack{\uparrow \\ \textbf{Conditional}}}{|}S)^{f_1} \cdot \ldots \cdot P(W_n|S)^{\overset{\mathbf{1}}{\downarrow}\, f_n}$$

Multivariate Bayes Example

Priors

$P(H) = \frac{1}{3} \quad P(S) = \frac{2}{3}$

Conditional

$P(W|S) = \frac{1+Count(S,W)}{2+Count(S)}$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	*?*

Multivariate Bayes Example

Priors

$P(H) = \frac{1}{3} \quad P(S) = \frac{2}{3}$

Conditional

$P(W|S) = \frac{1+Count(S,W)}{2+Count(S)}$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

$P(Purple|S) = (1+1)/(2+2) = 1/2$

$P(Circle|S) = 1/2$

$P(Purple|H) = (1+1)/(2+1) = 2/3$

$P(Circle|H) = 2/3$

Multivariate Bayes Example

Priors

$P(H) = \frac{1}{3} \quad P(S) = \frac{2}{3}$

Conditional

$P(W|S) = \frac{1+Count(S,W)}{2+Count(S)}$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

$P(Purple|S) = (1+1)/(2+2) = 1/2$

$P(Circle|S) = 1/2$

$P(S) = 2/3 * 1/2 * 1/2 \approx 0.166$

$P(Purple|H) = (1+1)/(2+1) = 2/3$

$P(Circle|H) = 2/3$

$P(H) = 1/3 * 2/3 * 2/3 \approx 0.148$

Multivariate Bayes Example

Priors

$P(H) = \frac{1}{3} \quad P(S) = \frac{2}{3}$

Conditional

$P(W|S) = \frac{1+Count(S,W)}{2+Count(S)}$

Doc	Content	
1	Purple Black Purple Circle	H
2	Circle Square Square Red	S
3	Square Purple	S
	Purple Purple Circle	?

$P(Purple|S) = (1+1)/(2+2) = 1/2$

$P(Circle|S) = 1/2$

$P(Purple|H) = (1+1)/(2+1) = 2/3$

$P(Circle|H) = 2/3$

$P(S) = 2/3 * 1/2 * 1/2 \approx 0.166$

$P(H) = 1/3 * 2/3 * 2/3 \approx 0.148$

Spam

Feature Selection

- Features:
  - Words
  - Lengths of words
  - Letters
  - Images
  - *N-grams of words*

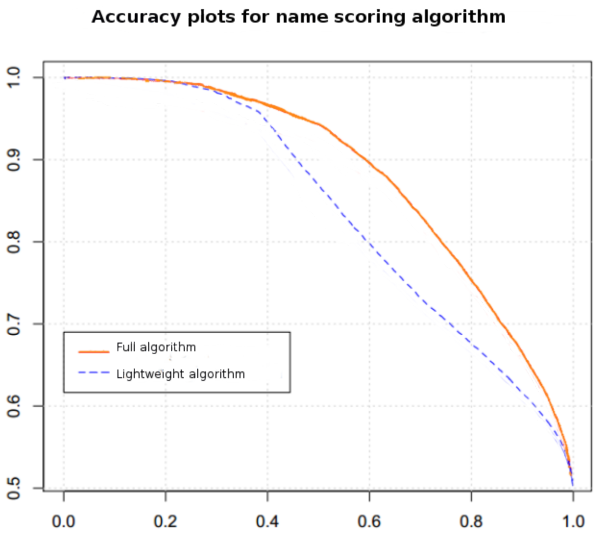
3-gram of “david”

(_da, dav, avi, vid, id_)

Multinomial Bayes Testing

- Freeman 2013
- LinkedIn account names
- 60 million accounts
  - 100,000 were chosen to be tested, 50,000 spam and ham
- N-gram values 3(Lightweight) and 5(Full)
- 110 MB vs 974 MB

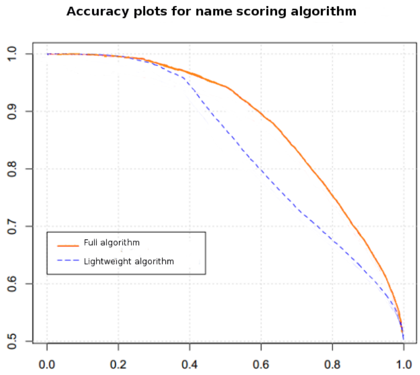
Results



Accuracy plots for name scoring algorithm
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Full algorithm
Lightweight algorithm

Results

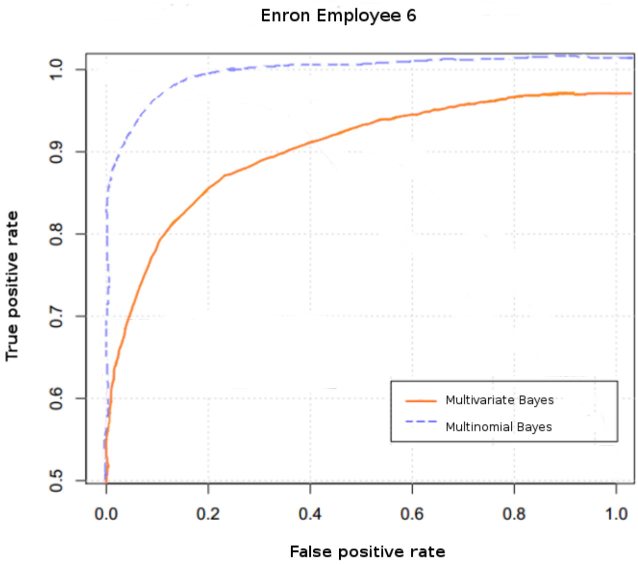
- Larger data sets $\Rightarrow$ Lightweight algorithm
  - Memory tradeoffs become more relevant
  - More reliable for more documents
- Both more effective than previous algorithm
  - Based on regular expressions
- Chose Full algorithm
- Cut false positive rate in half



Multivariate Bayes

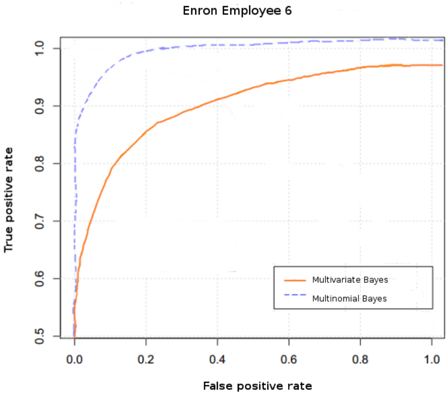
- Athens University of Economics and Business
- Data collected from Enron employees
  - Subject line and body
  - Ham only
- Mixed in unique generic spam emails
- Emulate real-time spam filtering
- Ordered the emails chronologically (complicated)
- 43,000 ham, 50,000 spam
- Clustered emails into chunks of 100
- Filter updated after each chunk

Results


Enron Employee 6
True positive rate
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Multivariate Bayes
Multinomial Bayes

Results

- Multivariate Bayes performed relatively poorly
- Multivariate Bayes still moderately effective
- Less effective than Multinomial Bayes
- Multinomial Bayes performed best in all cases



Advantages and Disadvantages of Bayesian Spam Filtering

Advantages

- Adjustable accuracy
- Different models for different needs
- User control
- Constantly adapts

Disadvantages

- Training data
- Training time and memory usage
- *Bayesian poisoning*

Questions?

eberh060@morris.umn.edu

References

David Mandell Freeman
Using Naive Bayes to Detect Spammy Names in Social Networks
AISec13, November 4, 2013, Berlin, Germany

V. Metsis, I. Androutsopoulos, G. Paliouras
Spam Filtering with Naive Bayes - Which Naive Bayes?
CEAS 2006 - Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California USA