A spam classifier based on Bayes network

Alberto Franzin, Fabio Palese

Sistemi Intelligenti

January 15th, 2013

INTRODUCTION

The Bayesian approach

Bayes rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A|B)$ is the *a posteriori* probability of event A, knowing the event B has already occurred.

The Bayesian approach

Bayes rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A|B)$ is the *a posteriori* probability of event A, knowing the event B has already occurred.
- $P(B|A)$ is the *likelihood*

The Bayesian approach

Bayes rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A|B)$ is the *a posteriori* probability of event A, knowing the event B has already occurred.
- $P(B|A)$ is the *likelihood*
- $P(B|A)P(A)$ is the *prior* probability

The Bayesian approach

Bayes rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A|B)$ is the *a posteriori* probability of event A, knowing the event B has already occurred.
- $P(B|A)$ is the *likelihood*
- $P(B|A)P(A)$ is the *prior* probability
- $P(B) = \sum_{a \in A} P(B|A=a)P(A=a)$ is the *total* probability

The Bayesian approach

Bayes rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- In other words, we can estimate the probability of an hypothesis, given that we know the consequences.

The Bayesian approach

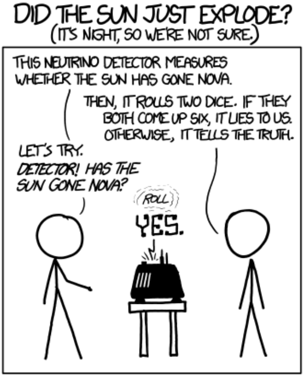
Bayes rule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- In other words, we can estimate the probability of an hypothesis, given that we know the consequences.
- This has led to two different interpretations of the theorem.

The Bayesian approach

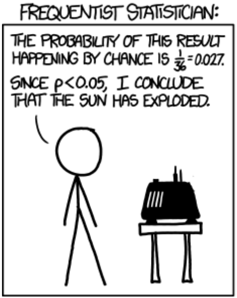
Frequentists vs. Bayesians



from http://xkcd.com/1132, see also
http://en.wikipedia.org/wiki/Sunrise_problem

The Bayesian approach

Frequentists vs. Bayesians



from http://xkcd.com/1132, see also
http://en.wikipedia.org/wiki/Sunrise_problem

The frequentist relies on the theoretical probability of the events.

The Bayesian approach

Frequentists vs. Bayesians

from http://xkcd.com/1132, see also
http://en.wikipedia.org/wiki/Sunrise_problem

The bayesian observes the past events occurred, and adapts the probability accordingly.

Bayesian networks

A Bayesian network is a way to describe causal relationships between events (J. Pearl, 1985).

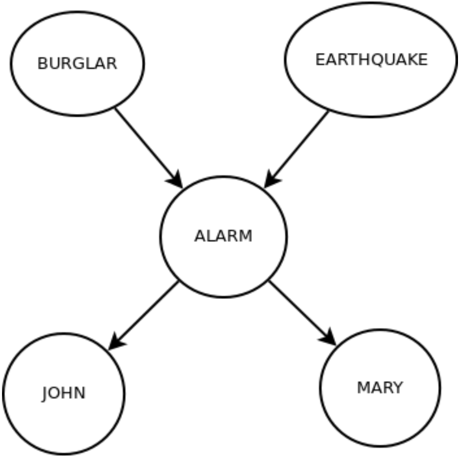
- Nodes = events
- (Directed) Edges = causal relationship

Bayesian networks

A Bayesian network is a way to describe causal relationships between events (J. Pearl, 1985).

- Nodes = events
- (Directed) Edges = causal relationship
- Two nodes are connected by an edge: the child of an arc is influenced by its ancestor in a probabilistic way

An example



Naive Bayes

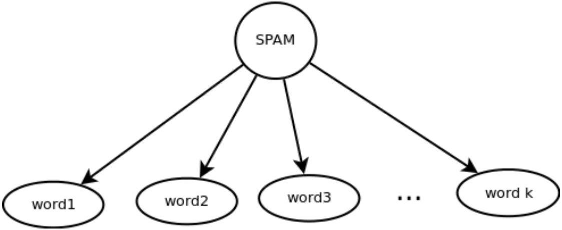
- Computing all the probabilities in a Bayesian network requires exponential time. We introduce the assumption of independence among variables.

NAIVE BAYES

- Computing all the probabilities in a Bayesian network requires exponential time. We introduce the assumption of independence among variables.
- It is called *naive*, since it's often unrealistic, but it yields good results.

Naive Bayes

- Computing all the probabilities in a Bayesian network requires exponential time. We introduce the assumption of independence among variables.
- It is called *naive*, since it's often unrealistic, but it yields good results.
- In spam classification:



NAIVE BAYES FOR SPAM CLASSIFICATION

We want to build a classifier that distinguishes good mails from undesired mails:

- good mails: *ham*
- undesired mails: *spam*

Naive Bayes for spam classification

Formulas:

- For each word:

$$P_{word|spam} = \frac{\text{\# occurrences of word in spam mails}}{\text{\# total occurrences of word}}$$

$$P_{spam|word} = \frac{P_{word|spam} P_{spam}}{P_{word}}$$

Naive Bayes for spam classification

Formulas:

- For each word:

$$P_{word|spam} = \frac{\text{\# occurrences of word in spam mails}}{\text{\# total occurrences of word}}$$

$$P_{spam|word} = \frac{P_{word|spam} P_{spam}}{P_{word}}$$

- Final for spam is:

$$P_{spam} = \prod_{words \in mail} P_{spam|word}$$

Naive Bayes for spam classification

Formulas:

- For each word:

$$P_{word|spam} = \frac{\text{\# occurrences of word in spam mails}}{\text{\# total occurrences of word}}$$

$$P_{spam|word} = \frac{P_{word|spam} P_{spam}}{P_{word}}$$

- Final for spam is:

$$P_{spam} = \prod_{words \in mail} P_{spam|word}$$

- same for ham

Naive Bayes for spam classification

Formulas:

- For each word:

$$P_{word|spam} = \frac{\text{\# occurrences of word in spam mails}}{\text{\# total occurrences of word}}$$

$$P_{spam|word} = \frac{P_{word|spam} P_{spam}}{P_{word}}$$

- Final for spam is:

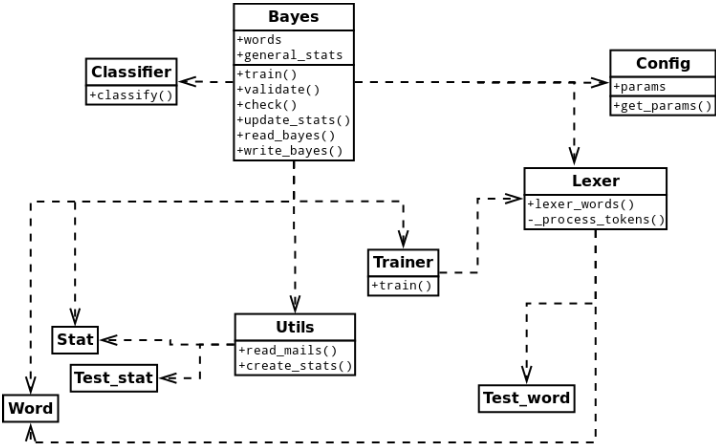
$$P_{spam} = \prod_{words \in mail} P_{spam|word}$$

- same for ham
- Outcome is the class that maximizes the probability of belonging to that class.

SpamBayes

- SpamBayes: applying Naive Bayes to spam classification
- Python with Ply and BeautifulSoup
- dataset: SpamAssassin archive
- Code and documentation available at `http://code.google.com/p/sist-int-2012project/`

SpamBayes


Bayes
+words
+general_stats
+train()
+validate()
+check()
+update_stats()
+read_bayes()
+write_bayes()
Classifier
+classify()
Config
+params
+get_params()
Lexer
+lexer_words()
-_process_tokens()
Trainer
+train()
Utils
+read_mails()
+create_stats()
Stat
Test_stat
Word
Test_word

Notes on implementation

- Smoothing:

$$P_{word|spam} = \frac{\#\text{ occurrences of word in spam mails} + k}{\#\text{ total occurrences of the word} + |C| \times k}$$

Notes on implementation

- Smoothing:

$$P_{word|spam} = \frac{\#\text{ occurrences of word in spam mails} + k}{\#\text{ total occurrences of the word} + |C| \times k}$$

- Calculations can be simplified: some words bring little contribution to the mail status

Notes on implementation

- Smoothing:

$$P_{word|spam} = \frac{\#\text{ occurrences of word in spam mails} + k}{\#\text{ total occurrences of the word} + |C| \times k}$$

- Calculations can be simplified: some words bring little contribution to the mail status
- Several mail features detected

Notes on implementation

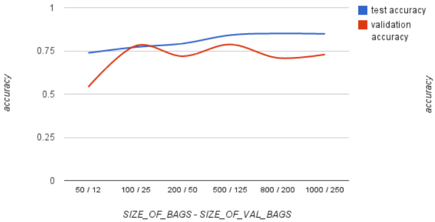
- Smoothing:

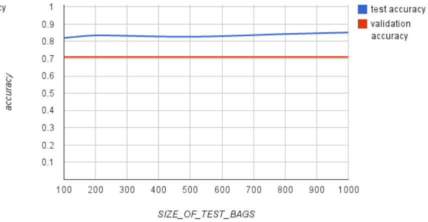
$$P_{word|spam} = \frac{\text{\# occurrences of word in spam mails} + k}{\text{\# total occurrences of the word} + |C| \times k}$$

- Calculations can be simplified: some words bring little contribution to the mail status
- Several mail features detected
- Several parameters to be tuned: we describe the more relevant ones

Parameters

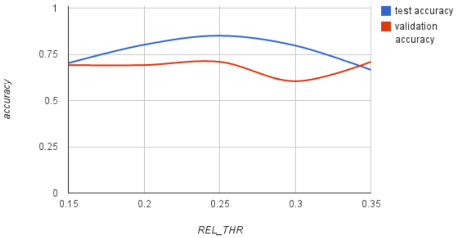
Size of training/validation/test sets:





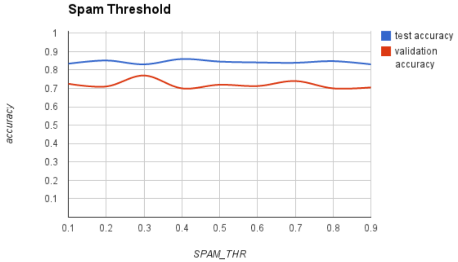
PARAMETERS

Relevance threshold:



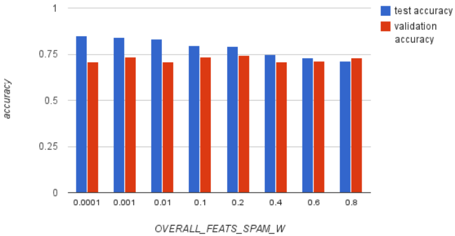
PARAMETERS

"Spamicity" threshold:



Parameters

Feature statistic threshold:



RESULTS

- About the dataset:

RESULTS

- About the dataset:
  - General features

RESULTS

- About the dataset:
  - General features
  - Single words

RESULTS

- About the dataset:
  - General features
  - Single words
- About the classification:

RESULTS

- About the dataset:
  - General features
  - Single words
- About the classification:
  - Accuracy

Results

- About the dataset:
  - General features
  - Single words
- About the classification:
  - Accuracy
  - How to improve?