A spam classifier based on Bayes network

Alberto Franzin, Fabio Palese

Sistemi Intelligenti

January 4, 2013

INTRODUCTION

THE BAYESIAN APPROACH

- Bayes rule:

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

defines the *a posteriori* probability of event A, knowing the event B has already occurred.

THE BAYESIAN APPROACH

- Bayes rule:

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

  defines the *a posteriori* probability of event A, knowing the event B has already occurred.
- In other words, we can estimate the probability of an hypothesis, given that we know the consequences.

THE BAYESIAN APPROACH

- Bayes rule:

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

  defines the *a posteriori* probability of event A, knowing the event B has already occurred.
- 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

- Bayes rule:

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

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

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
- $P(B|A)$ 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
- $P(B|A)$ is the *likelihood*
- $P(B|A)P(A)$ is the *prior* probability

The Bayesian approach

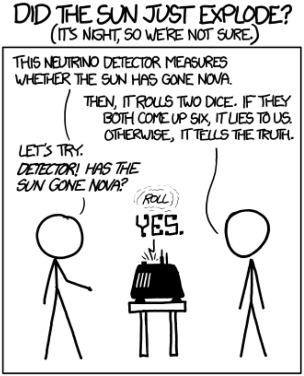
- Bayes rule:

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

- $P(A|B)$ is the *a posteriori* probability
- $P(B|A)$ 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

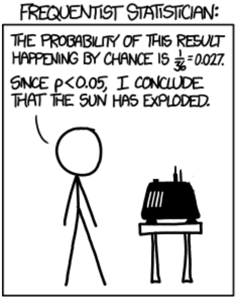
Frequentists vs. Bayesians



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.

What it is

A Bayes network is a way to describe causal relationships between events.

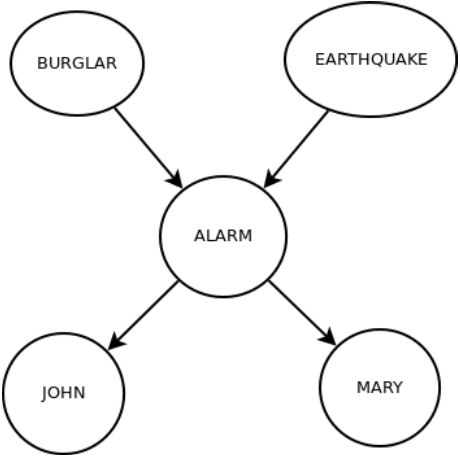
- Nodes = events
- (Directed) Edges = causal relationship

WHAT IT IS

A Bayes network is a way to describe causal relationships between events.

- 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



CONDITIONAL INDEPENDENCE

- If

$$P(A|B,C) = P(A|B)$$

then we say that B and C are *conditionally independent*.

CONDITIONAL INDEPENDENCE

- If
  $$P(A|B,C) = P(A|B)$$
  then we say that B and C are *conditionally independent*.
- Note that *conditional independence* $\neq$ *independence*

CONDITIONAL INDEPENDENCE

- If
  $$P(A|B,C) = P(A|B)$$
  then we say that B and C are *conditionally independent*.
- Note that *conditional independence* $\neq$ *independence*
- Explaining away: if we know that one possible cause of the event has happened, this may *explain away* the event, being all the other causes less probable once we know the one that happened.

NAIVE BAYES

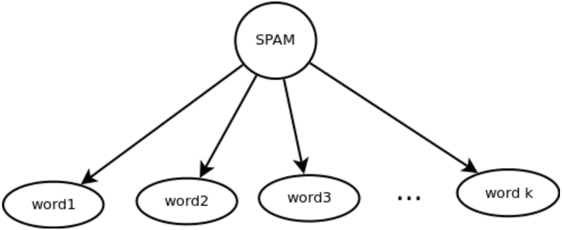
- Computing all the probabilities in a Bayesian network requires exponential time. We introduce the assumption of independence among variables.

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:



SPAMBAYES

Python, to use Ply and BeautifulSoup
dataset: SpamAssassin archive

FRAME 1