Bayes Theorem simply explained

With applications in Spam classifier and Autocorrect

Hung Tu Dinh

Jan 2018

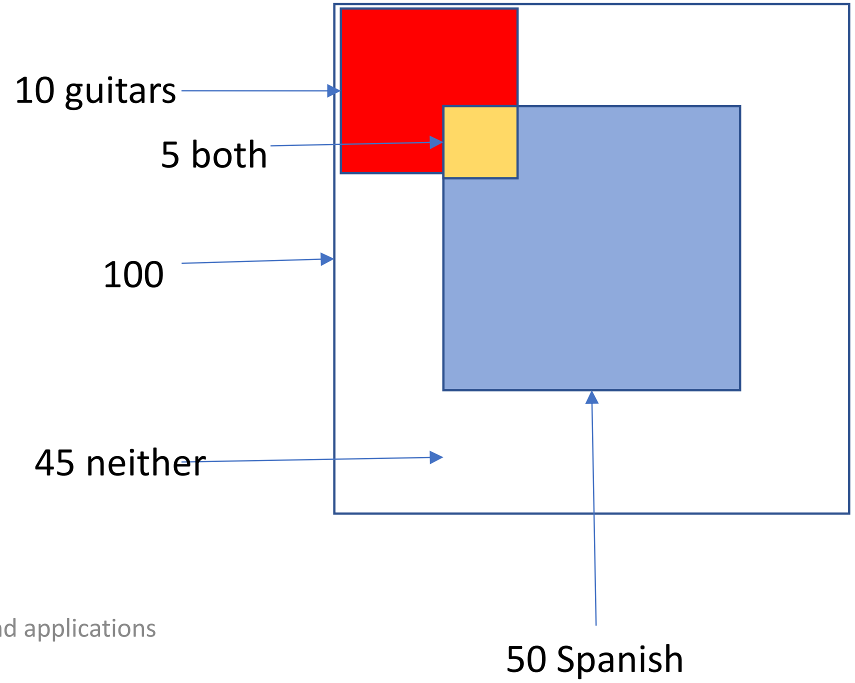
Agenda

- Part 1: Basic concepts of conditional probability and Bayes equation
- Part 2: How does spam filter work?
- Part 3: How does Auto-correct work?

Set up

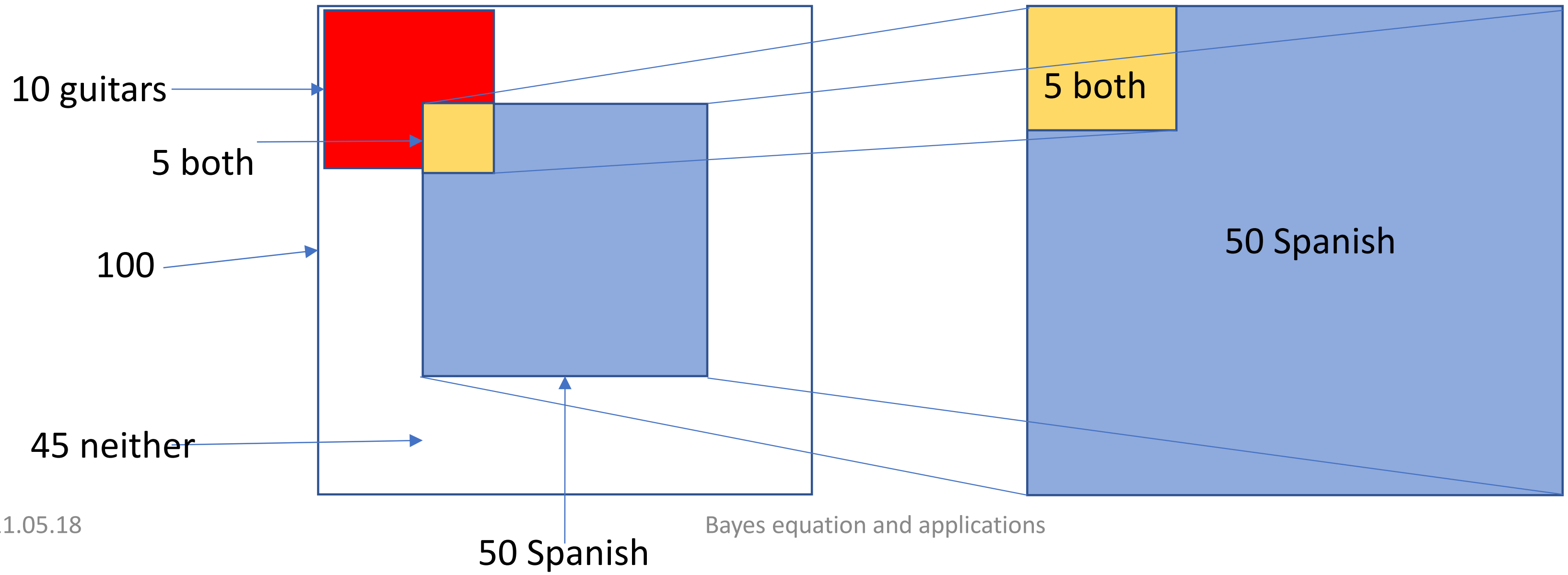
- In a school of 100 students
  - some take guitar in Music class
  - some take Spanish in Language class

Group	Value	Probability
Total	100	1
Guitar	10	0.1
Spanish	50	0.5
Both	5	0.05
Only guitar no Spanish	? 5	?
Neither	? 45	?



Conditional Probability

- P (Guitar & Spanish) = 5/100 = 0.05
- Among the 50 Spanish learners, only 5 (or 10%) also play guitar
- P(Guitar & Spanish | Spanish) = 5/50 = 0.10
- P(Guitar & Spanish | Guitar) = ? 5/10=0.5



Informal way of saying probability

- Probability is "chance" of a variable takes a value
- Music = {guitar, piano, violin}, Language = {Spanish, English, French}
- P(music=guitar)=0.1
- P(lang=Spanish)=0.5
- $\text{P(music=guitar|lang=Spanish)} = \frac{\text{Num(music=guitar \& lang=Spanish)}}{\text{Num(lang=Spanish)}} = \frac{\text{Num(music=guitar \& lang=Spanish) /Total}}{\text{Num(lang=Spanish) /Total}}$

- $\text{P(music=guitar|lang=Spanish)} = \frac{\text{P(music=guitar \& lang=Spanish)}}{\text{P(lang=Spanish)}}$

Rearrange the equation

- $\text{P(music=guitar|lang=Spanish)} = \frac{\text{P(music=guitar \& lang=Spanish)}}{\text{P(lang=Spanish)}}$
- Cross multiplication
  P(music=guitar & lang=Spanish) = P(lang=Spanish) x P(music=guitar|lang=Spanish)
  = 0.5 x 0.1 = 0.05 (also = 5/100=0.05).
- Intuitive:
  - Among the whole school, 50% learn Spanish.
  - Among the Spanish learners, 10% also learn guitar.
  - => Among the whole school, percentage of students doing both Spanish and guitar is 0.5 x 0.1 = 0.05.
- Likewise, switch the order of logic between language and music
- P(music=guitar & lang=Spanish)= P(lang=Spanish| music=guitar) x $\text{P}(\text{music} = \text{guitar})$=0.5 x 0.1=0.05

Bayes Equation

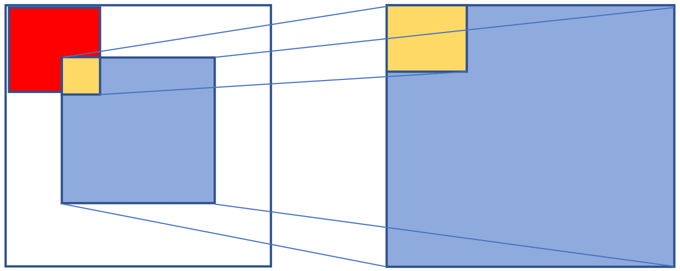
- Using symbol to make the equation more succint and general
- $P(Y=y|X=x)P(X=x) = P(X=x|Y=y)P(Y=y) = P(X=x \cap Y=y)$
- $P(Y=y|X=x) = \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)}$

Why the equation is needed?

- Because some probabilities is easier to compute than other.
- For example: Spam classifier
  - Objective: to estimate the probability of an email being a spam P(spam=True)
  - Not straightforward to get P(spam=True)
    - we look for other simpler probabilities
  - $P(spam = True|word_1 = money) = \frac{P(word_1 = money|spam = True) \times P(spam=True)}{P(word_1=money)}$
  - More straightforward to
    - Get P(spam=True) by counting spam emails/all emails available.
    - Among the spam, get the probability of the words e.g. chance of “money”
- More detailed of spam classifier in next part

Summary Part 1

- Probability is chance of a variable takes a (range) value
  - P(music=guitar), P(language=Spanish)
- Probability is associate with a population/sample
  - Among the whole school or only Spanish-learners
- Some probabilities are easier to compute
  - Use Bayes equation to compute the difficult one from the easier
  - $P(Y=y|X=x) = \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)}$



Part 2

- Frequentist vs Bayesian
  - Keyword: Updating belief
- Spam filtering application
  - Problem description & how it was solved
  - Propose method

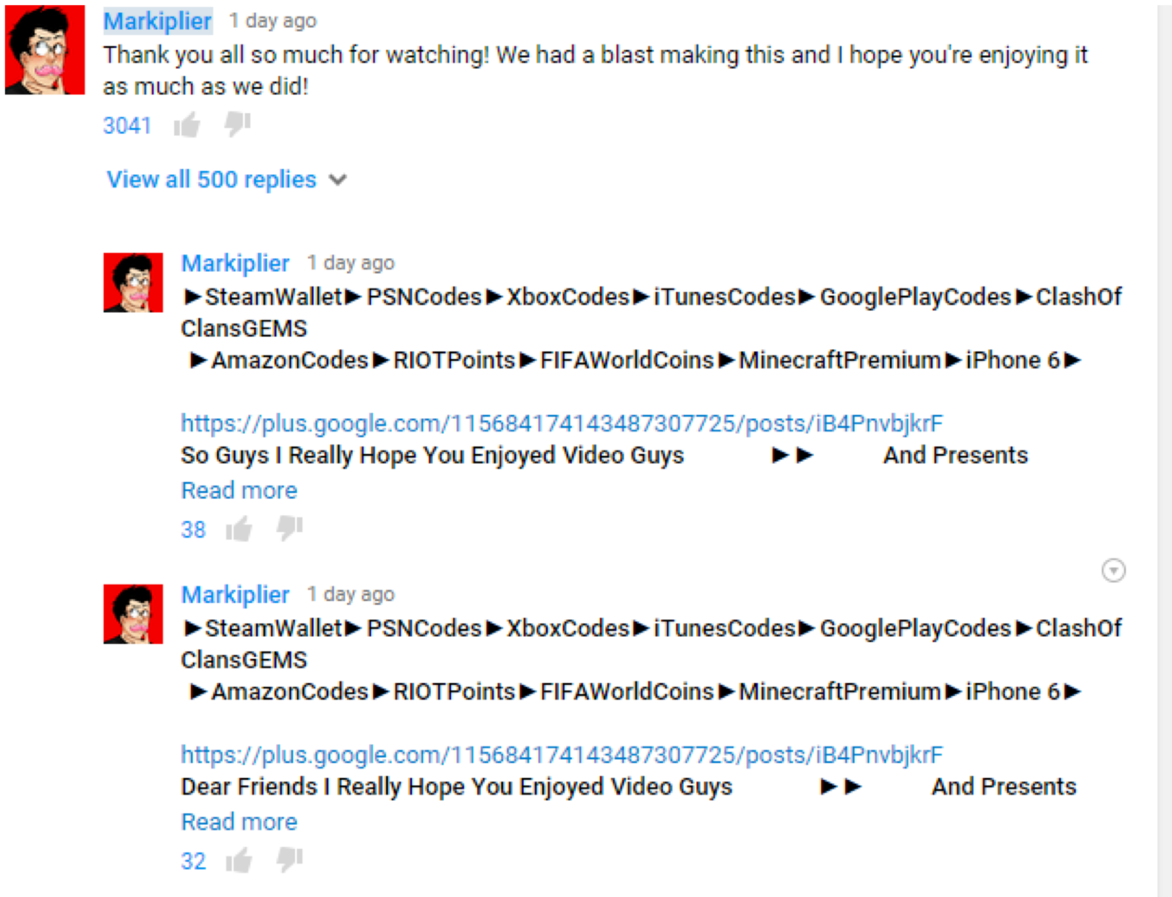
Two interpretations of probability

	Frequentist	**Bayesian (Belief)**
Example	Flip a coin for 100 times, on average, 50 heads and 50 tails	Weather forecast: quantify the belief that tomorrow is a sunny day
Interpretation	P=1 means the events always happen	P=1 means the belief is 100% certain
Situation	Common in scientific experiments	Belief changes before and after an evidence

Example: Vietnam U23 team in the tournament of 16 teams

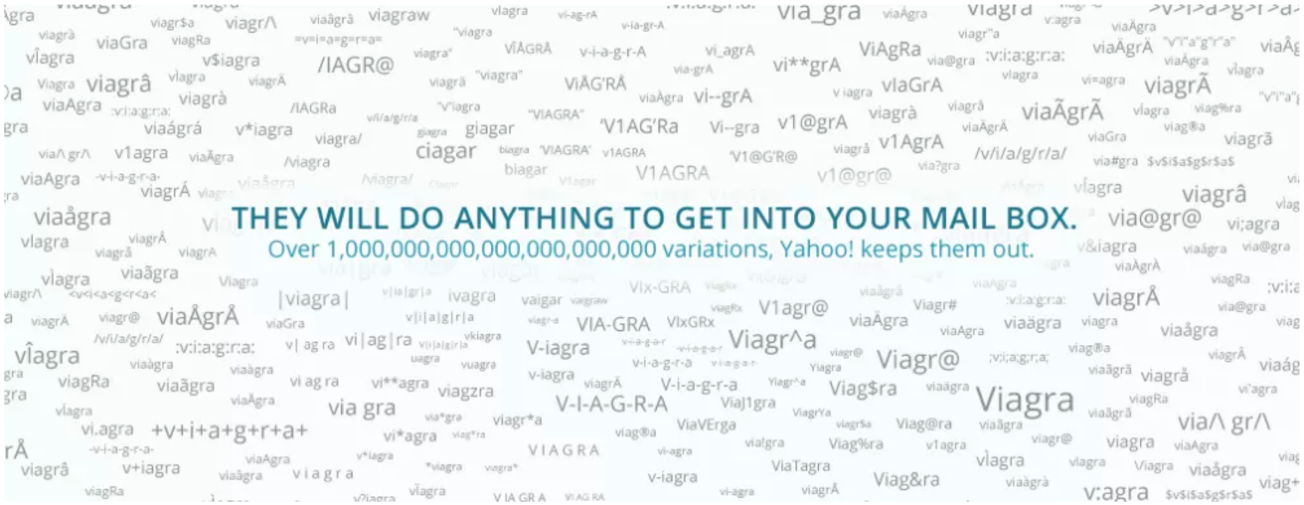
- Before the tournament, with no particular evidence, prior belief of winning ~ 1/16
- After some early games, with more evidence, update the belief of winning ~ 1/2

Spam filter is not only for email



How was it solved?

- Fighters: manually identify features to detect spam
  - Common words in spam: money, viagra, discount, deals, …
- Spammers circumvent



Source: https://flowingdata.com/2011/10/13/visualizing-yahoo-email-processing-in-real-time/

Some concepts

- "Spam" ~ undesired # "Scam" ~ fraudulent
- Measure of effectiveness
  - True positive vs false positive:
    - "true" if predicted correctly
    - "positive" is what you look for

    => "true positive" correct finding

    => "false negative" incorrect not finding

	Actual spam	Actual non-spam
Predicted as spam	True positive	**False positive**
Predicted as non-spam	False negative	True negative

Calculate spam score

- Find probability of an email being a spam, given a word in the email

$$P(spam = True|word_1 = money) = \frac{P(word_1 = money|spam = True)}{P(word_1 = money)} \times P(spam = True)$$

- Combine probabilities for all words in the email
  - Evidence 1: word "money" give $P_1 = P(spam = True|word_1 = money) = 0.9$
  - Evidence 1: word "hi" give $P_2 = P(spam = True|word_1 = hi) = 0.5$
  - …
  - Evidence n: word "hot" give $P_n = P(spam = True|word_n = hot) = 0.8$
  - $P_{combine} = \frac{P_1 \times \ldots \times P_n}{(P_1 \times \ldots \times P_n) + (1-P_1) \times \ldots \times (1-P_n)}$ simplified $\frac{abc}{abc + (1-a)(1-b)(1-c)}$

Implementation

$$\boldsymbol{P(spam = True|word_1 = money)} = \frac{\boldsymbol{P(word_1 = money|spam = True)}}{\boldsymbol{P(word_1 = money)}} \times \boldsymbol{P(spam = True)}$$

Train

Predict

1. Data sets: corpus of spam and corpus of non-spam $\boldsymbol{P(spam = True)}$
2. Tokenize (split into words) & count frequency of words in each corpus $P(word_1 = money|spam = True)$ & $P(word_1 = money|spam = False)$
3. Hash table: dict["money"]= $\frac{P(word_1 = money|spam = True)}{P(word_1 = money|spam = true \,\&\, false)} \times \boldsymbol{P(spam = True)}$
4. For each test email:
   1. Tokenize into words
   2. For each word, check in hash table to find $\boldsymbol{P(spam = True|word_1 = money)}$
   3. Combine probabilities from all words $\frac{abc}{abc+(1-a)(1-b)(1-c)}$

Updating the belief

- $P(spam = True|word_1 = money) = \frac{P(word_1 = money|spam = True)}{P(word_1 = money)} \times P(spam = True)$

- No evidence: global rate $P(spam = True)$ = 0.5.
- With only 1 evidence of word "money" the belief is $P_1 = 1.8 \times 0.5 = 0.9$
- With another evidence of "hi", $P_2 = 0.5$ the total belief is updated $\frac{0.9 \times 0.5}{0.9 \times 0.5 + (1 - 0,9) \times (1 - 0.5)} = 0.9$
- With another evidence of "hot", $P_3 = 0.8$ the total belief is updated $\frac{0.9 \times 0.8}{0.9 \times 0.5 + 0.1 \times 0.2} = 0.97$

$$\frac{abc}{abc + (1 - a)(1 - b)(1 - c)}$$

DEMO TIME

- Check the GitHub repository
  - https://github.com/browning/comment-troll-classifier

Strength of Bayesian filter

- Identify the features by itself (which words are good - bad)
- Automatically update the probability (“belief”) of spamming words
  - Robust against tricks of spammers
- Take into account both good words & suspicious words
- Customize for individual users (different training sets)

Summary Part 2

- Two interpretations
  - Frequentist: how many times an event occurs
  - Bayesian: how certain is the belief that an event occurring
- Bayes formula
  - Given a hypothesis H and evidence E
  - $P(H|E) = \frac{P(E|H)}{P(E)} \times P(H)$
  - Update belief about H after learning about the evidence E
- Spam filtering
  - Get evidence from each words using Bayes formula
  - Combine evidence from all words
    - Good words and bad words both contribute

Armchair philosophy

- *With more evidences, your belief about the world is updated.*
- *Among many possible evidences, choose wisely.*
- *Beware that we may reinforce our prior belief because of bias.*

Reference

- Brillant explanation https://brilliant.org/wiki/bayes-theorem/
- Some interesting articles from Paul Graham @ "Hackers & Painters"
  - http://www.paulgraham.com/spam.html (August 2002)
  - http://www.paulgraham.com/falsepositives.html
  - http://www.paulgraham.com/naivebayes.html
  - http://www.paulgraham.com/better.html (January 2003)