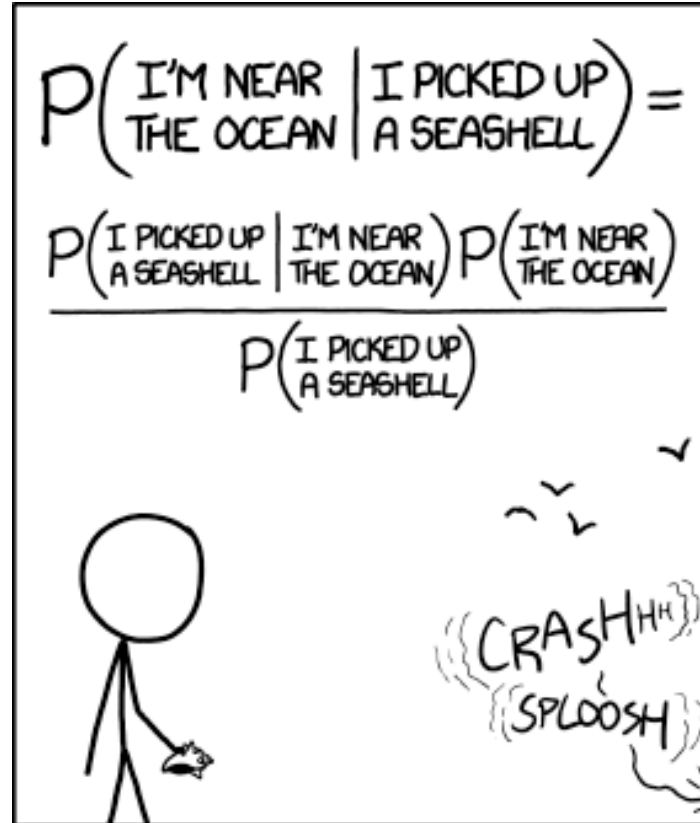


Bayesian inference, Naïve Bayes model



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND DON'T HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

Bayes Rule



Rev. Thomas Bayes
(1702-1761)

- The product rule gives us two ways to factor a joint probability:

$$P(A, B) =$$

- Therefore,
$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

- Why is this useful?
 - Can update our beliefs about A based on evidence B
 - $P(A)$ is the *prior* and $P(A|B)$ is the *posterior*
 - Key tool for probabilistic inference: can get *diagnostic probability* from *causal probability*
 - E.g., $P(\text{Cavity} = \text{true} | \text{Toothache} = \text{true})$ from $P(\text{Toothache} = \text{true} | \text{Cavity} = \text{true})$

Bayes Rule example

- Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year ($5/365 = 0.014$). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on Marie's wedding?

Bayes Rule example

- Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year ($5/365 = 0.014$). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on Marie's wedding?

$$\begin{aligned} P(\text{rain} \mid \text{predict}) &= \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict})} \\ &= \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict} \mid \text{rain})P(\text{rain}) + P(\text{predict} \mid \neg \text{rain})P(\neg \text{rain})} \end{aligned}$$

Law of total probability

$$\begin{aligned} P(X = x) &= \sum_{i=1}^n P(X = x, Y = y_i) \\ &= \sum_{i=1}^n P(X = x | Y = y_i) P(Y = y_i) \end{aligned}$$

Bayes Rule example

- Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year ($5/365 = 0.014$). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on Marie's wedding?

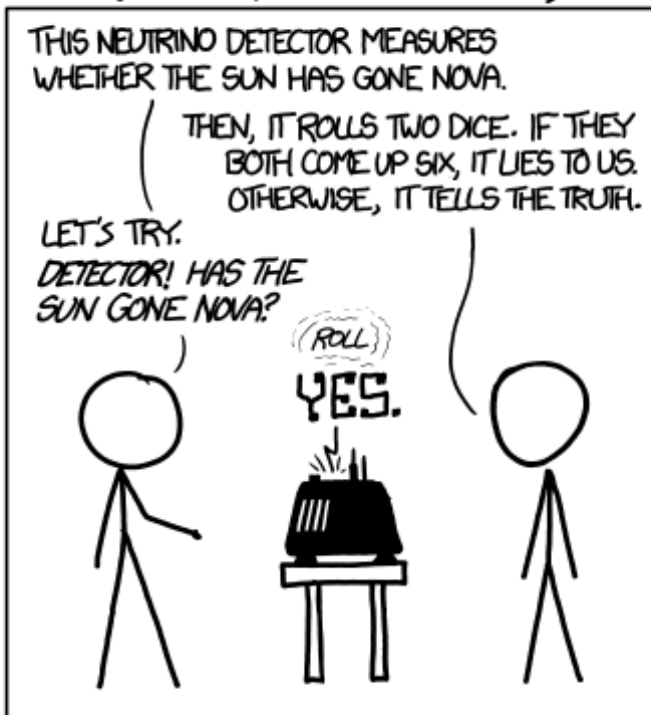
$$\begin{aligned} P(\text{rain} \mid \text{predict}) &= \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict})} \\ &= \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict} \mid \text{rain})P(\text{rain}) + P(\text{predict} \mid \neg \text{rain})P(\neg \text{rain})} \\ &= \frac{0.9 \times 0.014}{0.9 \times 0.014 + 0.1 \times 0.986} = \frac{0.0126}{0.0126 + 0.0986} = 0.111 \end{aligned}$$

Bayes rule: Example

- 1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

$$\begin{aligned}P(\text{cancer} \mid \text{positive}) &= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})} \\&= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive} \mid \text{cancer})P(\text{cancer}) + P(\text{positive} \mid \neg \text{cancer})P(\neg \text{cancer})} \\&= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.096 \times 0.99} = \frac{0.008}{0.008 + 0.095} = 0.0776\end{aligned}$$

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)



$$P(\text{nova} | \text{yes}) = \frac{P(\text{yes} | \text{nova})P(\text{nova})}{P(\text{yes})}$$

$$P(\neg \text{nova} | \text{yes}) = \frac{P(\text{yes} | \neg \text{nova})P(\neg \text{nova})}{P(\text{yes})}$$

FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.



BAYESIAN STATISTICIAN:

BET YOU \$50 IT HASN'T.



<https://xkcd.com/1132/>

See also: <https://xkcd.com/882/>

Probabilistic inference

- Suppose the agent has to make a decision about the value of an unobserved *query variable* X given some observed *evidence variable(s)* $E = e$
 - Partially observable, stochastic, episodic environment
 - Examples: $X = \{\text{spam, not spam}\}$, $e = \text{email message}$
 $X = \{\text{zebra, giraffe, hippo}\}$, $e = \text{image features}$



Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...

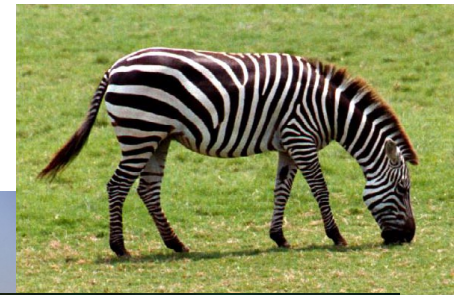


TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES
FOR ONLY \$99



Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.



Bayesian decision theory

- Let x be the value predicted by the agent and x^* be the true value of X .
- The agent has a **loss function**, which is 0 if $x = x^*$ and 1 otherwise
- Expected loss for predicting x :

$$\sum_{x^*} L(x, x^*) P(x^* | e)$$

- What is the estimate of X that minimizes the expected loss?
 - The one that has the greatest posterior probability $P(x|e)$
 - This is called the **Maximum a Posteriori (MAP)** decision

MAP decision

- Value x of X that has the highest posterior probability given the evidence $E = e$:

$$\hat{x} = \arg \max_x P(X = x | E = e) = \frac{P(E = e | X = x)P(X = x)}{P(E = e)}$$

$$\propto \arg \max_x P(E = e | X = x)P(X = x)$$

$$\underbrace{P(x | e)}_{\text{posterior}} \propto \underbrace{P(e | x)}_{\text{likelihood}} \underbrace{P(x)}_{\text{prior}}$$

Naïve Bayes model

- Suppose we have many different types of observations (symptoms, features) E_1, \dots, E_n that we want to use to obtain evidence about an underlying hypothesis X
- MAP decision:

$$\begin{aligned} P(X = x \mid E_1 = e_1, \dots, E_n = e_n) \\ \propto P(X = x)P(E_1 = e_1, \dots, E_n = e_n \mid X = x) \end{aligned}$$

Naïve Bayes model

- Suppose we have many different types of observations (symptoms, features) E_1, \dots, E_n that we want to use to obtain evidence about an underlying hypothesis X
- MAP decision:

$$\begin{aligned} P(X = x \mid E_1 = e_1, \dots, E_n = e_n) \\ \propto P(X = x)P(E_1 = e_1, \dots, E_n = e_n \mid X = x) \end{aligned}$$

- We can make the simplifying assumption that the different features are **conditionally independent given the hypothesis**:

$$P(E_1 = e_1, \dots, E_n = e_n \mid X = x) = \prod_{i=1}^n P(E_i = e_i \mid X = x)$$

Naïve Bayes model

- Posterior:

$$P(X = x \mid E_1 = e_1, \dots, E_n = e_n)$$

- MAP decision:

$$\hat{x} = \operatorname{argmax}_x \underbrace{P(x \mid e)}_{\text{posterior}} \propto \underbrace{P(x)}_{\text{prior}} \underbrace{\prod_{i=1}^n P(e_i \mid x)}_{\text{likelihood}}$$

Case study:

Text document classification

- **MAP decision:** assign a document to the class with the highest posterior $P(\text{class} \mid \text{document})$
- Example: spam classification
 - Classify a message as spam if $P(\text{spam} \mid \text{message}) > P(\neg \text{spam} \mid \text{message})$



Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...



TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES
FOR ONLY \$99



Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Case study:

Text document classification

- **MAP decision:** assign a document to the class with the highest posterior $P(\text{class} \mid \text{document})$
- We have $P(\text{class} \mid \text{document}) \propto P(\text{document} \mid \text{class})P(\text{class})$
- To enable classification, we need to be able to estimate the **likelihoods** $P(\text{document} \mid \text{class})$ for all classes and **priors** $P(\text{class})$

Naïve Bayes Representation

- Goal: estimate likelihoods $P(\text{document} \mid \text{class})$ and priors $P(\text{class})$
- Likelihood: **bag of words** representation
 - The document is a sequence of words (w_1, \dots, w_n)
 - The order of the words in the document is not important
 - Each word is conditionally independent of the others given document class



Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...



TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES
FOR ONLY \$99



Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Naïve Bayes Representation

- Goal: estimate likelihoods $P(\text{document} \mid \text{class})$ and priors $P(\text{class})$
- Likelihood: **bag of words** representation
 - The document is a sequence of words (w_1, \dots, w_n)
 - The order of the words in the document is not important
 - Each word is conditionally independent of the others given document class

$$P(\text{document} \mid \text{class}) = P(w_1, \dots, w_n \mid \text{class}) = \prod_{i=1}^n P(w_i \mid \text{class})$$

Bag of words illustration

2007-01-23: State of the Union Address

George W. Bush (2001-)

abandon accountable affordable afghanistan africa aided ally anbar armed army baghdad bless challenges chamber chaos
choices civilians coalition commanders commitment confident confront congressman constitution corps debates deduction
deficit deliver democratic deploy dikembe diplomacy disruptions earmarks economy einstein elections eliminates
expand extremists failing faithful families freedom fuel funding god haven ideology immigration impose
insurgents iran **iraq** islam julie lebanon love madam marine math medicare moderation neighborhoods nuclear offensive
palestinian payroll province pursuing **qaeda** radical regimes resolve retreat rieman sacrifices science sectarian senate
september shia stays strength students succeed sunni tax territories **terrorists** threats uphold victory
violence violent **war** washington weapons wesley

US Presidential Speeches Tag Cloud

<http://chir.ag/projects/preztags/>

Bag of words illustration

2007-01-23: State of the Union Address

George W. Bush (2001-)

1962-10-22: Soviet Missiles in Cuba

John F. Kennedy (1961-63)

abandon
choices c
deficit c
expand
insurgen
palestin
septemb
violenc

abandon achieving adversaries aggression agricultural appropriate armaments **arms** assessments atlantic ballistic berlin
buildup burdens cargo college commitment communist constitution consumers cooperation crisis **cuba** dangers
declined **defensive** deficit depended disarmament divisions domination doubled **economic** education
elimination emergence endangered equals **europa** expand exports fact false family forum **freedom** fulfill gromyko
halt hazards **hemisphere** hospitals ideals **independent** industries inflation labor latin limiting minister **missiles**
modernization neglect **nuclear** oas obligation observer **offensive** peril pledged predicted purchasing quarantine **quote**
recession rejection republics retaliatory safeguard sites solution **soviet** space spur stability standby **strength**
surveillance **tax** territory treaty undertakings unemployment **war** warhead **weapons** welfare western widen withdraw

US Presidential Speeches Tag Cloud

<http://chir.ag/projects/preztags/>

Bag of words illustration

2007-01-23: State of the Union Address

George W. Bush (2001-)

1962-10-22: Soviet Missiles in Cuba

John F. Kennedy (1961-63)

1941-12-08: Request for a Declaration of War

Franklin D. Roosevelt (1933-45)

abandoning acknowledge aggression aggressors airplanes armaments **armed** army assault assembly authorizations bombing
britain british cheerfully claiming constitution curtail december defeats defending delays **democratic** dictators disclose
economic empire endanger **facts** false forgotten fortunes france **freedom** fulfilled fullness fundamental gangsters
german germany **god** guam harbor hawaii **hemisphere** hint **hitler** hostilities immune improving indies innumerable
invasion **islands** isolate **japanese** labor metals midst midway **navy** nazis obligation offensive
officially **pacific** partisanship patriotism pearl peril perpetrated perpetual philippine preservation privilege reject
repaired **resisting** retain revealing rumors seas soldiers **speaks** speedy **stamina** **strength** sunday sunk supremacy tanks taxes
treachery true tyranny undertaken victory **war** wartime washington

US Presidential Speeches Tag Cloud

<http://chir.ag/projects/preztags/>

2016 convention speeches

Clinton



Trump

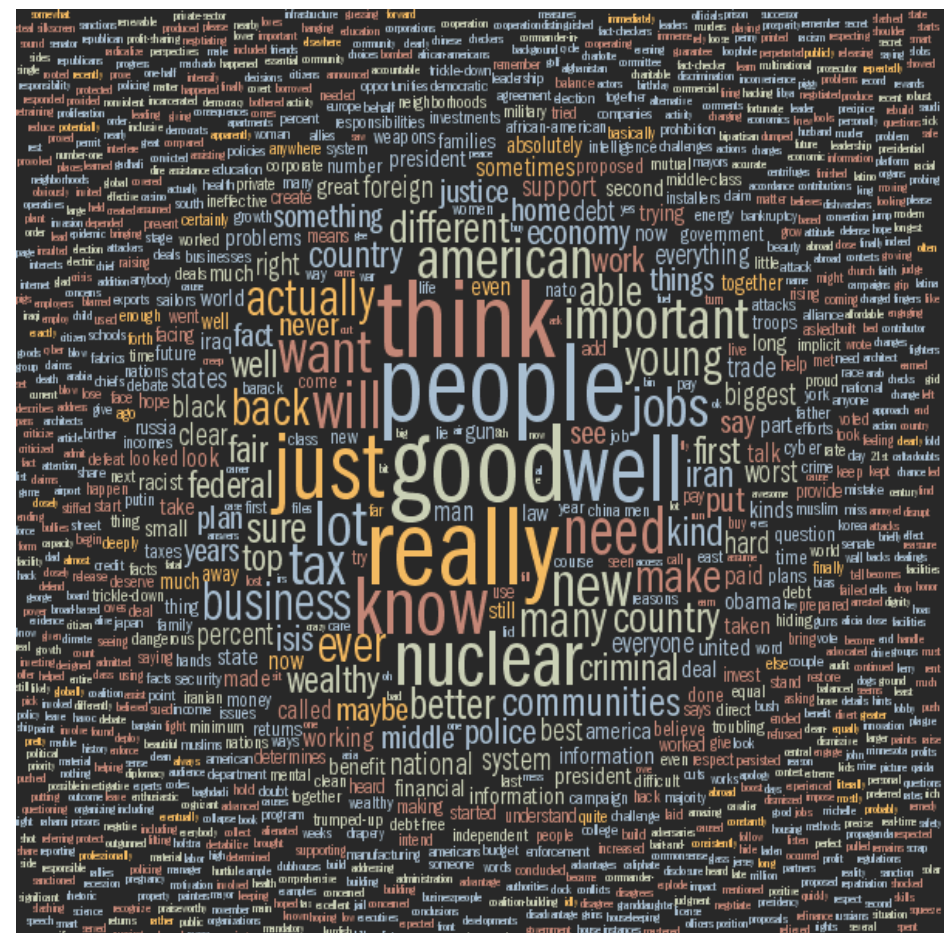
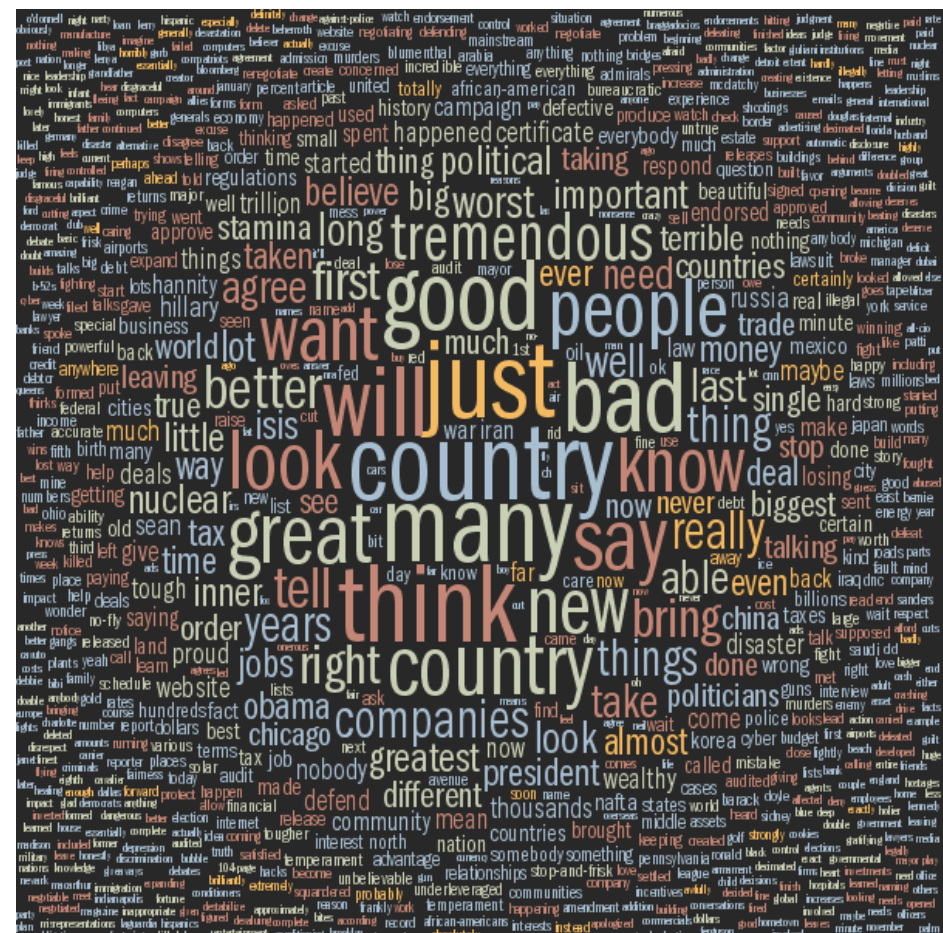


Source

2016 first presidential debate

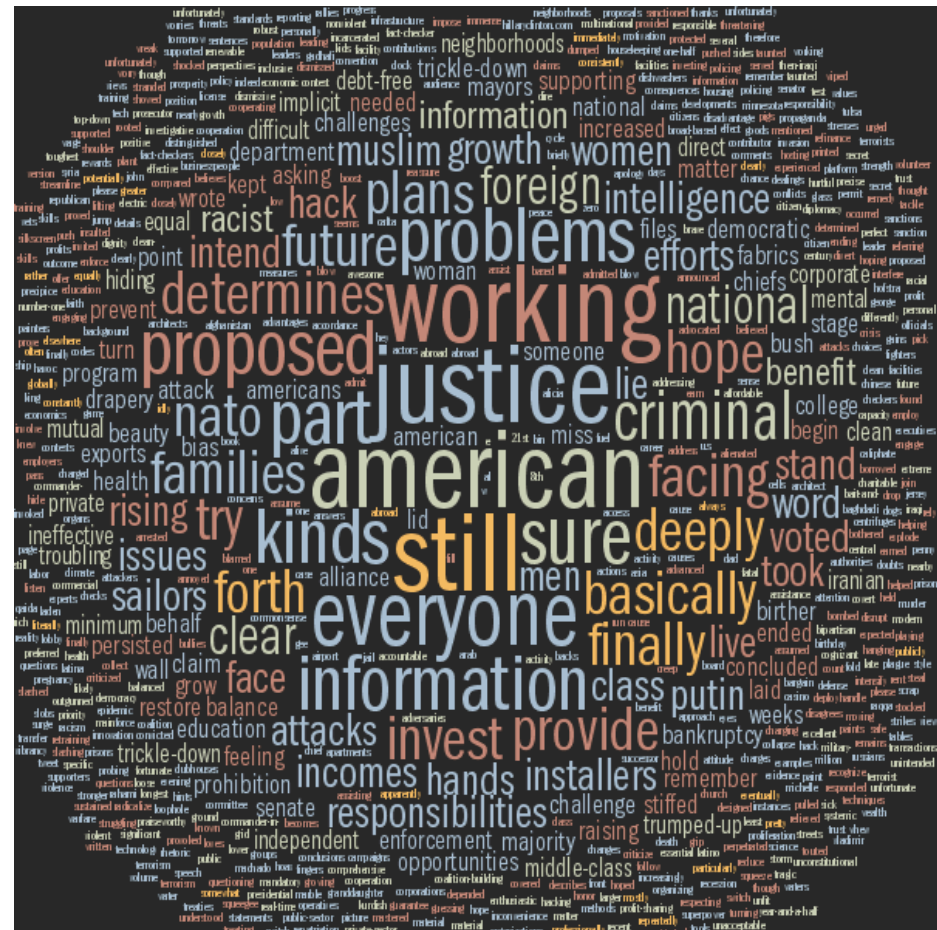
Trump

Clinton



Source

Clinton unique words



Source

Naïve Bayes Representation

- Goal: estimate likelihoods $P(\text{document} \mid \text{class})$ and $P(\text{class})$
- Likelihood: **bag of words** representation
 - The document is a sequence of words (w_1, \dots, w_n)
 - The order of the words in the document is not important
 - Each word is conditionally independent of the others given document class

$$P(\text{document} \mid \text{class}) = P(w_1, \dots, w_n \mid \text{class}) = \prod_{i=1}^n P(w_i \mid \text{class})$$

- Thus, the problem is reduced to estimating marginal likelihoods of individual words $P(w_i \mid \text{class})$

Parameter estimation

- Model parameters: feature likelihoods $P(\text{word} \mid \text{class})$ and priors $P(\text{class})$
 - How do we obtain the values of these parameters?

prior

spam:	0.33
¬spam:	0.67

$P(\text{word} \mid \text{spam})$

the :	0.0156
to :	0.0153
and :	0.0115
of :	0.0095
you :	0.0093
a :	0.0086
with:	0.0080
from:	0.0075
...	

$P(\text{word} \mid \neg\text{spam})$

the :	0.0210
to :	0.0133
of :	0.0119
2002:	0.0110
with:	0.0108
from:	0.0107
and :	0.0105
a :	0.0100
...	

Parameter estimation

- Model parameters: feature likelihoods $P(\text{word} \mid \text{class})$ and priors $P(\text{class})$
 - How do we obtain the values of these parameters?
 - Need *training set* of labeled samples from both classes

$$P(\text{word} \mid \text{class}) = \frac{\text{\# of occurrences of this word in docs from this class}}{\text{total \# of words in docs from this class}}$$

- This is the *maximum likelihood* (ML) estimate, or estimate that maximizes the likelihood of the training data:

$$\prod_{d=1}^D \prod_{i=1}^{n_d} P(w_{d,i} \mid \text{class}_{d,i})$$

d : index of training document, i : index of a word

Parameter estimation

- Parameter estimate:

$$P(\text{word} \mid \text{class}) = \frac{\text{\# of occurrences of this word in docs from this class}}{\text{total \# of words in docs from this class}}$$

- Parameter smoothing: dealing with words that were never seen or seen too few times
 - **Laplacian smoothing:** pretend you have seen every vocabulary word one more time than you actually did

$$P(\text{word} \mid \text{class}) = \frac{\text{\# of occurrences of this word in docs from this class} + 1}{\text{total \# of words in docs from this class} + V}$$

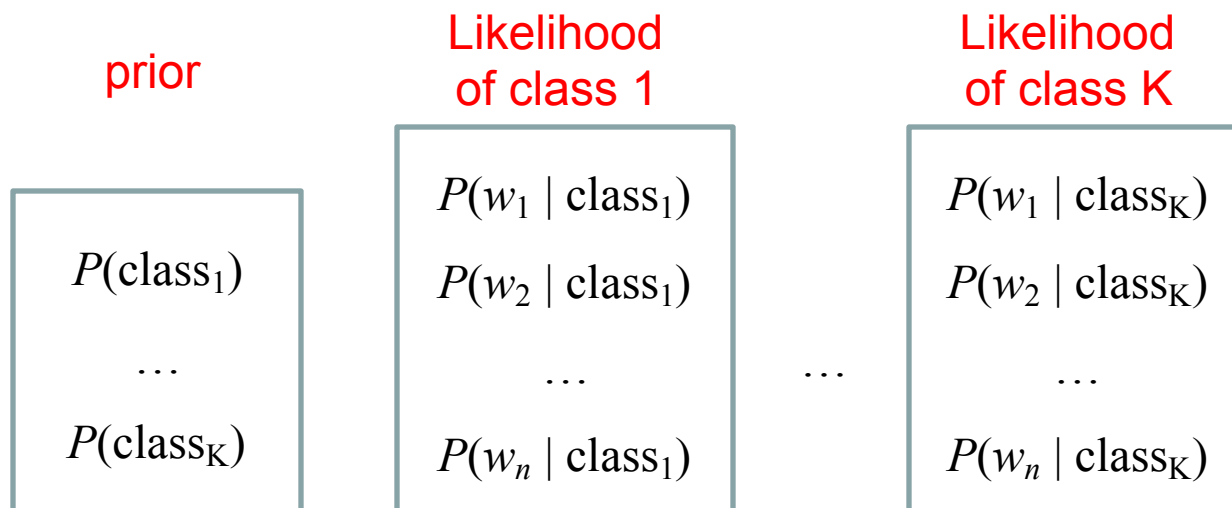
(V: total number of unique words)

Summary: Naïve Bayes for Document Classification

- Assign the document to the class with the highest posterior

$$P(\text{class} | \text{document}) \propto P(\text{class}) \prod_{i=1}^n P(w_i | \text{class})$$

- Model parameters:



Summary: Naïve Bayes for Document Classification

- Assign the document to the class with the highest posterior

$$P(class | document) \propto P(class) \prod_{i=1}^n P(w_i | class)$$

- Note: by convention, one typically works with logs of probabilities instead:

$$L(class | document) = \log P(class) + \sum_{i=1}^n \log P(w_i | class)$$

- Can help to avoid underflow

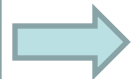
Learning and inference pipeline

Learning

Training Samples



Features



Training Labels



Training



Learned model

Inference



Test Sample



Features



Prediction

Learned model



Review: Bayesian decision making

- Suppose the agent has to make decisions about the value of an unobserved *query variable* X based on the values of an observed *evidence variable* E
- **Inference problem:** given some evidence $E = e$, what is $P(X | e)$?
- **Learning problem:** estimate the parameters of the probabilistic model $P(X | E)$ given a *training sample* $\{(x_1, e_1), \dots, (x_n, e_n)\}$