NAÏVE BAYES CLASSIFIERS

RECAP

Probabilistic agent has a numerical degree of belief between 0 (false) and 1 (true)

Unconditional Probability

- **P(a)**, the probability of "a" being true, or **P(a=True)**
- Does not depend on anything else to be true (**unconditional**)
- Represents the probability prior to further information that may adjust it (**prior**)

Conditional Probability

- **P(a|b)**, the probability of "a" being true, given that "b" is true
- Relies on "b" = true (**conditional**)
- Represents the prior probability adjusted based upon new information "b" (**posterior**)
- Can be generalized to more than 2 random variables: e.g. P(a|b, c, d)

Joint Probability

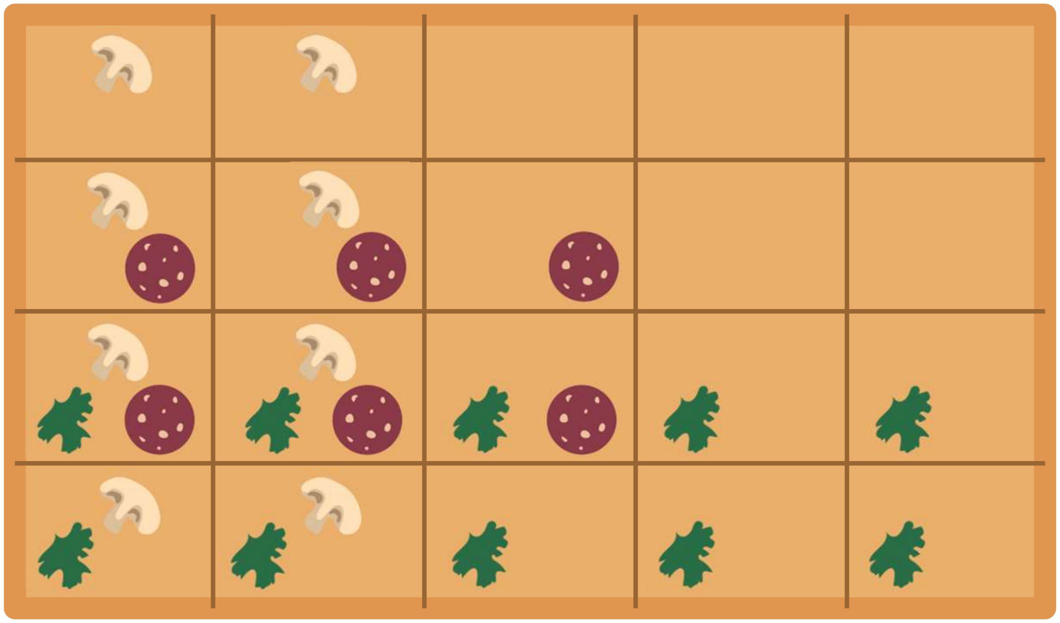
- **P(a, b) = P(a ^b)**, the probability of "a" and "b" both being true
- Can be generalized to more than 2 random variables: e.g. P(a, b, c, d)

ANSWER ANY QUERY FROM JOINT DISTRIBUTION

What is the probability of getting a slice with:

1) No mushrooms
2) Spinach and no mushrooms
3) Spinach, when asking for slice with no mushrooms

- Mushrooms
- Spinach
- No spinach
- No spinach and mushrooms
- No spinach when asking for no mushrooms
- No spinach when asking for mushrooms
- Spinach when asking for mushrooms
- No mushrooms and no spinach



Icons: CC, https://openclipart.org/detail/296791/pizza-slice

ANSWER ANY QUERY FROM JOINT DISTRIBUTION

You can answer all of these questions:

$P(M)$

m_1	12/20
m_2	

$P(S)$

s_1	
s_2	

$P(M,S)$

m_1	s_1	
m_1	s_2	6/20
m_2	s_1	
m_2	s_2	

$P(M|\mathrm{s}_1)$

m_1	
m_2	

$P(M|\mathrm{s}_2)$

m_1	
m_2	

$P(S|m_1)$

s_1	
s_2	6/12

$P(S|\mathrm{m}_2)$

s_1	
s_2	

ANSWER ANY QUERY FROM JOINT DISTRIBUTION

P(Weather)?

P(Weather | winter)?

P(Weather | winter, hot)?

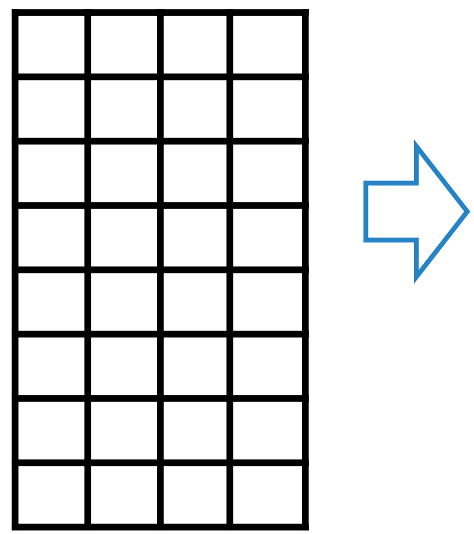
Season	Temp	Weather	P(S, T, W)
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

ANSWER ANY QUERY FROM JOINT DISTRIBUTION

Joint distributions are the best!

Problems with joints

- Huge
  - n variables with d values
  - d^n entries
- We aren't given the joint table
  - Usually some set of conditional probability tables

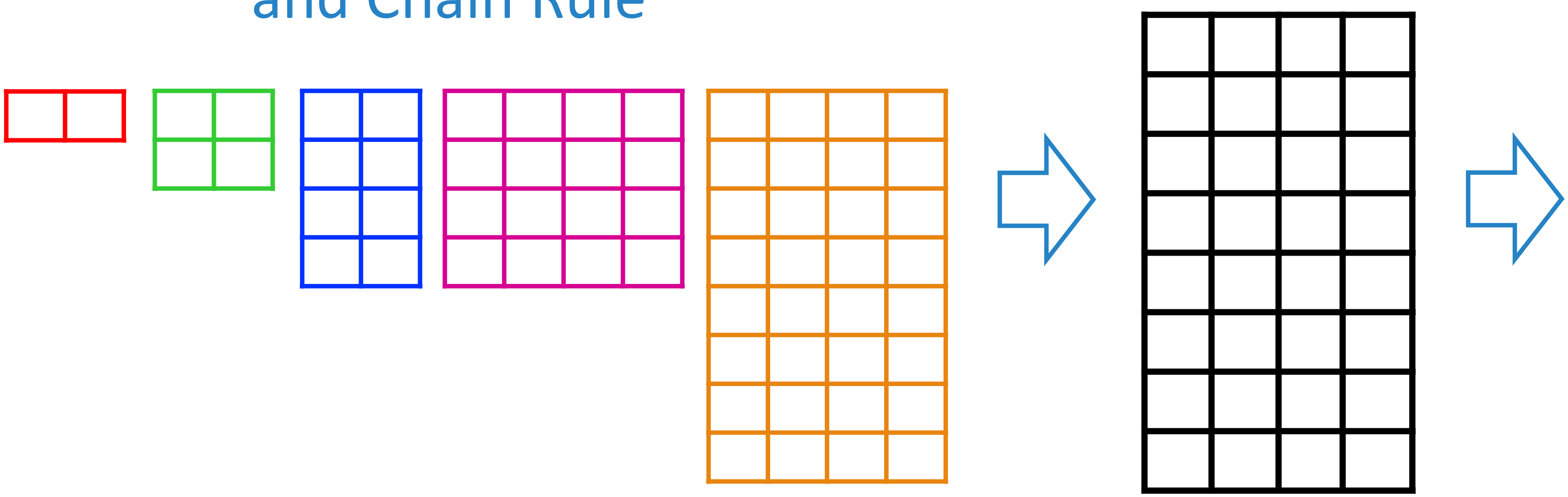


BUILD JOINT DISTRIBUTION USING CHAIN RULE

Conditional Probability Tables and Chain Rule

Joint

Query



$P(a \mid e)$

$P(A)\ P(B|A)\ P(C|A,B)\ P(D|A,B,C)\ P(E|A,B,C,D)$

BUILD JOINT DISTRIBUTION USING CHAIN RULE

Two tools to construct joint distribution

1. Product rule

$$P(A,B) = P(A \mid B)P(B)$$
$$P(A,B) = P(B \mid A)P(A)$$

2. Chain rule

$$P(X_1, X_2, \dots, X_n) = \prod_i P(X_i \mid X_1, \dots, X_{i-1})$$

$P(A,B,C) = P(A)P(B \mid A)P(C \mid A,B)$ for ordering A, B, C

$P(A,B,C) = P(A)P(C \mid A)P(B \mid A,C)$ for ordering A, C, B

$P(A,B,C) = P(C)P(B \mid C)P(A \mid C,B)$ for ordering C, B, A

...

USING THE PRODUCT RULE

- **Applies to any number of variables:**
  - $P(a, b, c) = P(a, b|c)\, P(c) = P(a|b, c)\, P(b, c)$
  - $P(a, b, c|d, e) = P(a|b, c, d, e)\, P(b, c|d, e)$

- **Factoring**: (AKA **Chain Rule** for probabilities)
  - By the product rule, we can always write:

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b, c, \ldots z)$$

We often use comma to abbreviate AND.

  - Repeatedly applying this idea, we can write:

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b \mid c, .. z)\, P(c| .. z) .. P(z)$$

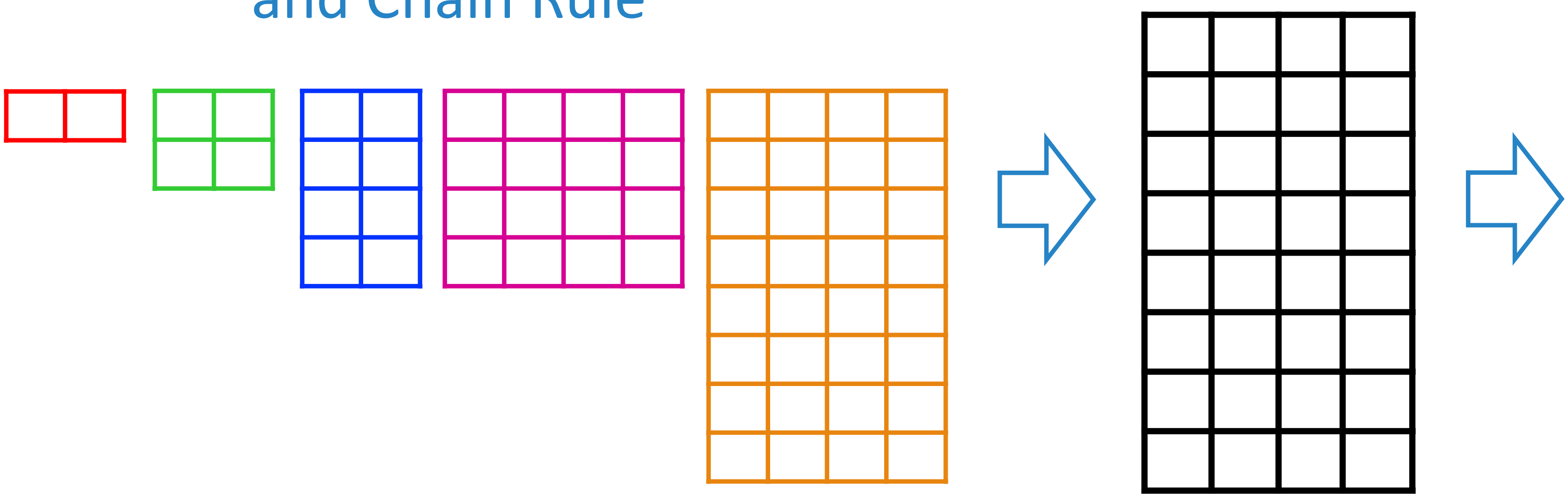
  - This holds for any ordering of the variables

ANSWER ANY QUERY FROM CONDITION PROBABILITY TABLES

Conditional Probability Tables
and Chain Rule

Joint

Query



$P(a \mid e)$

$P(A)$ $P(B|A)$ $P(C|A,B)$ $P(D|A,B,C)$ $P(E|A,B,C,D)$

ANSWER ANY QUERY FROM CONDITION PROBABILITY TABLES

Process to go from (specific) conditional probability tables to query

Construct the joint distribution

- Product Rule or Chain Rule

Answer query from joint

- Definition of conditional probability
- Law of total probability (marginalization, summing out)

ANSWER ANY QUERY FROM CONDITION PROBABILITY TABLES

Bayes' rule as an example

Given: $P(E|Q),\ P(Q)$ Query: $P(Q \mid e)$

Construct the joint distribution

- Product Rule or Chain Rule
- $P(E, Q) = P(E|Q)P(Q)$

Answer query from joint

- Definition of conditional probability
- $P(\,Q \mid e\,) = \frac{P(e,Q)}{P(e)}$
- Law of total probability (marginalization, summing out)

$$P(\,Q \mid e\,) = \frac{P(e,Q)}{\sum_q P(e,q)}$$

BAYESIAN NETWORK

TYPES OF CLASSIFIERS

- We can divide the large variety of classification approaches into three major types

1. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors
2. Generative:
   - build a generative statistical model
   - e.g., Bayesian networks
3. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., decision tree

BAYES DECISION RULE

- If we know the conditional probability P(X | y) we can determine the appropriate class by using Bayes rule:

$$P(y=i \mid X) = \frac{P(X \mid y=i)P(y=i)}{P(X)} \stackrel{def}{=} q_i(X)$$

But how do we determine p(X|y)?

COMPUTING P(X | Y)

Recall…

y – the class label

X – input attributes (features)

- Consider a dataset with 16 attributes (lets assume they are all binary). How many parameters to we need to estimate to fully determine p(X|y)?

age	employme	education	edun	marital	…	job	relation	race	gender	hour	country	wealth
					…							
39	State_gov	Bachelors	13	Never_mar	…	Adm_cleri	Not_in_fam	White	Male	40	United_Sta	poor
51	Self_emp_	Bachelors	13	Married	…	Exec_man	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	…	Handlers_	Not_in_fam	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	…	Handlers_	Husband	Black	Male	40	United_Sta	poor
28	Private	Bachelors	13	Married	…	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	…	Exec_man	Wife	White	Female	40	United_Sta	poor
50	Private	9th	5	Married_sp	…	Other_serv	Not_in_fam	Black	Female	16	Jamaica	poor
52	Self_emp_	HS_grad	9	Married	…	Exec_man	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	…	Prof_speci	Not_in_fam	White	Female	50	United_Sta	rich
42	Private	Bachelors	13	Married	…	Exec_man	Husband	White	Male	40	United_Sta	rich
37	Private	Some_coll	10	Married	…	Exec_man	Husband	Black	Male	80	United_Sta	rich
30	State_gov	Bachelors	13	Married	…	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	…	Adm_cleri	Own_child	White	Female	30	United_Sta	poor
33	Private	Assoc_ac	12	Never_mar	…	Sales	Not_in_fam	Black	Male	50	United_Sta	poor
41	Private	Assoc_voc	11	Married	…	Craft_repai	Husband	Asian	Male	40	*MissingV	rich
34	Private	7th_8th	4	Married	…	Transport_	Husband	Amer_Indi	Male	45	Mexico	poor
26	Self_emp_	HS_grad	9	Never_mar	…	Farming_fi	Own_child	White	Male	35	United_Sta	poor
33	Private	HS_grad	9	Never_mar	…	Machine_o	Unmarried	White	Male	40	United_Sta	poor
38	Private	11th	7	Married	…	Sales	Husband	White	Male	50	United_Sta	poor
44	Self_emp_	Masters	14	Divorced	…	Exec_man	Unmarried	White	Female	45	United_Sta	rich
41	Private	Doctorate	16	Married	…	Prof_speci	Husband	White	Male	60	United_Sta	rich

Learning the values for the full conditional probability table would require enormous amounts of data

NAÏVE BAYES CLASSIFIER

- Naïve Bayes classifiers assume that given the class label (Y) the attributes are **conditionally independent** of each other:

$$X = \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix}$$

$$\underbrace{p(X \mid y) = \prod_j}_{\text{Product of probability terms}} \underbrace{p_j(x^j \mid y)}_{\text{Specific model for attribute } j}$$

Product of probability terms

Specific model for attribute *j*

- Using this idea the full classification rule becomes:

$$\begin{aligned} \hat{y} &= \arg\max_v p(y = v \mid X) \\ &= \arg\max_v \frac{p(X \mid y = v)\, p(y = v)}{p(X)} \\ &= \arg\max_v \prod_j p_j(x^j \mid y = v)\, p(y = v) \end{aligned}$$

v are the classes we have

CONDITIONAL LIKELIHOOD: FULL VERSION

$$L(X_i \mid y_i = 1, \Theta) = \prod_j p(x_i^j \mid y_i = 1, \theta_1^j)$$

Vector of binary attributes for sample *i*

The set of all parameters in the NB model

The specific parameters for attribute *j* in class 1

Note the following:

1. We assume conditional independence between attributes **given** the class label
2. We learn a **different** set of parameters for the two classes (class 1 and class 2).

LEARNING PARAMETERS

$$L(X_i \mid y_i = 1, \Theta) = \prod_j p(x_i^j \mid y_i = 1, \theta_1^j)$$

- Let $X_1 \ldots X_{k1}$ be the set of input samples with label'y=1'
- Assume all attributes are **binary**
- To determine the MLE parameters for $p(x^j = 1 \mid y = 1)$ we simply count how many times the j'th entry of those samples in class 1 is 0 (termed n0) and how many times its 1 (n1). Then we set:

$$p(x^j = 1 \mid y = 1) = \frac{n1}{n0 + n1}$$

FINAL CLASSIFICATION

- Once we computed all parameters for attributes in both classes we can easily decide on the label of a **new** sample X.

Can be easily be extended to multi-class classification

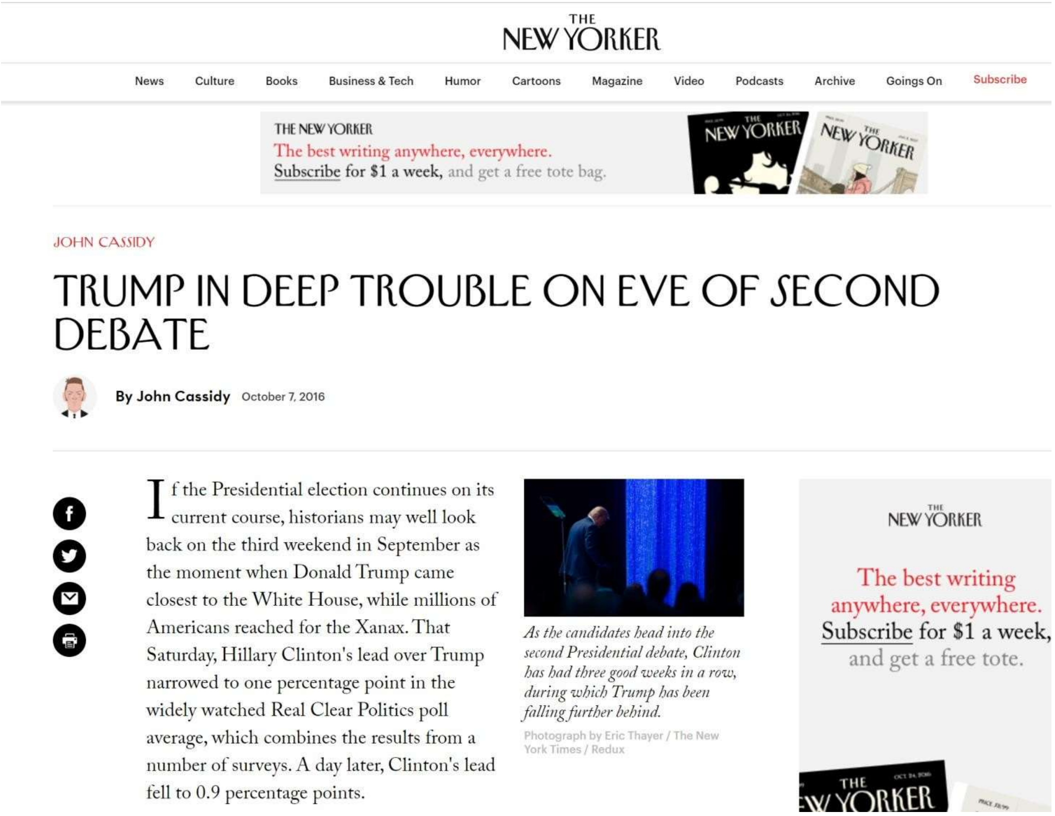
$$\hat{y} = \arg\max_v p(y = v \mid X)$$
$$= \arg\max_v \frac{p(X \mid y = v) p(y = v)}{p(X)}$$
$$= \arg\max_v \prod_j p_j(x^j \mid y = v) p(y = v)$$

Perform this computation for both class 1 and class 2 and select the class that leads to a higher probability as your decision

Prior on the prevalence of samples from each class

EXAMPLE: TEXT CLASSIFICATION

- What is the major topic of this article?



THE NEW YORKER

News | Culture | Books | Business & Tech | Humor | Cartoons | Magazine | Video | Podcasts | Archive | Goings On | Subscribe

JOHN CASSIDY

TRUMP IN DEEP TROUBLE ON EVE OF SECOND DEBATE

By John Cassidy October 7, 2016

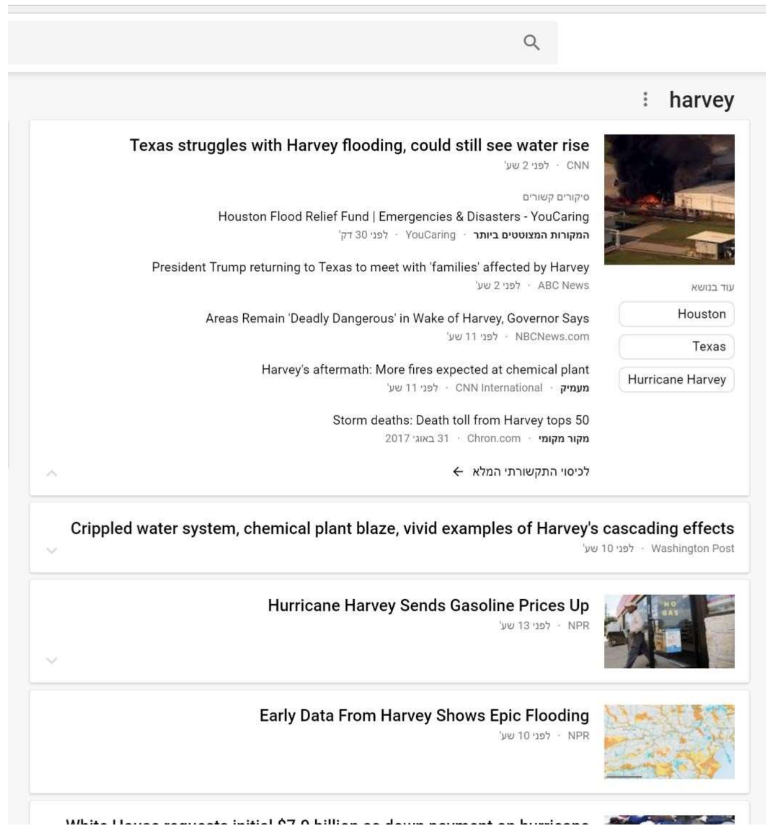
If the Presidential election continues on its current course, historians may well look back on the third weekend in September as the moment when Donald Trump came closest to the White House, while millions of Americans reached for the Xanax. That Saturday, Hillary Clinton's lead over Trump narrowed to one percentage point in the widely watched Real Clear Politics poll average, which combines the results from a number of surveys. A day later, Clinton's lead fell to 0.9 percentage points.

As the candidates head into the second Presidential debate, Clinton has had three good weeks in a row, during which Trump has been falling further behind.

Photograph by Eric Thayer / The New York Times / Redux

EXAMPLE: TEXT CLASSIFICATION

- Text classification is all around us



FEATURE TRANSFORMATION

- How do we encode the set of features (words) in the document?
- What type of information do we wish to represent? What can we ignore?
- Most common encoding: '**Bag of Words**'
- Treat document as a collection of words and encode each document as a vector based on some dictionary
- The vector can either be binary (present / absent information for each word) or discrete (number of appearances)

- Google is a good example
- Other applications include job search adds, spam filtering and many more.

FEATURE TRANSFORMATION: BAG OF WORDS

- In this example we will use a binary vector
- For document X_i we will use a vector of *m** indicator features $\{\phi^j(X_i)\}$ for whether a word appears in the document
  - $\phi^j(X_i) = 1$, if word *j* appears in document X_i;
    $\phi^j(X_i) = 0$ if it does not appear in the document
- $\Phi(X_i) = [\phi^1(X_i) \ldots \phi^m(X_i)]^T$ is the resulting feature vector for the entire dictionary for document X_i
- For notational simplicity we will replace each document X_i with a fixed length vector $\Phi_i = [\phi^1 \ldots \phi^m]^T$, where $\phi^j = \phi^j(X_i)$.

EXAMPLE

Assume we would like to classify documents as election related or not.

Dictionary

- Washington
- Congress

…

54. Trump

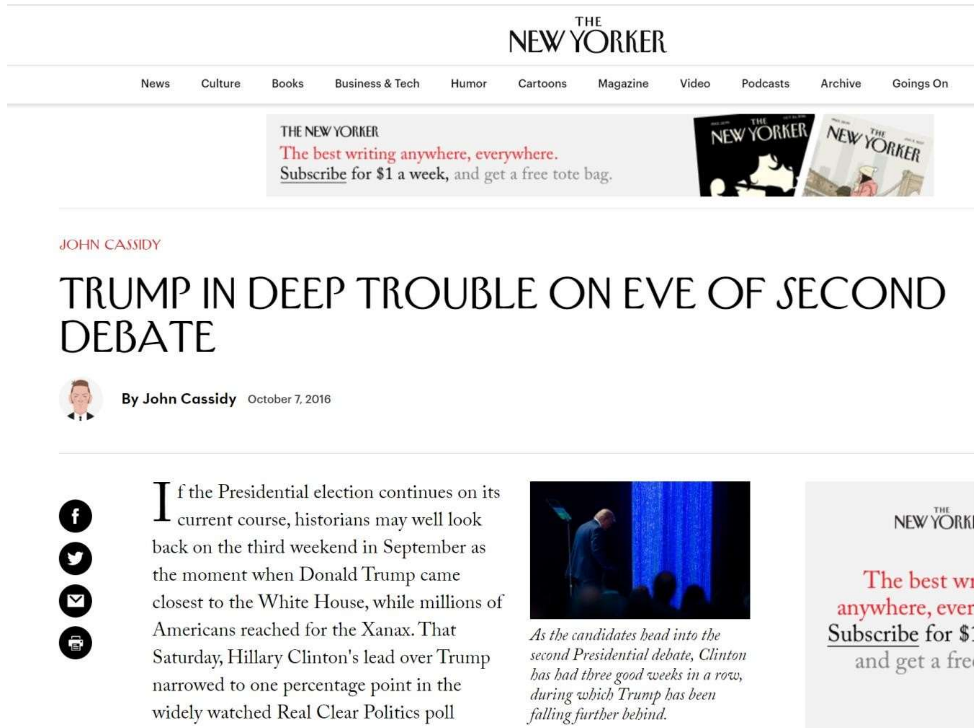
55. Clinton

56. Russia

$$\phi^{54}=\phi^{54}(X_i) = 1$$

$$\phi^{55}=\phi^{55}(X_i) = 1$$

$$\phi^{56}=\phi^{56}(X_i) = 0$$



THE NEW YORKER

News Culture Books Business & Tech Humor Cartoons Magazine Video Podcasts Archive Goings On

THE NEW YORKER
The best writing anywhere, everywhere.
Subscribe for $1 a week, and get a free tote bag.

JOHN CASSIDY

TRUMP IN DEEP TROUBLE ON EVE OF SECOND DEBATE

By John Cassidy October 7, 2016

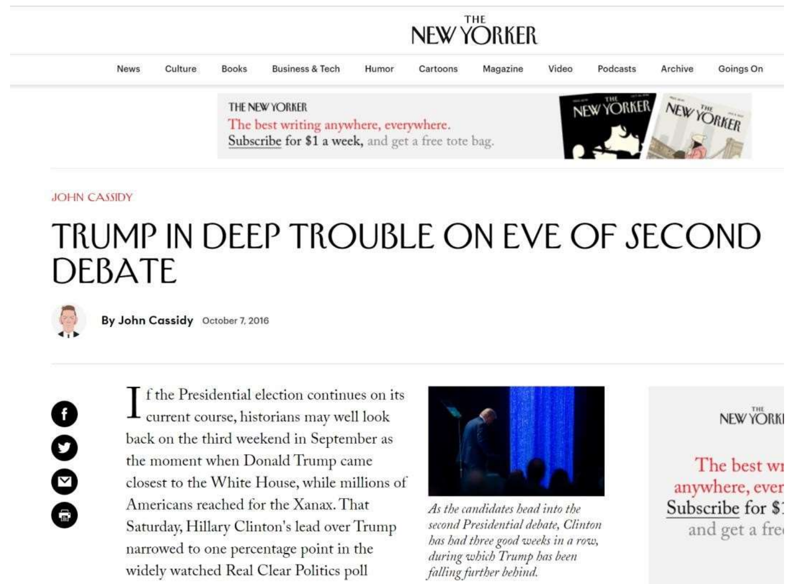
If the Presidential election continues on its current course, historians may well look back on the third weekend in September as the moment when Donald Trump came closest to the White House, while millions of Americans reached for the Xanax. That Saturday, Hillary Clinton's lead over Trump narrowed to one percentage point in the widely watched Real Clear Politics poll

As the candidates head into the second Presidential debate, Clinton has had three good weeks in a row, during which Trump has been falling further behind.

EXAMPLE: CONT.

We would like to classify documents as election related or not.

- Given a collection of documents with their labels (usually termed ‘training data’) we learn the parameters for our model.
- For example, if we see the word ‘Trump’ in *n1* out of the *n* documents labeled as ‘election’ we set *p(‘Trump’|’election’)=n1/n*
- Similarly we compute the priors (*p(‘election’)*) based on the proportion of the documents from both classes.


THE NEW YORKER

News Culture Books Business & Tech Humor Cartoons Magazine Video Podcasts Archive Goings On

THE NEW YORKER
The best writing anywhere, everywhere.
Subscribe for $1 a week, and get a free tote bag.

JOHN CASSIDY

TRUMP IN DEEP TROUBLE ON EVE OF SECOND DEBATE

By John Cassidy October 7, 2016

If the Presidential election continues on its current course, historians may well look back on the third weekend in September as the moment when Donald Trump came closest to the White House, while millions of Americans reached for the Xanax. That Saturday, Hillary Clinton's lead over Trump narrowed to one percentage point in the widely watched Real Clear Politics poll

As the candidates head into the second Presidential debate, Clinton has had three good weeks in a row, during which Trump has been falling further behind.

EXAMPLE: CLASSIFYING ELECTION (E) or Sports (S)

Assume we learned the following model

$P(\phi^{trump}=1 \mid E) = 0.8$, $P(\phi^{trumo}=1 \mid S) = 0.1$ $P(S) = 0.5$

$P(\phi^{russia}=1 \mid E) = 0.9$, $P(\phi^{russia}=1 \mid S) = 0.05$ $P(E) = 0.5$

$P(\phi^{clinton}=1 \mid E) = 0.9$, $P(\phi^{clinton}=1 \mid S) = 0.05$

$P(\phi^{football}=1 \mid E) = 0.1$, $P(\phi^{football}=1 \mid S) = 0.7$

Assume we have the following feature vector for a document:

$\phi^{trump} = 1,\ \phi^{russia} = 1,\ \phi^{clinton} = 1,\ \phi^{football} = 0$

$P(y = E \mid 1,1,1,0) \propto 0.8*0.9*0.9*0.9*0.5 = 0.5832$

$P(y = S \mid 1,1,1,0) \propto 0.1*0.05*0.05*0.3*0.5 = 0.000075$

So the document is classified as ‘Election’

NAÏVE BAYES CLASSIFIERS FOR CONTINUOUS VALUES

So far we assumed a binomial or discrete distribution for the data given the model ($p(X_i|y)$)

However, in many cases the data contains continuous features:

- Height, weight
- Levels of genes in cells
- Brain activity

For these types of data we often use a Gaussian model

In this model we assume that the observed input vector X is generated from the following distribution

$$X \sim N(\mu, \Sigma)$$

POSSIBLE PROBLEMS WITH NAÏVE BAYES CLASSIFIERS: ASSUMPTIONS

- In most cases, the assumption of conditional independence given the class label is violated

  - much more likely to find the word ‘Donald’ if we saw the word ‘Trump’ regardless of the class
- This is, unfortunately, a major shortcoming which makes these classifiers inferior in many real world applications
- There are models that can improve upon this assumption without using the full conditional model.

USING THE PRODUCT RULE

- **Applies to any number of variables:**
  - $P(a, b, c) = P(a, b|c)\, P(c) = P(a|b, c)\, P(b, c)$
  - $P(a, b, c|d, e) = P(a|b, c, d, e)\, P(b, c|d, e)$

- **Factoring**: (AKA **Chain Rule** for probabilities)
  - By the product rule, we can always write:

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b, c, \ldots z)$$

We often use comma to abbreviate AND.

  - Repeatedly applying this idea, we can write:

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b \mid c, .. z)\, P(c \mid .. z) .. P(z)$$

  - This holds for any ordering of the variables