Probabilistic Approach to AI

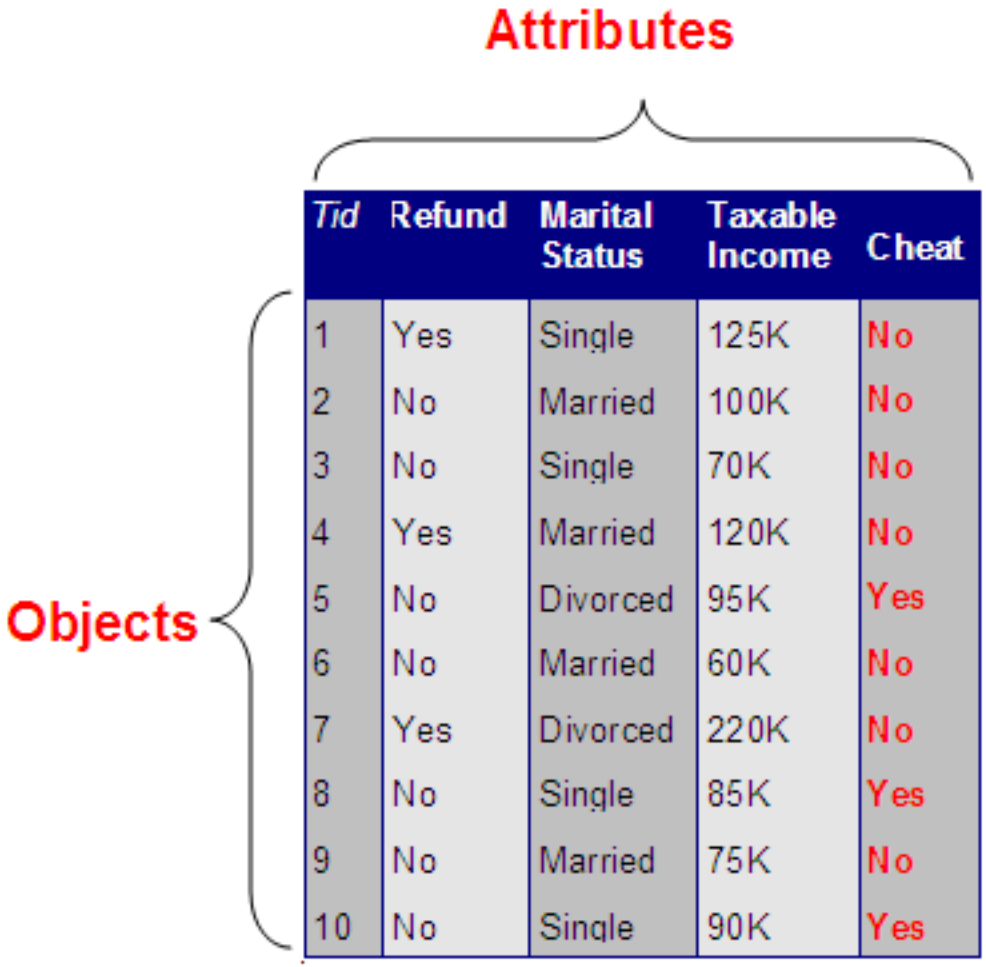
Chapter 13, Russell and Norvig

- Naïve Bayes algorithm
  - Maximum likelihood estimator
  - Nominal features
  - Laplacian correction
  - Real-valued features
- Bayesian networks

Data contains attributes and the class

Called training set

Decision to be made: Given the attributes of a new tax record, should the person be audited?



Another example – weather data

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Attribute	Rules	Errors	Total errors
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Windy	False → Yes	2/8	5/14
	True → No*	3/6	

* indicates a tie

ML Algorithms: The basic methods

- Neural networks
- Naïve Bayes, Bayesian networks
- Decision trees
- K-Nearest neighbors
- Support-vector machines

Outline

- Background
- Probability Basics
- Probabilistic Classification
- Naïve Bayes
- Examples
- Applying naïve Bayes to text classification
- Conclusions

Basic Axioms and Theorems of Probability

- Axioms:
  - 0 <= P(A) <= 1
  - P(True) = 1
  - P(False) = 0
  - P(A or B) = P(A) + P(B) – P(A and B)

- Theorems:
  - P(not A) = P(~A) = 1 – P(A)
  - P(A) = P(A ^ B) + P(A ^ ~B)

Some basic definitions

- Events
- Mutually exclusive events
- Conditional probability
- Independent events
- Laws of addition and multiplication

Some basic definitions

- Sample space, Events and Random Variables:

S = { (1,1), (1,2), … , (6, 6) }

R = { 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 }

$$p(R) = \{ \frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}, \frac{1}{36} \}$$

- Mutually exclusive events: {getting at least one 1}, {getting both even numbers }. Are they mutually exclusive?

- Independent events: {getting at least one even roll}, {sum of rolls is odd}

- Conditional probability: given the sum of the rolls is even, what is the probability of getting a 3?

Deriving the Bayes Rule

Conditional Probability:

Chain rule:

Bayes Rule:

Another version of Bayes' theorem:

BAYES' THEOREM

Let $A_1, \ldots, A_k$ be a collection of mutually exclusive and exhaustive events with $P(A_i) > 0$ for $i = 1, \ldots, k$. Then for any other event B for which $P(B) > 0$,

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P(A_i)} \qquad j = 1, \ldots, k \quad (1.5)$$

Example 1.33 *Incidence of a rare disease*. In the book's Introduction, we presented the following example as a common misunderstanding of probability in everyday life. Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time. If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

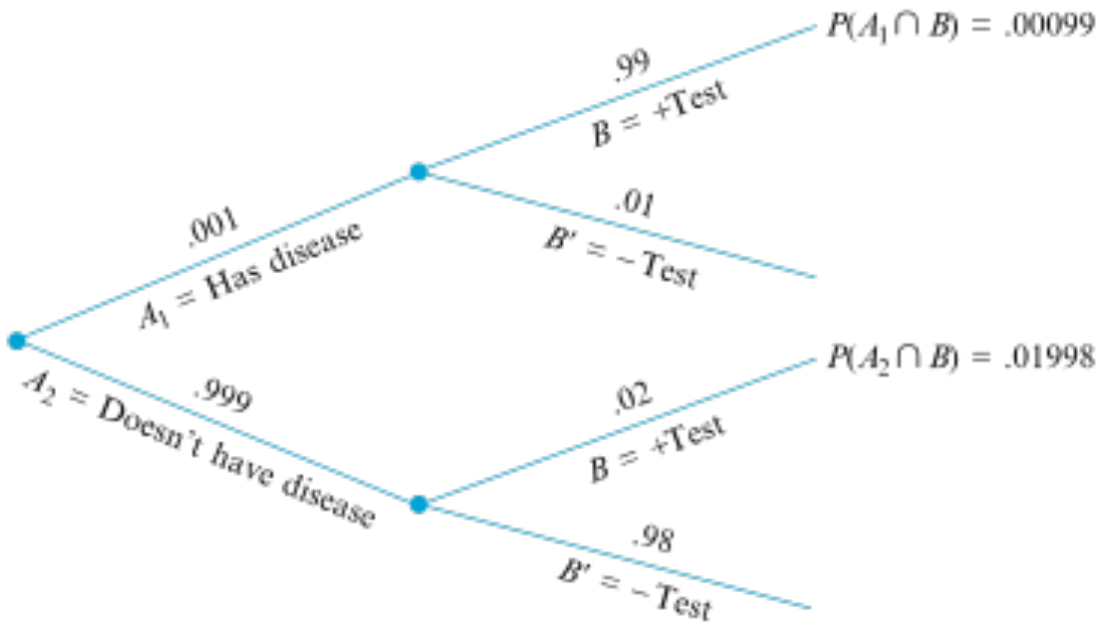


Fig. 1.13 Tree diagram for the rare-disease problem

Next to each branch corresponding to a positive test result, the Multiplication Rule yields the recorded probabilities. Therefore, $P(B) = .00099 + .01998 = .02097$, from which we have

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.00099}{.02097} = .047$$

Boxes and Balls with Bayes Rule

Assuming I'm inherently more likely to select the red box (66.6%) than the blue box (33.3%).

If I selected an orange ball, what is the likelihood that I selected

the red box?

the blue box?

Boxes and Balls

$$p(B = r|L = o) = \frac{p(L = o|B = r)p(B = r)}{p(L = o)}$$

$$= \frac{\frac{6}{8}\frac{2}{3}}{\frac{7}{12}} = \frac{6}{7}$$

$$p(B = b|L = o) = \frac{p(L = o|B = b)p(B = b)}{p(L = o)}$$

$$= \frac{\frac{1}{4}\frac{1}{3}}{\frac{7}{12}} = \frac{1}{7}$$

Bayesian Classifiers

- Consider each attribute and class label as random variables

- Given a record with attributes (A_1, A_2,…,A_n)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)

- Can we estimate P(C| A_1, A_2,…,A_n) directly from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$

  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

- Assume conditional independence among attributes A_i (conditional on the class):
  - $P(A_1, A_2, \ldots, A_n | C_j) = P(A_1 | C_j)\, P(A_2 | C_j) \ldots P(A_n | C_j)$
  - Can estimate $P(A_i | C_j)$ for all A_i and C_j ?
  - New point is classified to C_j if $P(C_j) \prod P(A_i | C_j)$ is maximal.

How to Estimate Probabilities from Data?

- Class: $P(C) = N_c/N$
  - e.g., P(No) = 7/10,
    P(Yes) = 3/10

- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_c$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k

Examples:

  - P(Status=Married|No) =

Most likely class – dropping denominator P(d)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

Example of Naïve Bayes Classifier

A: attributes

M: mammals

N: non-mammals

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Taking the log

- Multiplying lots of small probabilities can result in floating point underflow.
- Since $\log(xy) = \log(x) + \log(y)$, we can sum log probabilities instead of multiplying probabilities.
- Since log is a monotonic function, the class with the highest score does not change.
- So what we usually compute in practice is:

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

21

How to Estimate Probabilities from Data?

- Normal distribution:

  - One for each (A_i , c_j) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

Example of Naïve Bayes Classifier

Given a Test Record:

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K|
  Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K|
  Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

Naïve Bayes Classifier – Laplace correction

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

c: number of classes

p: prior probability

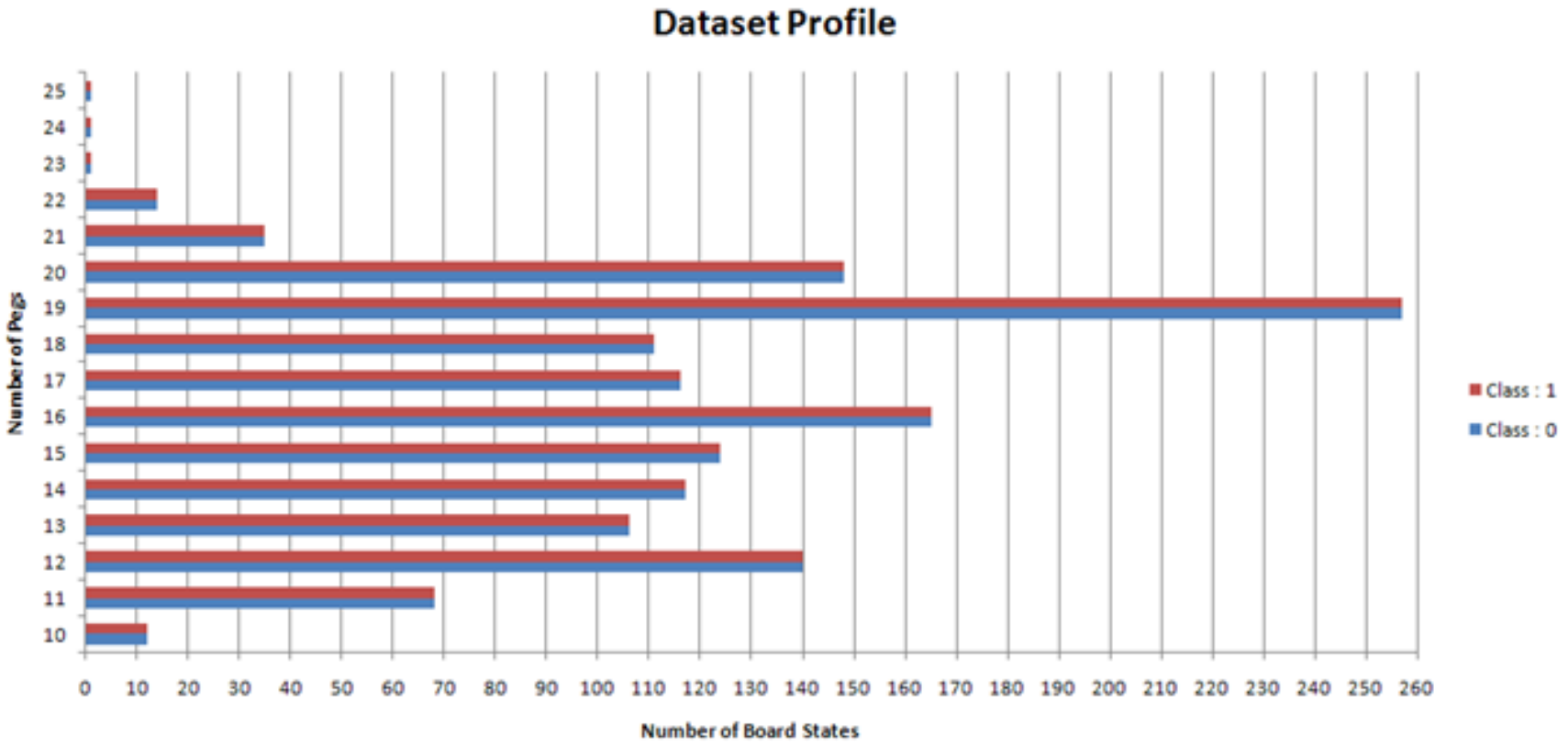
m: parameter

Naïve Bayes classifier – a case study

Peg-solitaire: standard board and our board

8	7	6	5	4	3	2	1
16	15	14	13	12	11	10	9
24	23	22	21	20	19	18	17
32	31	30	29	28	27	26	25
40	39	38	37	36	35	34	33
48	47	46	45	44	43	42	41
56	55	54	53	52	51	50	49
64	63	62	61	60	59	58	57

Can we train a classifier to predict solvable positions?



Feature selection

A1. Number of pegs (pegs).

A2. Number of first moves for any peg on the board (first_moves).

A3. Number of rows having 4 pegs separated by single vacant positions (ideal_row).

A4. Number of columns having 4 pegs separated by single vacant positions (ideal_col).

A5. Number of the first two moves for any peg on the board (first_two).

A6. Percentage of the total number of pegs in quadrant one (quad_one).

A7. Percentage of the total number of pegs in quadrant two (quad_two).

A8. Percentage of the total number of pegs in quadrant three (quad_three).

A9. Percentage of the total number of pegs in quadrant four (quad_four).

A10. Number of pegs isolated by one vacant position (island_one).

A11. Number of pegs isolated by two vacant positions (island_two).

A12. Number of rows having 3 pegs separated by single vacant positions (ideal_row_three).

A13. Number of columns having 3 pegs separated by single vacant positions (ideal_col_three).

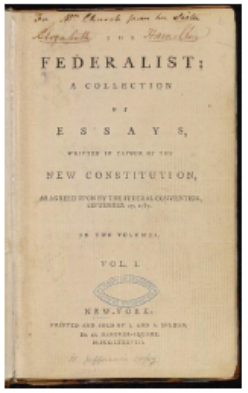
Success rate for various feature sets

Attribute[1]	Profile 1	Profile 2	Profile 3	Profile 4	Profile 5
A1. pegs	x	x	x	x	x
A2. first_moves	x	x	x		x
A3. ideal_row	x				
A4. ideal_col	x				
A5. first_two	x	x	x		x
A6. quad_one	x	x	x	x	
A7. quad_two	x	x	x	x	
A8. quad_three	x	x	x	x	
A9. quad_four	x	x	x	x	
A10. island_one	x	x	x		x
A11. island_two	x	x	x		x
A12. ideal_row_three	x	x			
A13. ideal_col_three	x	x			
A14. c^2	x	x	x	x	x
Naïve Bayes					
% Split[3]	66	66	66	66	66
% Correct	93.5618	93.5618	93.9772	78.92	90.0312
% Split			80		
% Correct			94.0035		
% Split			90		
% Correct			94.7183		
% Split			95		
% Correct			96.4789		

Bayesian classifier in authorship resolution

Dan Jurafsky

Who wrote which Federalist papers?



- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods

James Madison

Alexander Hamilton

Dan Jurafsky

Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

Classifying movie review

Dan Jurafsky

Positive or negative movie review?

- unbelievably disappointing

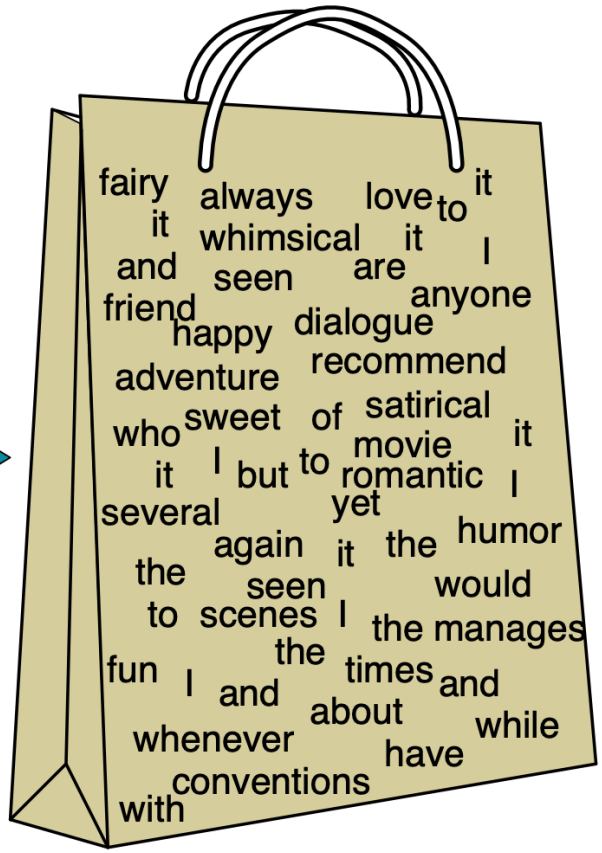
- Full of zany characters and richly applied satire, and some great plot twists

- this is the greatest screwball comedy ever filmed

- It was pathetic. The worst part about it was the boxing scenes.

Bag of words model

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Bag of words model

- Document is treated as a collection of words
- Each word is treated as a feature
- The number of occurrences (frequency) is the value associated with the word. (known as the multinomial model)
- (In binary model, simply 1 or 0 is used for occurrence or not.)
- Two key simplifications: order is ignored, dependencies between words is ignored (Naïve)
- Can use selected subset, instead of all the words

- From training corpus, extract *Vocabulary*

- Calculate $P(c_j)$ terms
  - For each c_j in C do

    $docs_j \leftarrow$ all docs with class $= c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Suppose n_1, n_2, ..., n_k is the number of times word i occurs in the document, and P_1, P_2, ..., P_k is the probability of obtaining word i when sampling from all the documents in category H. Assume that the probability is independent of the word's context and position in the document. These assumptions lead to a *multinomial distribution* for document probabilities. For this distribution, the probability of a document E given its class H—in other words, the formula for computing the probability $\Pr[E \mid H]$ in Bayes' rule—is

$$\Pr[E \mid H] = N! \times \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_i!}$$

where $N = n_1 + n_2 + \ldots + n_k$ is the number of words in the document. The reason for the factorials is to account for the fact that the ordering of the occurrences of each word is immaterial according to the bag-of-words model. P_i is estimated by computing the relative frequency of word i in the text of all training documents pertaining to category H. In reality, there could be a further term that gives the probability that the model for category H generates a document whose length is the same as the length of E, but it is common to assume that this is the same for all classes and hence can be dropped.

Naive Bayes: Training

function TRAIN NAIVE BAYES(D, C) **returns** log $P(c)$ and log $P(w|c)$

for each class $c \in C$ # Calculate $P(c)$ terms
  N_{doc} = number of documents in D
  N_c = number of documents from D in class c
  $logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$
  $V \leftarrow$ vocabulary of D
  $bigdoc[c] \leftarrow$ **append**(d) **for** d $\in$ D **with** class c
  for each word w in V # Calculate $P(w|c)$ terms
    $count(w,c) \leftarrow$ # of occurrences of w in $bigdoc[c]$
    $loglikelihood[w,c] \leftarrow \log \frac{count(w,c) + 1}{\sum_{w' \text{ in } V} (count(w',c) + 1)}$
return *logprior*, *loglikelihood*, V

Naive Bayes: Testing

function TEST NAIVE BAYES(*testdoc*, *logprior*, *loglikelihood*, C, V) **returns** best *c*

for each class $c \in C$
  $sum[c] \leftarrow logprior[c]$
  for each position *i* in *testdoc*
    $word \leftarrow testdoc[i]$
    if $word \in V$
      $sum[c] \leftarrow sum[c] + loglikelihood[word,c]$
return $\text{argmax}_c\ sum[c]$

37

Exercise

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

- Estimate parameters of Naive Bayes classifier
- Classify test document

38

Example: Parameter estimates

Priors: $\hat{P}(c) = 3/4$ and $\hat{P}(\overline{c}) = 1/4$ Conditional probabilities:

$$\begin{aligned}
\hat{P}(\textsc{Chinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\textsc{Tokyo}|c) = \hat{P}(\textsc{Japan}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\textsc{Chinese}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\textsc{Tokyo}|\overline{c}) = \hat{P}(\textsc{Japan}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

The denominators are (8 + 6) and (3 + 6) because the lengths of $text_c$ and $text_{\overline{c}}$ are 8 and 3, respectively, and because the constant B is 6 as the vocabulary consists of six terms.

39

Example: Classification

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$
$$\hat{P}(\overline{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

Thus, the classifier assigns the test document to c = *China*. The reason for this classification decision is that the three occurrences of the positive indicator `CHINESE` in d_5 outweigh the occurrences of the two negative indicators `JAPAN` and `TOKYO`.

40

Naive Bayes is not so naive

- More robust to nonrelevant features than some more complex learning methods
- More robust to concept drift (changing of definition of class over time) than some more complex learning methods
- Better than methods like decision trees when we have many equally important features
- A good dependable baseline for text classification (but not the best)
- Optimal if independence assumptions hold (never true for text, but true for some domains)
- Very fast
- Low storage requirements

Naïve Bayes (Summary)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Networks