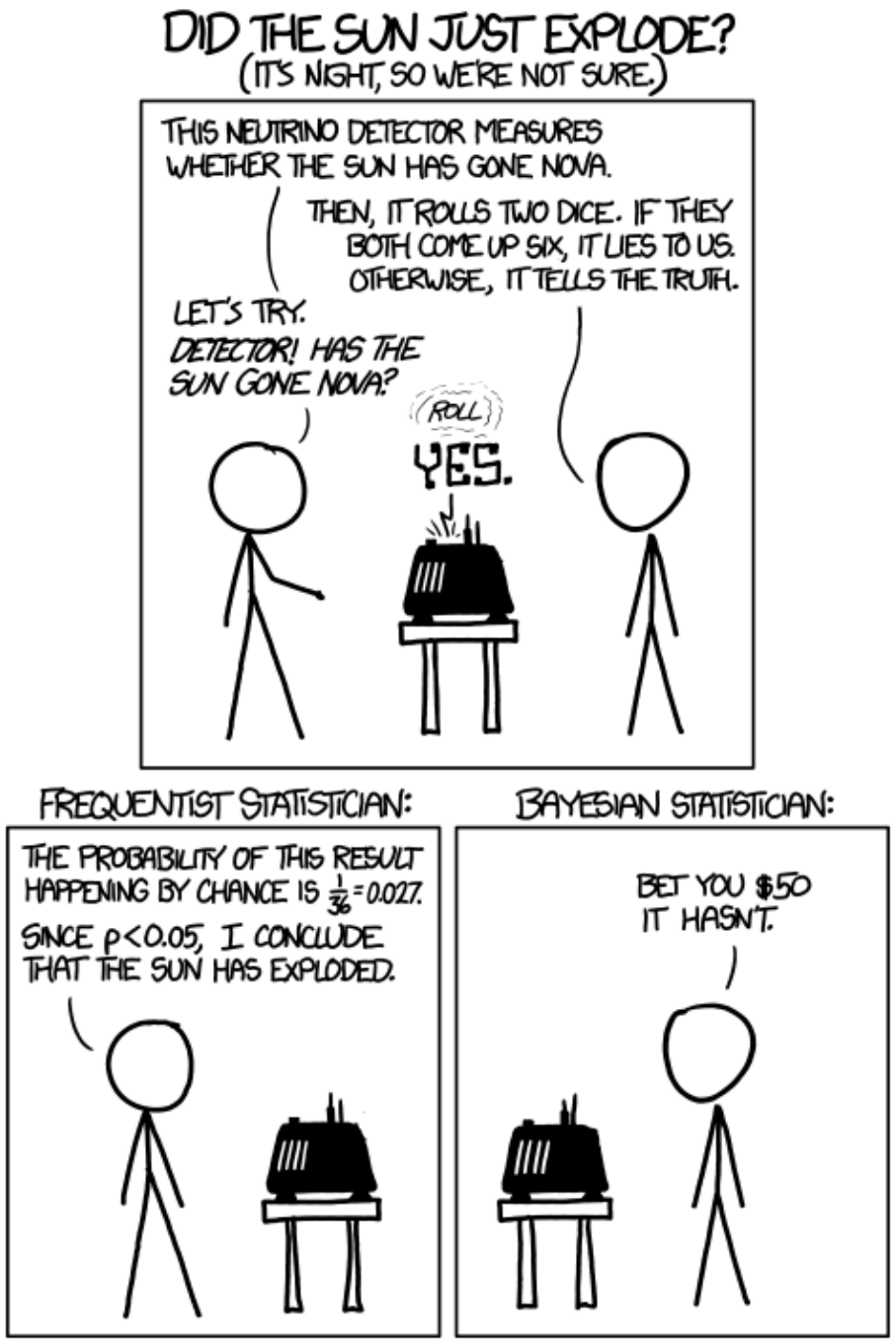
CS440/ECE448 Lecture 14: Naïve Bayes

Mark Hasegawa-Johnson, 2/2020

Including slides by Svetlana Lazebnik, 9/2016



https://www.xkcd.com/1132/

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Maximum A Posteriori estimation of parameters
  - Laplace Smoothing

Bayes' Rule

Rev. Thomas Bayes
(1702-1761)

By Unknown - [2][3], Public Domain, https://commons.wikimedia.org/w/index.php?curid=14532025

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

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

- Therefore,

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

- Why is this useful?
  - "A" is something we care about, but P(A|B) is really really hard to measure (example: the sun exploded)
  - "B" is something less interesting, but P(B|A) is easy to measure (example: the amount of light falling on a solar cell)
  - Bayes' rule tells us how to compute the probability we want (P(A|B)) from probabilities that are much, much easier to measure (P(B|A)).

Bayes Rule example

Eliot & Karson are getting married tomorrow, at an outdoor ceremony in the desert. Unfortunately, the weatherman has predicted rain for tomorrow.

- In recent years, it has rained (event R) only 5 days each year (5/365 = 0.014).

$$P(R) = 0.014$$

- When it actually rains, the weatherman forecasts rain (event F) 90% of the time.

$$P(F|R) = 0.9$$

- When it doesn't rain, he forecasts rain (event F) only 10% of the time.

$$P(F|\neg R) = 0.1$$

- What is the probability that it will rain on Eliot's wedding?

$$P(R|F) = \frac{P(F|R)P(R)}{P(F)} = \frac{P(F,R)P(R)}{P(F,R) + P(F,\neg R)} = \frac{P(F|R)P(R)}{P(F|R)P(R) + P(F|\neg R)P(\neg R)}$$

$$= \frac{(0.9)(0.014)}{(0.9)(0.014) + (0.1)(0.956)} = 0.116$$

The More Useful Version of Bayes' Rule

Rev. Thomas Bayes (1702-1761)

By Unknown - [2][3], Public Domain, https://commons.wikimedia.org/w/index.php?curid=14532025

This version is what you memorize.

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

- Remember, P(B|A) is easy to measure (the probability that light hits our solar cell, if the sun still exists and it's daytime). Let's assume we also know P(A) (the probability the sun still exists).
- But suppose we don't really know P(B) (what is the probability light hits our solar cell, if we don't really know whether the sun still exists or not?)

This version is what you actually use.

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

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Maximum A Posteriori estimation of parameters
  - Laplace Smoothing

The Misdiagnosis Problem

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?

$$P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})}$$

$$= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive} \mid \text{cancer})P(\text{cancer}) + P(\text{positive} \mid \neg\text{cancer})P(\neg\text{Cancer})}$$

$$= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.096 \times 0.99} = \frac{0.008}{0.008 + 0.095} = 0.0776$$

Second Opinions

If your doctor tells you that you have a health problem or suggests a treatment for an illness or injury, you might want a second opinion. This is especially true when you're considering surgery or major procedures.

Asking another doctor to review your case can be useful for many reasons:

HEALTH INSURANCE AND MEDICARE HOME

News
Reference
Quizzes
Videos
Message Boards
Find a Doctor

RELATED TO HEALTH

TODAY ON WEBMD

Clinical Trials
What qualifies you for one?

Working During Cancer Treatment
Know your benefits.

Going to the Dentist?
How to save money.

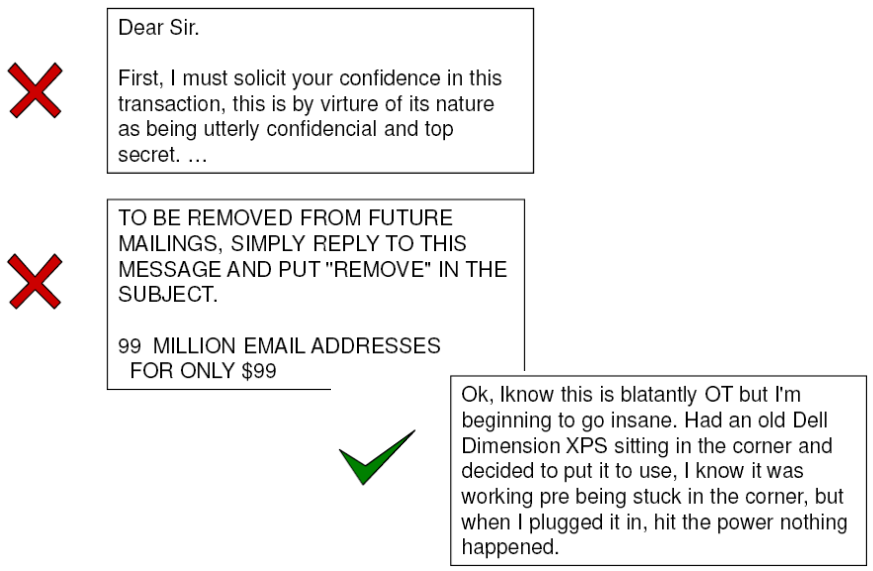
Enrolling in Medicare
How to get started.

The Bayesian Decision

The agent is given some evidence, E.

The agent has to make a decision about the value of an unobserved variable Y. Y is called the "query variable" or the "class variable" or the "category."

- Partially observable, stochastic, episodic environment
- Example: $Y \in$ {spam, not spam}, $E =$ email message.
- Example: $Y \in$ {zebra, giraffe, hippo}, $E =$ image features



Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Maximum A Posteriori estimation of parameters
  - Laplace Smoothing

Classification using probabilities

- Suppose you know that you have a toothache.
- Should you conclude that you have a cavity?
- Goal: make a decision that **minimizes your probability of error**.
- Equivalent: make a decision that **maximizes the probability of being correct**. This is called a MAP (maximum a posteriori) decision. You decide that you have a cavity if and only if

$$P(Cavity|Toothache) > P(\neg Cavity|Toothache)$$

Bayesian Decisions

- What if we don't know $P(Cavity|Toothache)$? Instead, we only know $P(Toothache|Cavity)$, $P(Cavity)$, and $P(Toothache)$?
- Then we choose to believe we have a Cavity if and only if

$$P(Cavity|Toothache) > P(\neg Cavity|Toothache)$$

Which can be re-written as

$$\frac{P(Toothache|Cavity)P(Cavity)}{P(Toothache)} > \frac{P(Toothache|\neg Cavity)P(\neg Cavity)}{P(Toothache)}$$

MAP decision

The action, “a”, should be the value of C that has the highest posterior probability given the observation E=e:

$$a = \text{argmax} P(Y = a|E = e) = \text{argmax} \frac{P(E = e|Y = a)P(Y = a)}{P(E = e)}$$
$$= \text{argmax} P(E = e|Y = a)P(Y = a)$$

$$P(Y = a|E = e) \propto P(E = e|Y = a)P(Y = a)$$

posterior likelihood prior

The Bayesian Terms

- $P(Y = y)$ is called the "**prior**" (*a priori*, in Latin) because it represents your belief about the query variable before you see any observation.
- $P(Y = y|E = e)$ is called the "**posterior**" (*a posteriori*, in Latin), because it represents your belief about the query variable after you see the observation.
- $P(E = e|Y = y)$ is called the "**likelihood**" because it tells you how much the observation, E=e, is like the observations you expect if Y=y.
- $P(E = e)$ is called the "**evidence distribution**" because E is the evidence variable, and $P(E = e)$ is its marginal distribution.

$$P(y|e) = \frac{P(e|y)P(y)}{P(e)}$$

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Maximum A Posteriori estimation of parameters
  - Laplace Smoothing

Naïve Bayes model

- Suppose we have many different types of observations (symptoms, features) $X_1, \ldots, X_n$ that we want to use to obtain evidence about an underlying hypothesis C
- MAP decision:

$$P(Y = y | E_1 = e_1, \ldots, E_n = e_n) \propto P(Y = y) P(E_1 = e_1, \ldots, E_n = e_n | Y = y)$$

- If each feature E_i can take on k values, how many entries are in the probability table $P(E_1 = e_1, \ldots, E_n = e_n | Y = y)$?

Naïve Bayes model

Suppose we have many different types of observations (symptoms, features) $E_1, \ldots, E_n$ that we want to use to obtain evidence about an underlying hypothesis Y

The Naïve Bayes decision:

$$a = \text{argmax}\, p(Y = a | E_1 = e_1, \ldots, E_n = e_n)$$

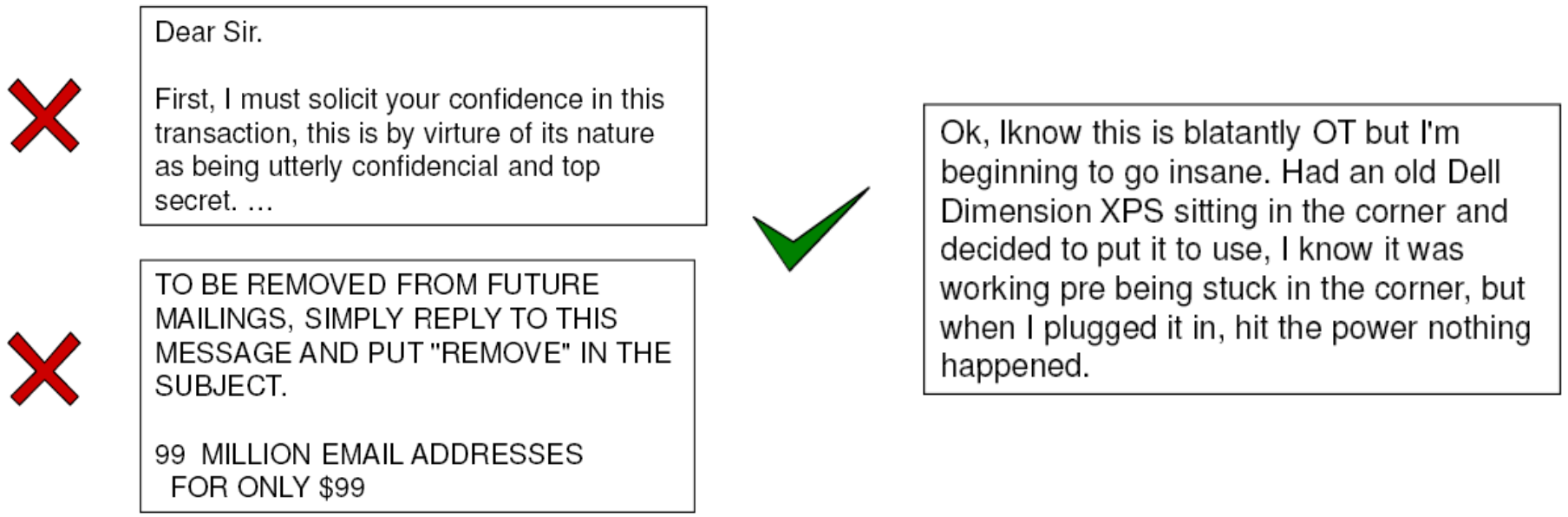
$$= \text{argmax}\, p(Y = a) p(E_1 = e_1, \ldots, E_n = e_n | Y = a)$$

$$\approx \text{argmax}\, p(Y = a) p(E_1 = e_1 | Y = a) \ldots p(E_n = e_n | Y = a)$$

Case study:
Text document classification

- **MAP decision:** assign a document to the class with the highest posterior P(class | document)

- Example: spam classification
  - Classify a message as spam if P(spam | message) > P(¬spam | message)



Case study:
Text document classification

- **MAP decision:** assign a document to the class with the highest posterior P(class | document)

- We have P(class | document) $\propto$ P(document | class)P(class)

- To enable classification, we need to be able to estimate the likelihoods P(document | class) for all classes and
priors P(class)

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Maximum A Posteriori estimation of parameters
  - Laplace Smoothing

Naïve Bayes Representation

- Goal: estimate likelihoods P(document | class) and priors P(class)
- Likelihood: ***bag of words*** representation
  - The document is a sequence of words (w_1, …, 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 virture of its nature as being utterly confidencial 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, Iknow 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(document | class) and priors P(class)
- Likelihood: ***bag of words*** representation
  - The document is a sequence of words ($E_1 = w_1, \ldots, E_n = w_n$)
  - The order of the words in the document is not important
  - Each word is conditionally independent of the others given document class

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

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

Parameter estimation

- Model parameters: feature likelihoods p(word | class) and priors p(class)
  - How do we obtain the values of these parameters?

prior

spam:	0.33
¬spam:	0.67

P(word | 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(word | ¬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
...	

Bag of words illustration

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

Bag of words illustration

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

Bag of words illustration

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Maximum A Posteriori estimation of parameters
  - Laplace Smoothing

Bag of words representation of a document

Consider the following movie review (8114_3.txt in your dataset):

I'm warning you, it's pretty pathetic

What is its BoW representation?

Word	# times it occurs
I'm	1
it's	1
pathetic	1
pretty	1
warning	1
you	1
(every other word)	0

- "pathetic" probably means it's a negative review.
- …but "pretty" means it's a positive review, right?

Bigram representation of a document

A “bigram” is just a pair of words that occur together, in sequence. For example, the following review has the bigrams shown at left:

I’m warning you, it’s pretty pathetic

Word	# times it occurs
I’m warning	1
it’s pretty	1
pretty pathetic	1
warning you	1
you it’s	1
(every other bigram)	0

{ ”I’m warning”, “warning you”, “pretty pathetic” } == negative

{ “it’s pretty” } == positive, but maybe we can ignore that.

Naïve Bayes with Bigrams

- Goal: estimate likelihoods P(document | class) and priors P(class)
- Likelihood: ***bigrams*** representation
  - The document is a sequence of bigrams ($E_1 = b_1, \ldots, E_n = b_n$)
  - The order of the bigrams in the document is not important
  - Each bigram is conditionally independent of the others given document class

$$P(document|class) = P(b_1, \ldots, b_n|class) = \prod_{i=1}^{n} P(b_i|class)$$

  - Thus, the problem is reduced to estimating marginal likelihoods of individual words $p(b_i \mid class)$

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Laplace Smoothing

Bayesian Learning

- Model parameters: feature likelihoods P(word | class) and priors P(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. It is the estimate that maximizes the probability of the training data, which is defined as:

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

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

Bayesian Learning

The data likelihood

$$\mathrm{P}(training\ data) = \prod_{d=1}^{D} \prod_{i=1}^{\#\ words\ in\ d} P(E = w_i | Y = c_d)$$

is maximized (subject to the constraint that sum_w p(w|c)=1) if we choose:

$$P(E = w | Y = c) = \frac{\#\ \text{occurrences of word } w \text{ in documents of type } c}{\text{total number of words in all documents of type } c}$$

$$P(Y = c) = \frac{\#\ \text{documents of type } c}{\text{total number of documents}}$$

Bayesian Learning

The data likelihood

$$\mathrm{P}(training\ data) = \prod_{d=1}^{D} \prod_{i=1}^{\#\ unique\ words\ in\ d} P(E = w_i | Y = c_d)$$

is maximized (subject to the constraint that sum_w p(w|c)=1) if we choose:

$$P(E = w | Y = c) = \frac{\#\ \text{documents of type } c \text{ containing word } w}{\text{total number of documents of type } c}$$

$$P(Y = c) = \frac{\#\ \text{documents of type } c}{\text{total number of documents}}$$

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
  - Bigrams
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Laplace Smoothing

What is the probability that the sun will fail to rise tomorrow?

- # times we have observed the sun to rise = 100,000,000
- # times we have observed the sun not to rise = 0
- Estimated probability the sun will not rise = $\frac{0}{0+100,000,000} = 0$

Oops….

Laplace Smoothing

- The basic idea: add 1 “unobserved observation” to every possible event
- # times the sun has risen or might have ever risen = 100,000,000+1 = 100,000,001
- # times the sun has failed to rise or might have ever failed to rise = 0+1 = 1
- Estimated probability the sun will not rise = $\frac{1}{1+100,000,001} = 0.000000009999998$

Parameter estimation

- ML (Maximum Likelihood) parameter estimate:

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

- Laplacian Smoothing estimate
  - How can you estimate the probability of a word you never saw in the training set? (Hint: what happens if you give it probability 0, then it actually occurs in a test document?)
  - **Laplacian smoothing:** pretend you have seen every vocabulary word one more time than you actually did

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

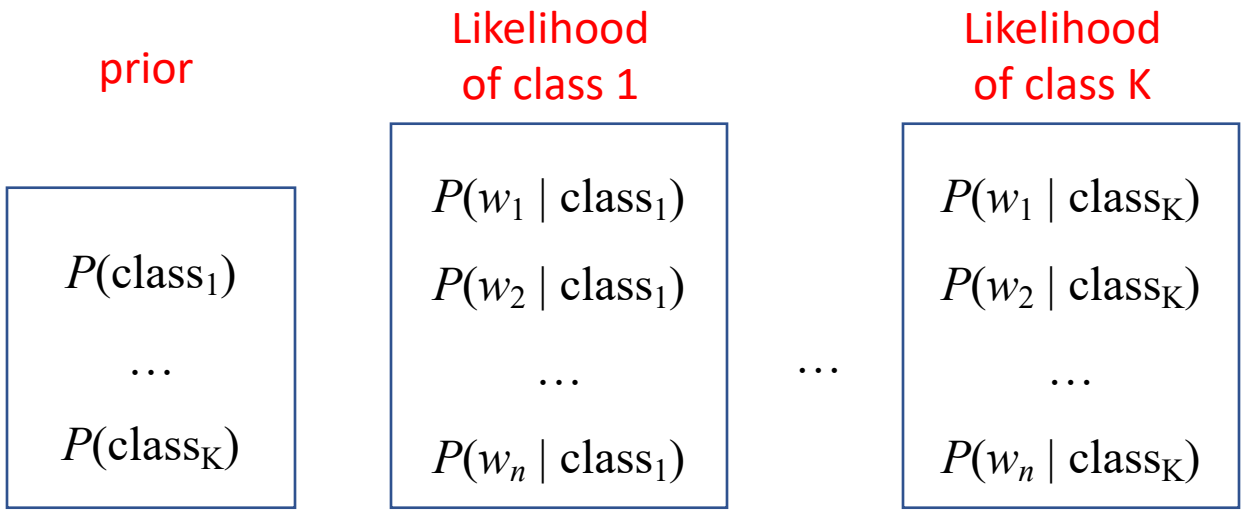
(V: total number of unique words)

Summary: Naïve Bayes for Document Classification

- Naïve Bayes model: assign the document to the class with the highest posterior

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

- Model parameters:



Review: Bayesian decision making

- Suppose the agent has to make decisions about the value of an unobserved *query variable* Y based on the values of an observed *evidence variable* E
- **Inference problem:** given some observation $E = e$, what is $P(Y \mid E=e)$?
- **Learning problem:** estimate the parameters of the probabilistic model $P(y \mid e)$ given a *training sample* $\{(e_1,y_1), \ldots, (e_n,y_n)\}$