CS 471: Introduction to AI

Module 5: Quantifying Uncertainty

Acting Under Uncertainty

Agents in the real world need to handle uncertainty, may be due to partial observability

Eg: An automated taxi has the goal of delivering a passenger to the airport on time.

The taxi forms a plan, A90, that involves leaving home 90 minutes before the flight departs and driving at a reasonable speed.

Even though the airport is only 5 miles away, a logical agent will not be able to conclude with absolute certainty that “Plan A90 will get us to the airport in time.”

Instead, it reaches the weaker conclusion “Plan A90 will get us to the airport in time, as long as the car doesn’t break down, and I don’t get into an accident, and the road isn’t closed, and . . . .”

Example: Uncertainty

Diagnosing a dental patient’s toothache.

- Diagnosis almost always involves uncertainty. Consider the following simple rule:

  Toothache ⇒ Cavity

- The problem is that this rule is wrong. Not all patients with toothaches have cavities; some of them have gum disease, or one of several other problems: We could try turning the rule:

  Cavity ⇒ Toothache

  But this rule is not right either; not all cavities cause pain.

The only way to fix the rule is to make it logically exhaustive: to augment the left-hand side with all the qualifications required for a cavity to cause a toothache.

Example: Uncertainty

Using logic for medical diagnosis fails for three main reasons:

- Laziness: It is too much work to list the complete set of possibilities.
- Theoretical ignorance: Medical science has no complete theory for the domain.
- Practical ignorance: Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.

Example: Uncertainty

- The connection between toothaches and cavities is not a strict logical consequence in either direction.
- This is true for other judgmental domains too
- Tool for dealing with degree of belief is probability theory

Logical agent => true or false or has no opinion Probabilistic agent => degree of belief between 0 (certainly false) and 1 (certainly true).

Eg: We might say that there is an 80% chance, a probability of 0.8 that the patient who has a toothache has a cavity. This belief could be derived from statistical data: 80% of the toothache patients seen so far have had cavities, or from some general dental knowledge, or from a combination of evidence sources.

Probability

- Sample space: The set of all possible worlds
- Eg: if we roll two dice, sample space has 36 possible outcomes: (1,1), (1,2), . . ., (6,6).
- Event: Any possible outcome
- The Greek letter Ω (uppercase omega) => sample space ω (lowercase omega) => elements of the space
- Every possible outcome has a probability between 0 and 1 and that the total probability of the set of possible worlds is 1:

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} P(\omega) = 1\,.$$

- Eg: For two dice problem, probability of each possible outcome (1,1), (1,2), . . ., (6,6) is 1/36.

Probability

- We typically care about sets of outcomes. When rolling fair dice, what is the probability that the two dice add up to 11?

Probabilities such as P(Total=11) and P(doubles) are called unconditional or prior probabilities => degrees of belief in the absence of any other information.

Probability

- Most of the time, we have some information that has already been revealed, called evidence.
- Eg: The first die may already be showing a 5 and we are waiting for the other one to stop spinning.

  We are interested in the conditional or posterior probability (or just “posterior”) of rolling doubles given that the first dice is a 5.

  P(doubles | Dice_1= 5); “|” is pronounced “given”.

Probability Basics

- Going to the dentist for a regularly scheduled checkup,

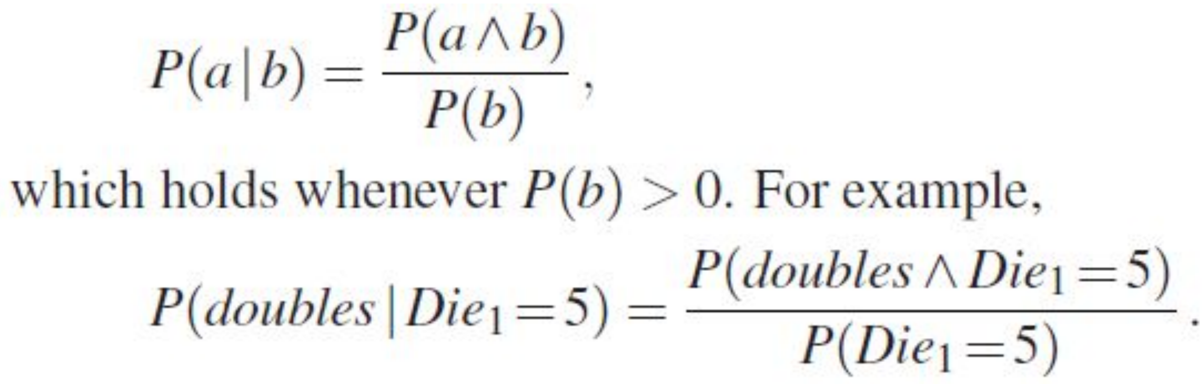
  then the prior probability => P(cavity) = 0.2

  Going to the dentist because I have a toothache,

  then the conditional probability P(cavity | toothache) = 0.6

Probability Basics

- Conditional probabilities are defined in terms of unconditional probabilities as: for any events a and b,



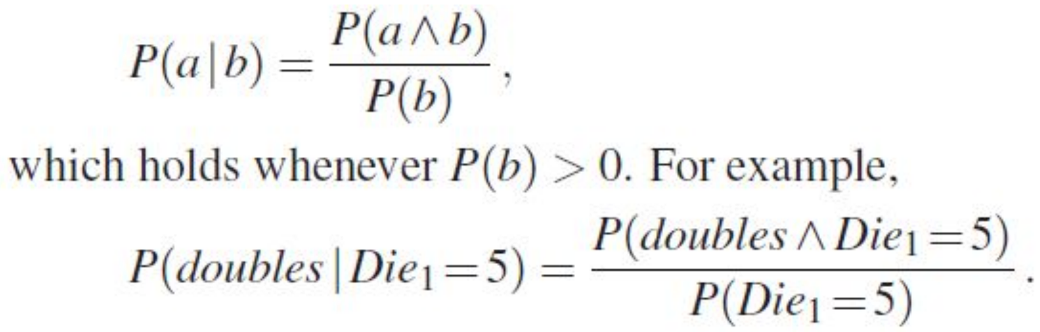
$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)},$$

which holds whenever $P(b) > 0$. For example,

$$P(doubles \mid Die_1 = 5) = \frac{P(doubles \wedge Die_1 = 5)}{P(Die_1 = 5)}.$$

Observing b rules out all those possible worlds where b is false, leaving a set whose total probability is just P(b).

Activity



$$P(a \mid b) = \frac{P(a \land b)}{P(b)},$$

which holds whenever $P(b) > 0$. For example,

$$P(doubles \mid Die_1 = 5) = \frac{P(doubles \land Die_1 = 5)}{P(Die_1 = 5)}.$$

Solve for P(doubles | Die$_1$ = 5)

Probability Basics

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

Conditional probability can be written in a different form called the product rule:

$$P(a \wedge b) = P(a|b)P(b)$$

For a and b to be true, we need b to be true, and we also need a to be true given b.

Random Variables

- Random variable: represents an event whose outcome is unknown
- Probability distribution: is an assignment of weights to outcomes
- Example: Traffic on freeway
  - Random variable: T = whether there’s traffic
  - Outcomes: T in {none, light, heavy}
  - Distribution: P(T=none) = 0.25, P(T=light) = 0.5, P(T=heavy) = 0.25

Representations

Probability distribution on multiple variables:

P(Weather=sun) = 0.6

P(Weather=rain) = 0.1

P(Weather=cloud) = 0.29

P(Weather=snow) = 0.01

P(Weather, Cavity) denotes the probabilities of all combinations of the values of Weather and Cavity. This is a 4×2 table of probabilities called the joint probability distribution of Weather and Cavity.

Why 4x2 table and how does it look like?

Probability Axioms

- Basic axioms of probability:

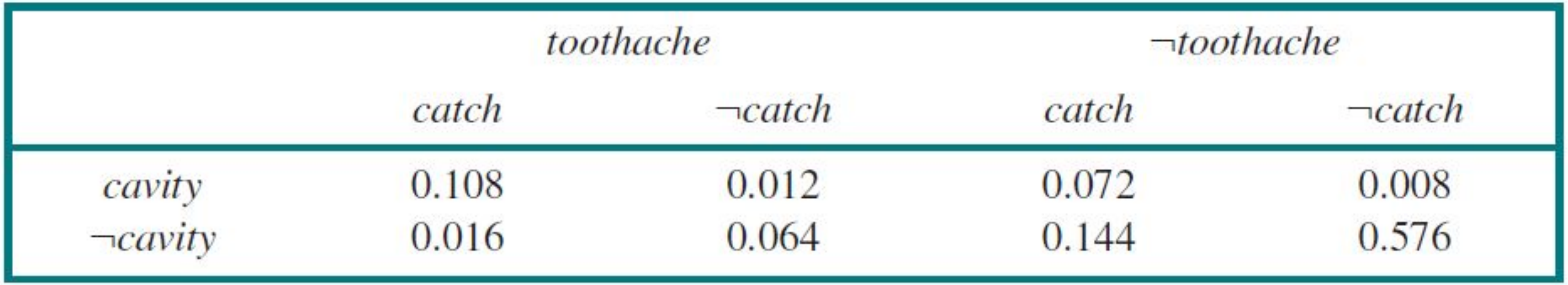
$$
\begin{aligned}
P(\neg a) &= \textstyle\sum_{\omega \in \neg a} P(\omega) \\
&= \textstyle\sum_{\omega \in \neg a} P(\omega) + \sum_{\omega \in a} P(\omega) - \sum_{\omega \in a} P(\omega) \\
&= \textstyle\sum_{\omega \in \Omega} P(\omega) - \sum_{\omega \in a} P(\omega) \\
&= 1 - P(a)
\end{aligned}
$$

- Inclusion-exclusion principle:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)\,.$$

Inference using Full Joint Distributions

- Eg: A domain consisting of three Boolean variables Toothache, Cavity, and Catch. The full joint distribution is a 2×2×2 table.



	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Probabilities in the joint distribution must sum to 1.

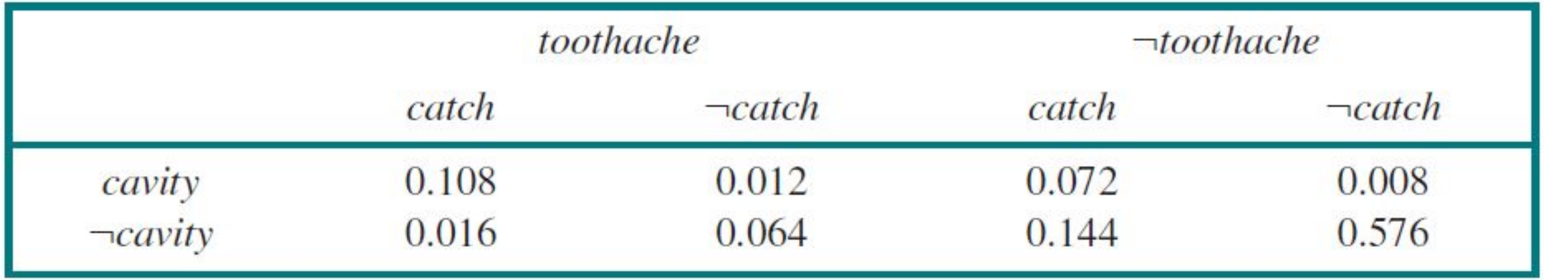
Inference using Full Joint Distributions

- Common task is to compute the distribution of a single variable.

What is the unconditional or marginal probability of cavity?

$$P(\text{cavity}) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2.$$

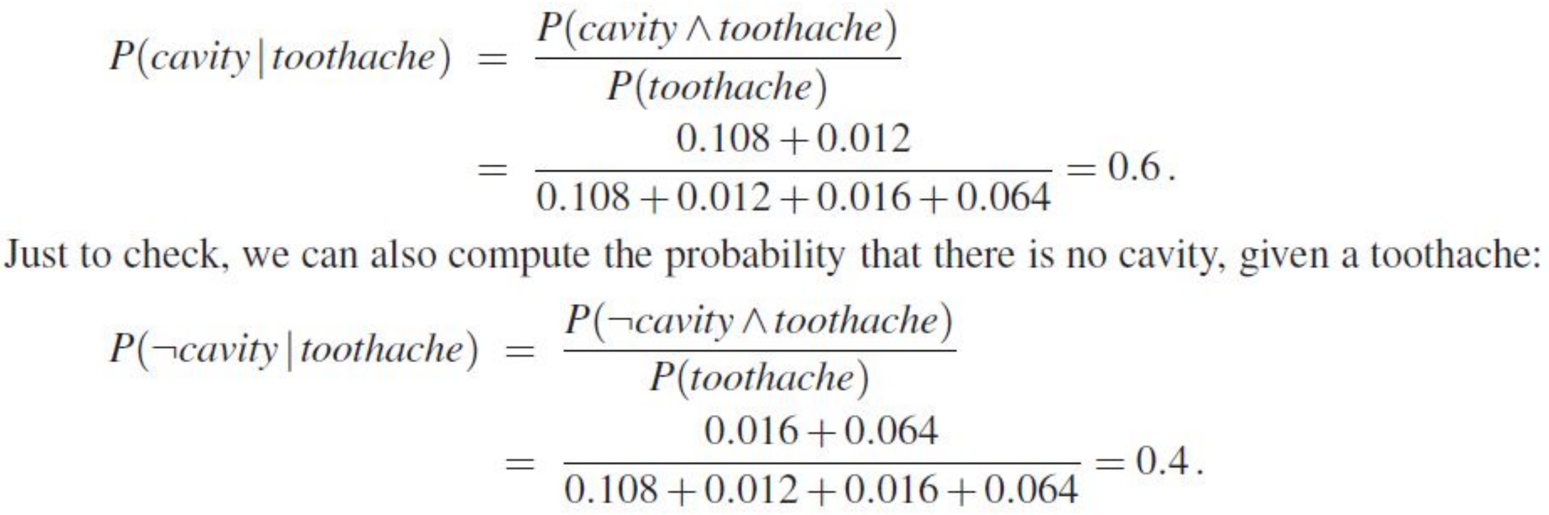
What is the unconditional or marginal probability of ¬cavity?



	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Inference using Full Joint Distributions

We can compute the probability of a cavity, given evidence of a toothache:



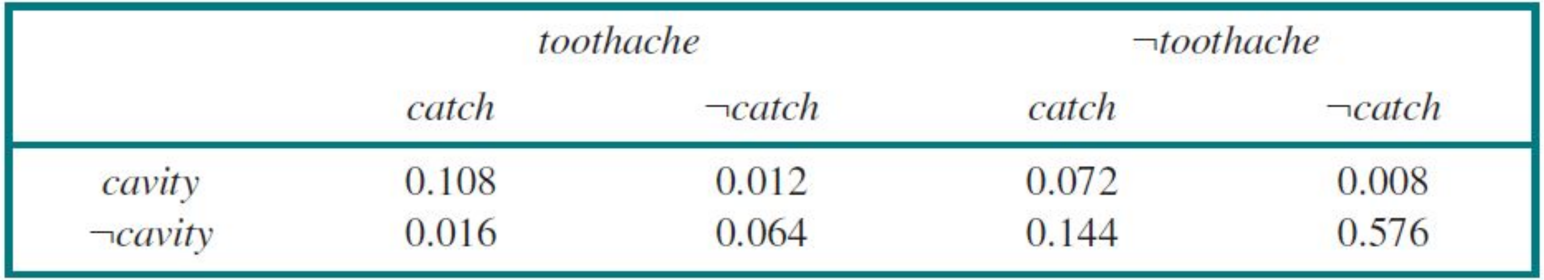
$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6\,.$$

Just to check, we can also compute the probability that there is no cavity, given a toothache:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4\,.$$



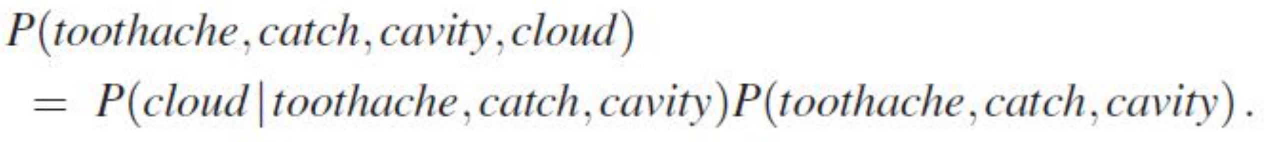
	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Independence

- Expand the full joint distribution by adding a fourth variable, Weather.
- Full joint distribution becomes P(Toothache,Catch,Cavity,Weather), with 2×2×2×4 = 32 entries.
    - Contains four “editions” of the table, one for each kind of weather.

How is the value of P(toothache,catch,cavity,cloud) related to the value of P(toothache,catch,cavity)?

- We can use the product rule:



$$\begin{aligned} &P(toothache, catch, cavity, cloud) \\ &\quad = P(cloud \mid toothache, catch, cavity) P(toothache, catch, cavity) . \end{aligned}$$

Independence

- Typically one will not imagine that one’s dental problems influence the weather.

$$P(cloud \mid toothache, catch, cavity) = P(cloud) .$$

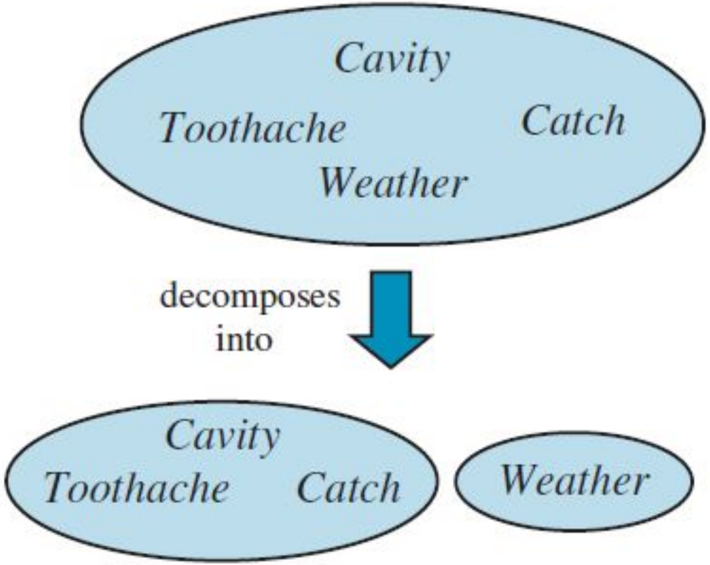
$$\begin{aligned} & P(toothache, catch, cavity, cloud) \\ & \quad = P(cloud \mid toothache, catch, cavity) P(toothache, catch, cavity) . \end{aligned}$$

From this, we can deduce

$$P(toothache, catch, cavity, cloud) = P(cloud) P(toothache, catch, cavity) .$$

Independence

$$P(toothache, catch, cavity, cloud) = P(cloud)P(toothache, catch, cavity)$$



Example of factoring a large joint distribution into smaller distributions, using absolute independence.

Independence

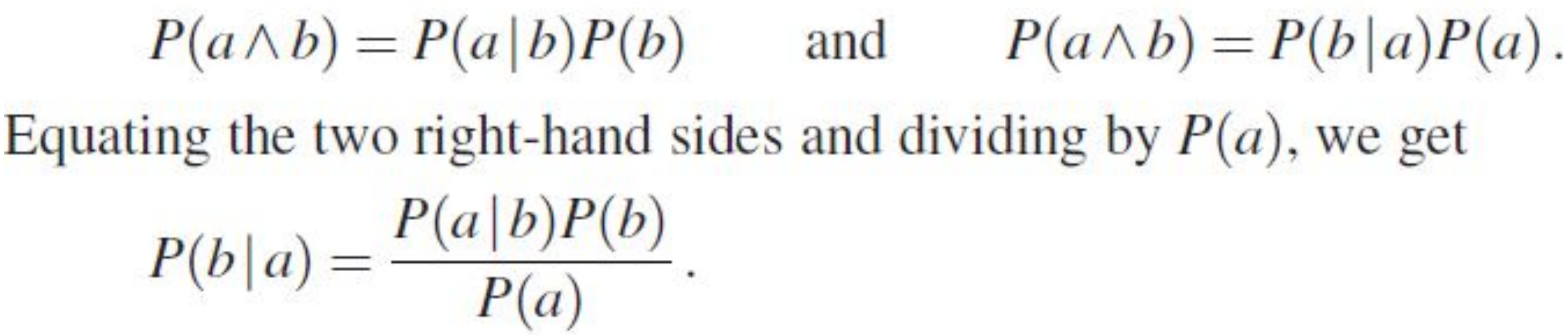
$$\mathbf{P}(Toothache, Catch, Cavity, Weather) = \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

- Weather is independent of one’s dental problems: this property is called <u>independence</u>
- Independence between variables a and b can be written as:

$$P(a|b) = P(a) \quad \text{or} \quad P(b|a) = P(b) \quad \text{or} \quad P(a \wedge b) = P(a)P(b)\,.$$

Bayes’ Rule

Product rule



$$P(a \wedge b) = P(a|b)P(b) \qquad \text{and} \qquad P(a \wedge b) = P(b|a)P(a)\,.$$

Equating the two right-hand sides and dividing by $P(a)$, we get

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}\,.$$

This equation is known as Bayes’ rule.

Applying Bayes’ rule: The Simple Case

- Bayes’ rule allows us to compute the single term P(b|a) in terms of three terms: P(a|b), P(b), and P(a).
- In a task such as medical diagnosis:
  - The doctor knows P(symptoms|disease) and want to derive a diagnosis, P(disease|symptoms).

Example: Applying Bayes' rule

- Doctor knows that the disease meningitis causes a patient to have a stiff neck, 70% of the time.
- Doctor also knows some unconditional facts: the prior probability that any patient has meningitis is 1/50,000, and the prior probability that any patient has a stiff neck is 1%.
- Let s be the proposition that the patient has a stiff neck and m be the proposition that the patient has meningitis:
- Compute what percentage of patients with a stiff neck has meningitis?

Naive Bayes Models

- Naive Bayes model => a single cause directly influence a number of effects, all of which are conditionally independent, given the cause.
- Full joint distribution:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause) .$$

- Called “naive” because of its simplifying assumption
- In practice, naive Bayes systems often work very well, even when the conditional independence assumption is not strictly true.

Text Classification with Naive Bayes

- Naive Bayes model can be used for the task of text classification: given a text, decide which of a predefined set of classes or categories it belongs to.
- Two example sentences:

Sentence 1: Stocks rallied on Monday, with major indexes gaining 1% as optimism persisted over the first quarter earnings season.

Sentence 2: Heavy rain continued to pound much of the east coast on Monday, with flood warnings issued in New York City and other locations.

- The task is to classify each sentence into a Category; the major sections of the newspaper: news, sports, business, weather, or entertainment.

Text Classification with Naive Bayes

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause) .$$

- “cause” => Category variable “effect” => presence or absence of certain keywords, HasWord_i.
- Naive Bayes consists of prior probabilities P(Category) and conditional probabilities $\text{P}(\text{HasWord}_i|\text{Category})$.

- Example: P(Category=weather)=0.09 => 9% of articles are about weather
- P(HasWord (stocks) = true|Category=business) = 0.37 => 37% of articles about business contain word “stocks”.

Document	Text	Class
1	I loved the movie	positive
2	I hated the movie	negative
3	A great movie. Good movie	positive
4	Poor acting	negative
5	Great acting. A good movie	positive

Doc.	I	loved	the	movie	hated	a	great	poor	acting	good	Class
1	1	1	1	1							positive
2	1		1	1	1						negative
3				2		1	1			1	positive
4								1	1		negative
5				1		1	1		1	1	positive

- test = “I hated the poor acting”

  P(positive / test) = P(test / positive) x p(positive) / p(test) [Bayes rule]

  ∝ P(test / positive) x p(positive) [assuming p(test) as normalization factor]

  ∝ P(“I”/positive) x P(“hated”/positive) x P(“the”/positive) x P(“poor”/positive) x P(“acting”/positive) x p(positive)

  [assuming the words in the test document are independent]

- test = “I hated the poor acting”

  P(negative / test) = P(test / negative) x p(negative) / p(test) [Bayes rule]

  ∝ P(test / negative) x p(negative) [assuming p(test) as normalization factor]

  ∝ P(“I”/negative) x P(“hated”/negative) x P(“the”/negative) x P(“poor”/negative) x P(“acting”/negative) x p(negative)

  [assuming the words in the test document are independent]

Document	Text	Class
1	I loved the movie	positive
2	I hated the movie	negative
3	A great movie. Good movie	positive
4	Poor acting	negative
5	Great acting. A good movie	positive

P(positive / test) ∝ P(“I”/positive) x P(“hated”/positive) x P(“the”/positive) x P(“poor”/positive) x P(“acting”/positive) x p(positive)

Step 1: Compute prior probabilities

P(positive) = number of positive documents / total documents = 3/5

P(negative) = number of negative documents / total documents = 2/5

Doc.	I	loved	the	movie	hated	a	great	poor	acting	good	Class
1	1	1	1	1							positive
2	1		1	1	1						negative
3				2		1	1			1	positive
4								1	1		negative
5				1		1	1		1	1	positive

P(positive / test) $\propto$ P(“I”/positive) x P(“hated”/positive) x P(“the”/positive) x P(“poor”/positive) x P(“acting”/positive) x p(positive)

Step 2: Compute posterior probabilities

P(“I” / positive) = number of times “I” occurs in positive documents / total number of words in positive documents = 1/14

P(“I” / negative) = number of times “I” occurs in negative documents / total number of words in negative documents = 1/6

Doc.	I	loved	the	movie	hated	a	great	poor	acting	good	Class
1	1	1	1	1							positive
2	1		1	1	1						negative
3				2		1	1			1	positive
4								1	1		negative
5				1		1	1		1	1	positive

P(positive / test) ∝ P(“I”/positive) x P(“hated”/positive) x P(“the”/positive) x P(“poor”/positive) x P(“acting”/positive) x p(positive)

Step 2: Similarly, compute posterior probability for word “hated”

P(“hated” / positive) = number of times “hated” occurs in positive documents / total number of words in positive documents = 0/14

P(“hated” / negative) = number of times “hated” occurs in negative documents / total number of words in negative documents = 1/6

What is the problem?

Doc.	I	loved	the	movie	hated	a	great	poor	acting	good	Class
1	1	1	1	1							positive
2	1		1	1	1						negative
3				2		1	1			1	positive
4								1	1		negative
5				1		1	1		1	1	positive

P(positive / test) ∝ P(“I”/positive) x P(“hated”/positive) x P(“the”/positive) x P(“poor”/positive) x P(“acting”/positive) x p(positive)

To avoid zero posterior probability, we perform laplace smoothing, add 1 in the numerator and add number of unique words in the training set in your denominator

P(“hated” / positive) = number of times “hated” occurs in positive documents / total number of words in positive documents = **(0 + 1) / (14 + 10)**

P(“hated” / negative) = number of times “hated” occurs in negative documents / total number of words in negative documents = (1 + 1) / (6 + 10)

Doc.	I	loved	the	movie	hated	a	great	poor	acting	good	Class
1	1	1	1	1							positive
2	1		1	1	1						negative
3				2		1	1			1	positive
4								1	1		negative
5				1		1	1		1	1	positive

P(positive / test) ∝ P("I"/positive) x P("hated"/positive) x P("the"/positive) x P("poor"/positive) x P("acting"/positive) x p(positive)

= (1+1)/(14+10) x (0+1)/(14+10) x (1+1)/(14+10) x (0+1)/(14+10) x (1+1)/(14+10) x 3/5

= 6.028×10^{-7}

P(negative / test) ∝ P("I"/negative) x P("hated"/negative) x P("the"/negative) x P("poor"/negative) x P("acting"/negative) x p(negative)

= (1+1)/(6+10) x (1+1)/(6+10) x (1+1)/(6+10) x (1+1)/(6+10) x (1+1)/(6+10) x ⅖

= 1.22×10^{-5}

Text Classification with Naive Bayes

- Naive Bayes model assumes that words occur independently in documents
- This independence assumption is clearly violated in practice
- Even with these errors, the ranking of the possible categories is often quite accurate
- Naive Bayes models are widely used for language determination, document retrieval, spam filtering, and other classification tasks.

Summary

- Uncertainty arises because of both laziness and ignorance.
- Probabilities helps us to quantify the uncertainty.
- Bayes’ rule allows unknown probabilities to be computed from known conditional probabilities.
- The naive Bayes model assumes the conditional independence of all effect variables, given a single cause variable.

THANK YOU!