CS 5/7320
Artificial Intelligence

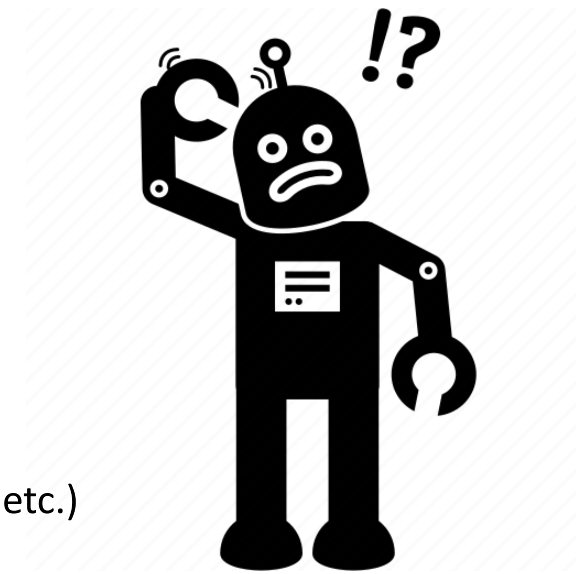
Uncertainty and Probabilities

AIMA Chapter 12

Slides by Michael Hahsler
based on slides by Svetlana Lazepnik
with figures from the AIMA textbook

"Dice" by Steve A Johnson

Uncertainty is Bad for Agents Based on Logic



Example: Catching a Flight

Let action A_t = leave for airport t minutes before flight

Question: Will A_t get me there on time?

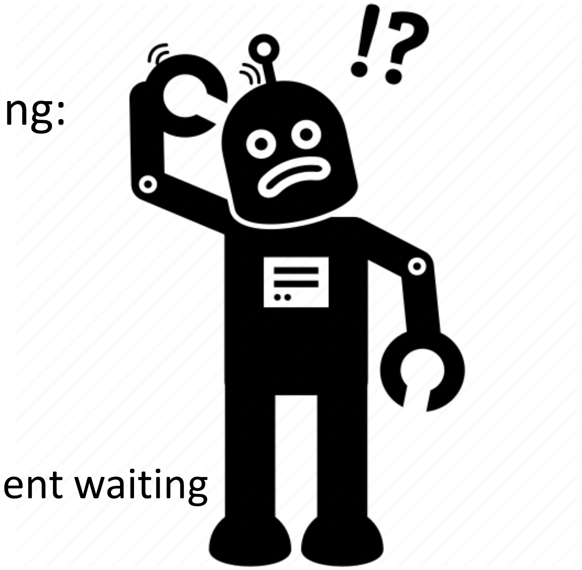
Problems:

- Partial observability (road state, other drivers' plans, etc.)
- Noisy sensors (traffic reports)
- Uncertainty in action outcomes (flat tire, etc.)
- Complexity of modeling and predicting traffic

A purely logical approach leads to conclusions that are too weak for effective decision making:

- A_{25} will get me there on time if there is no accident on the bridge and it doesn't rain and my tires remain intact, etc., etc.
- A_{Inf} guarantees to get there in time, but who lives forever?

Making Decisions Under Uncertainty



Probabilities: Suppose the agent believes the following:

$$P(A_{25}\ gets\ me\ there\ on\ time) = 0.04$$
$$P(A_{90}\ gets\ me\ there\ on\ time) = 0.80$$
$$P(A_{120}\ gets\ me\ there\ on\ time) = 0.99$$
$$P(A_{1440}\ gets\ me\ there\ on\ time) = 0.9999$$

Which action should the agent choose?

- Depends on **preferences** for missing flight vs. time spent waiting
- Encapsulated by a utility function $U(action)$

The agent should choose the action that maximizes the **expected utility**:

$$argmax_{A_t}\ [\,P(A_t\ succeeds)\ U(A_t\ succeeds) + P(A_t\ fails)\ U(A_t\ fails)\,]$$

- **Utility theory** is used to represent and infer preferences.
- **Decision theory = probability theory + utility theory**

Sources of Uncertainty

Probabilistic assertions summarize effects of:

Randomness
- Intrinsically random behavior

Ignorance
- Lack of explicit theories, relevant facts, observability, etc.

Laziness
- Failure to enumerate exceptions, qualifications, etc.

Example: What is the source of uncertainty for a coin toss?

A Quick Review of Probability Theory

What are Probabilities?

Random variables

Events

Joint probabilities

Marginal probabilities

Conditional probabilities

Bayes' Rule

Conditional independence

What are Probabilities?

Frequentism (Objective; Positivist)

Probabilities are long-run relative frequencies determined by observation.

- For example, if we toss a coin **many times**, $P(heads)$ is estimated as the proportion of the time the coin will come up heads
- But what if we are dealing with events that only happen once? E.g., what is the probability that a Republican will win the presidency in 2024? How do we define comparable elections? **Reference class problem**.

Bayesian Statistics (Subjective)

Probabilities are degrees of belief based on prior knowledge and updated by evidence.

Provides tools to:

- How do we assign belief values to statements without evidence?
- How do we update our degrees of belief given observations?

Random variables

Random Variable

- We describe the (uncertain) state of the world using *random variables.*
- Random variables are denoted by capital letters.

- **R**: *Is it raining?*
- **W**: *What's the weather?*
- **Die**: *What is the outcome of rolling two dice?*
- **V**: *What is the speed of my car (in MPH)?*

Domain

- Random variables take on values in a *domain D.*
- Domain values must be mutually exclusive and exhaustive.

- **R** ∈ {True, False}
- **W** ∈ {Sunny, Cloudy, Rainy, Snow}
- **Die** ∈ {(1,1), (1,2), … (6,6)}
- **V** ∈ [0, 200]

Events and Propositions

Probabilistic statements are defined over **events**, world states or sets of states

- *“It is raining”*
- *“The weather is either cloudy or snowy”*
- *“The sum of the two dice rolls is 11”*
- *“My car is going between 30 and 50 miles per hour”*

Events are described using **propositions**:

- R = True
- W = “Cloudy” $\vee$ W = “Snowy”
- D $\in$ {(5,6), (6,5)}
- 30 $\leq$ S $\leq$ 50

Notation:

- For events: $P(A)$ is the probability that event A happens.
- For propositions: $P(A = true), P(a)$ is the probability of the set of possible worlds in which proposition A holds.
- For random variables: $P(X = x)$ or $P_X(x)$ or $P(x)$ for short, is the probability of the event that random variable X has taken on the value x.

Kolmogorov's 3 Axioms of Probability

Three axioms are sufficient to define probability theory:

1. Probabilities are non-negative real numbers.
2. The probability that at least one atomic event happens is 1.
3. The probability of mutually exclusive events is additive.

This leads to important properties (A and B are sets of events):

- Numeric bound: $0 \leq P(A) \leq 1$
- Monotonicity: if $A \subseteq B$ then $P(A) \leq P(B)$
- Addition law: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Probability of the empty set: $P(\emptyset) = 0$
- Complement rule: $P(\neg A) = 1 - P(A)$

- Continuous variables need in addition the definition of density functions.

Atomic events

- Atomic event: a complete specification of the state of the world, or a complete assignment of domain values **to all random variables.**

- Atomic events are mutually exclusive and exhaustive.

- E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

$$Cavity = false \wedge Toothache = false$$
$$Cavity = false \wedge Toothache = true$$
$$Cavity = true \wedge Toothache = false$$
$$Cavity = true \wedge Toothache = true$$

Joint probability distributions

- A **joint distribution** is an assignment of probabilities to every possible atomic event

Atomic event	P
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

Sum: 1.00

- Notation:
  - $P(x), P(X = x)$ is the **probability** that random variable X takes on value x
  - $\boldsymbol{P}(X)$ is the **distribution of probabilities** for all possible values of X. Often we are lazy or forget to make P bold.

Marginal probability distributions

- Sometimes we are only interested in one variable. This is called the *marginal distribution* $\boldsymbol{P}(Y)$

P(Cavity, Toothache)	
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

Marginal Prob. Distr.

P(Cavity)	
Cavity = false	?
Cavity = true	?

P(Toothache)	
Toothache = false	?
Toothache = true	?

Marginal probability distributions

- Suppose we have the joint distribution $\boldsymbol{P}(X, Y)$ and we want to find the *marginal distribution* $\boldsymbol{P}(Y)$

$$P(X = x) = P((X = x \land Y = y_1) \lor \cdots \lor (X = x \land Y = y_n))$$
$$= P((x, y_1) \lor \cdots \lor (x, y_n)) = \sum_{i=1}^{n} P(x, y_i)$$

- **General rule**: to find $P(X = x)$, sum the probabilities of all atomic events where $X = x$. This is called "summing out" or marginalization.

Marginal probability distributions

- Suppose we have the joint distribution $\boldsymbol{P}(X, Y)$ and we want to find the *marginal distribution* $\boldsymbol{P}(Y)$

P(Cavity, Toothache)	
Cavity = false ∧ *Toothache = false*	0.8
Cavity = false ∧ *Toothache = true*	0.1
Cavity = true ∧ *Toothache = false*	0.05
Cavity = true ∧ *Toothache = true*	0.05

Marginal Prob. Distr.

P(Cavity)	
Cavity = false	0.8+0.1 = 0.9
Cavity = true	0.05+0.05=0.1

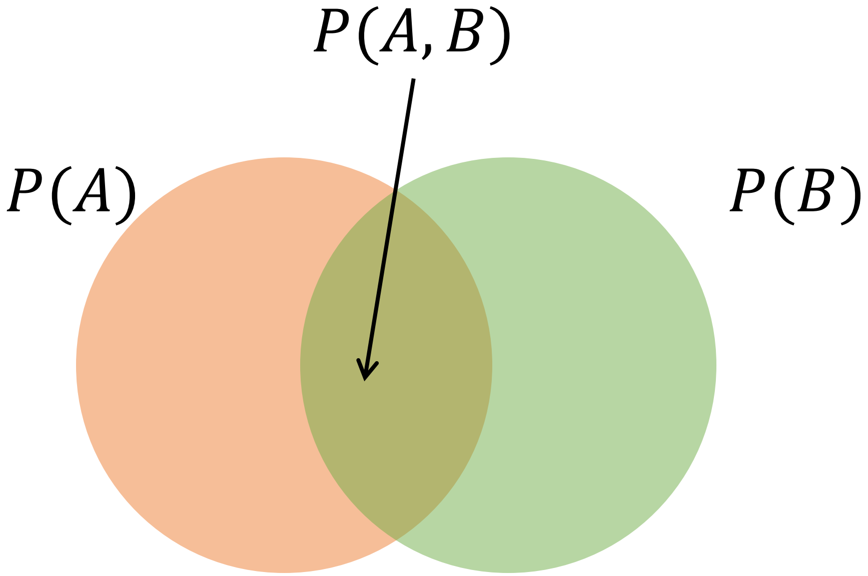
P(Toothache)	
Toothache = false	0.8+0.0.5= 0.85
Toothache = true	0.1+0.05= 0.15

Conditional probability

- Probability of cavity given toothache:

  $\mathrm{P}(\mathrm{Cavity} = \mathrm{true} \mid \mathrm{Toothache} = \mathrm{true})$

- For any two events A and B, $P(A \mid B) = \frac{P(A,B)}{P(B)}$



Conditional probability

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

Joint Prob. Distr.

P(Cavity, Toothache)	
Cavity = false ∧ Toothache = false	0.8
Cavity = false ∧ Toothache = true	0.1
Cavity = true ∧ Toothache = false	0.05
Cavity = true ∧ Toothache = true	0.05

Marginal Prob. Distr.

P(Cavity)	
Cavity = false	0.9
Cavity = true	0.1

P(Toothache)	
Toothache = false	0.85
Toothache = true	0.15

- What is P(*Cavity = true | Toothache = false*)?

  0.05 / 0.85 = 0.059

- What is P(*Cavity = false | Toothache = true*)?

  0.1 / 0.15 = 0.667

Conditional distributions

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

P(Cavity, Toothache)	
Cavity = false ∧ *Toothache = false*	0.8
Cavity = false ∧ *Toothache = true*	0.1
Cavity = true ∧ *Toothache = false*	0.05
Cavity = true ∧ *Toothache = true*	0.05

A conditional distribution is a distribution over the values of one variable given fixed values of other variables

P(Cavity | Toothache = true)	
Cavity = false	0.667
Cavity = true	0.333

P(Cavity | Toothache = false)	
Cavity = false	0.941
Cavity = true	0.059

P(Toothache | Cavity = true)	
Toothache= false	0.5
Toothache = true	0.5

P(Toothache | Cavity = false)	
Toothache= false	0.889
Toothache = true	0.111

Normalization trick

- To get the whole conditional distribution $\mathbf{P}(X \mid Y = y)$ at once, select all entries in the joint distribution matching $Y = y$ and renormalize them to sum to one.

P(Cavity, Toothache)	
Cavity = false ∧ *Toothache = false*	0.8
Cavity = false ∧ *Toothache = true*	0.1
Cavity = true ∧ *Toothache = false*	0.05
Cavity = true ∧ *Toothache = true*	0.05

Select P(X, Y = y)

Toothache, Cavity = false	
Toothache= false	0.8
Toothache = true	0.1

Sum is $P(Y = y) = 0.9$

Renormalize sum to 1 (= divide by $P(Y = y)$)

P(Toothache | Cavity = false)	
Toothache= false	0.889
Toothache = true	0.111

Equivalent to
$\mathbf{P}(X \mid Y = y) = \alpha\, \mathbf{P}(X, Y = y)$
with $\alpha = 1/P(Y = y)$

Bayes' Rule

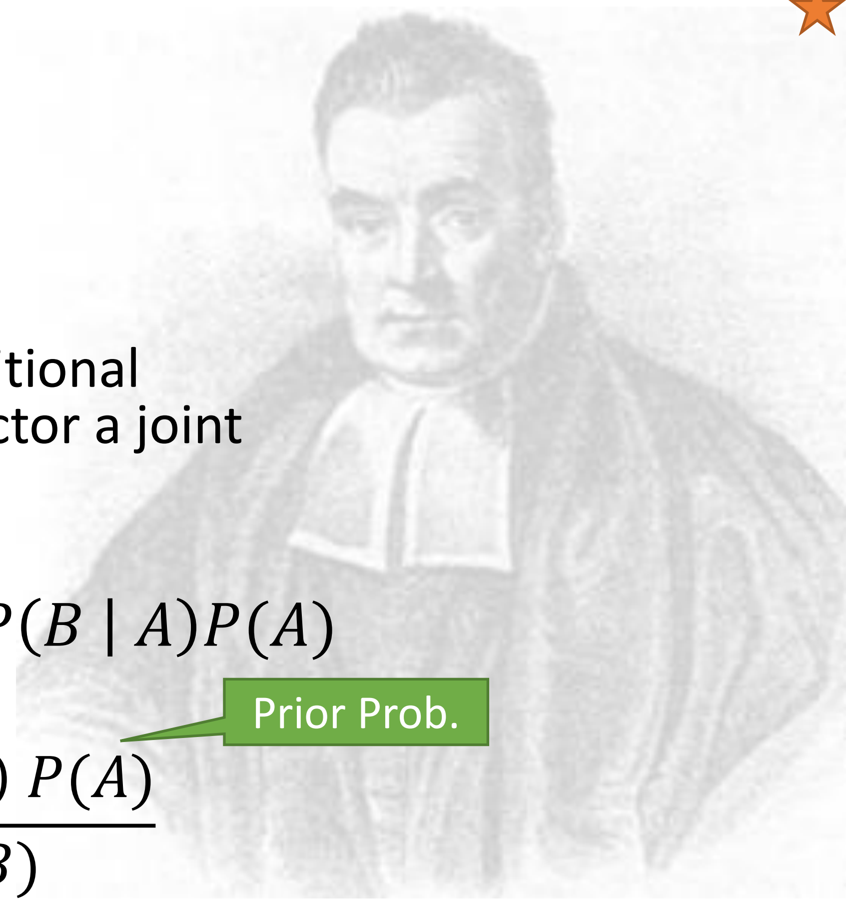
- The **product rule** (definition of conditional distribution) gives us two ways to factor a joint distribution for events A and B:

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

Posterior Prob.

Prior Prob.

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



Rev. Thomas Bayes
(1702-1761)

- Why is this useful?
  - Can get *diagnostic probability* P(Cavity | Toothache) from *causal probability* P(Toothache | Cavity)
  - We can update our beliefs based on evidence.
  - Important tool for probabilistic inference .

Example: Getting Married in the Desert

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 Marie’s belief for the probability that it will rain on her wedding day?

Example: Getting Married in the Desert

New Evidence

Prior Probability

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 Marie's belief for the **probability that it will rain** on her wedding day?

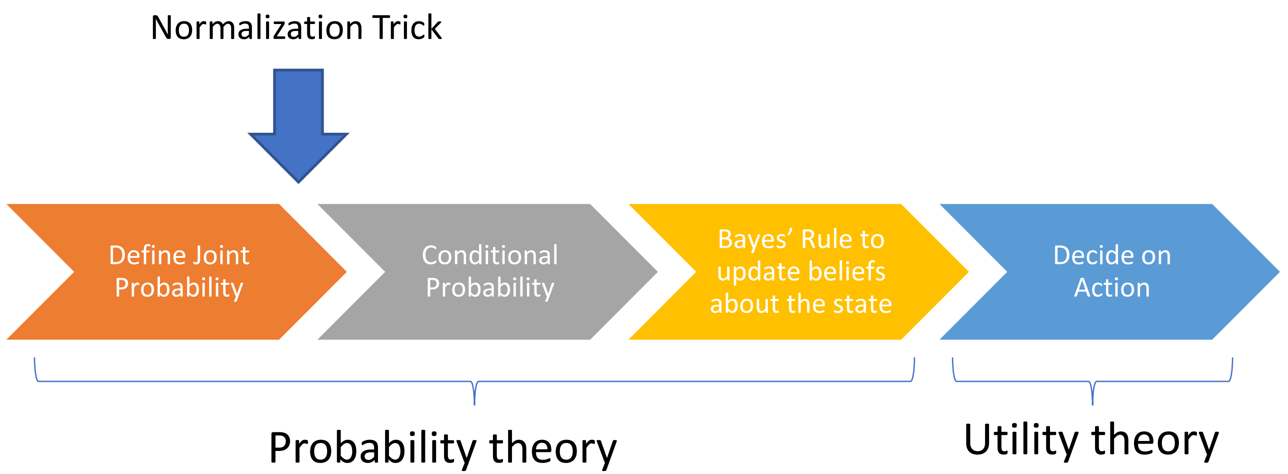
Posteriori Probability

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$$\begin{aligned} P(\text{Rain}|\text{Predict}) &= \frac{P(\text{Predict}|\text{Rain})P(\text{Rain})}{P(\text{Predict})} \\ &= \frac{P(\text{Predict}|\text{Rain})P(\text{Rain})}{P(\text{Predict}|\text{Rain})P(\text{Rain}) + P(\text{Predict}|\neg\text{Rain})P(\neg\text{Rain})} \\ &= \frac{0.9 * 0.014}{0.9 * 0.014 + 0.1 * 0.986} = 0.111 \end{aligned}$$

The weather forecast updates her belief from 0.014 to 0.111

Approach to Choose Actions



- **Problem: the joint probability table is typically too large!** For n random variables with a domain size of d each, we have a table of size $O(d^n)$. This is a problem for
  - storing the table and
  - estimating the probabilities from data (we need lots of data).

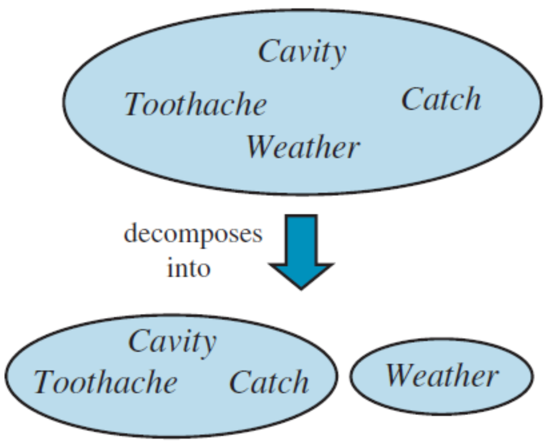
- **Solution**: Decomposition of joint probability distributions using **independence** and conditional independence between events. A large table can be broken into several much smaller tables.

Independence Between Events

- Two events A and B are **independent** if and only if

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

- This is equivalent to $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$
- Independence is an important simplifying assumption for modeling, e.g., *Cavity* and *Weather* can be assumed to be independent



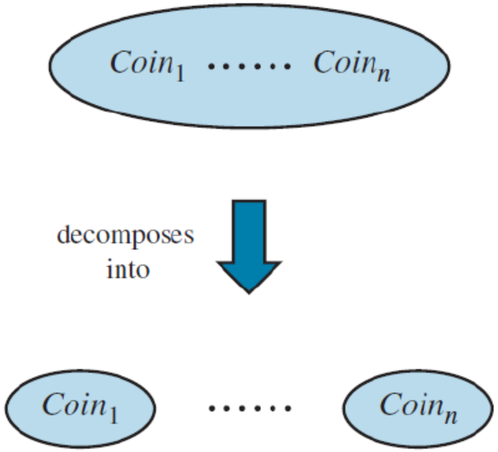
Independence ⇒ P(Cavity, Weather) = P(Cavity)P(Weather)
P(Cavity | Weather) = P(Cavity)

Decomposition of the Joint Probability Distribution

- **Independence:** The joint probability can be decomposed into

$$\boldsymbol{P}(Coin_1, \dots, Coin_n)$$
$$= \boldsymbol{P}(Coin_1) \times \cdots \times \boldsymbol{P}(Coin_n) = \prod_i \boldsymbol{P}(Coin_i)$$

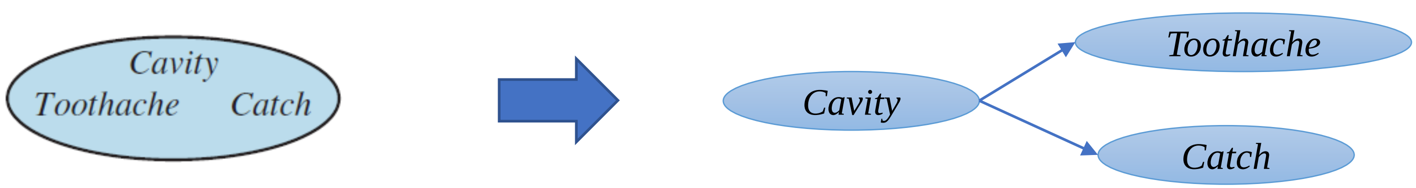
- We need for each coin one parameter (chance of getting H).
- Independence reduces the numbers needed to specify the joint distribution from $2^n - 1$ to n.
- Note: If we have identical (iid) coins, then we even only need 2 numbers, probability of H and number of coins.



Conditional Independence

- **Conditional independence**: A and B are *conditionally independent* given C (i.e., if we know c) iff

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



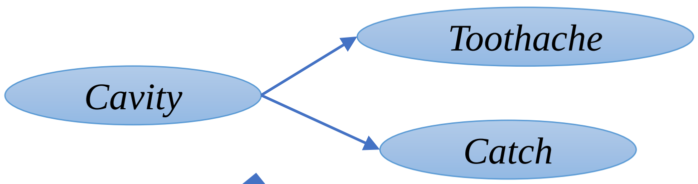
- If the patient has a cavity, the probability that the probe catches in it does not depend on whether he/she has a toothache

  P(*Catch* | *Toothache, Cavity*) = P(*Catch* | *Cavity*)

- Therefore*, Catch* is conditionally independent of *Toothache* given *Cavity*
- Likewise, *Toothache* is conditionally independent of *Catch* given *Cavity*

  P(*Toothache* | *Catch, Cavity*) = P(*Toothache* | *Cavity*)

Decomposition of the Joint Probability Distribution



- **Conditional independence** using the chain rule:

  P(Toothache, Catch, Cavity) =
  P(Cavity) P(Catch | Cavity) **P**(Toothache | ~~Catch~~, Cavity) =
  P(Cavity) P(Catch | Cavity) **P**(Toothache | Cavity)

- The full joint probability distribution needs $2^3 - 1 = 7$ independent numbers (-1 because the 2^3 numbers have to sum up to 1).
- Conditional independence reduces this to $1 + 2 + 2 = 5$
- In many practical applications, conditional independence reduces the space requirements from $O(2^n)$ to $O(n)$.

Chain rule: Example for 4 variables.

$$\begin{aligned}
\mathrm{P}(X_4, X_3, X_2, X_1) &= \mathrm{P}(X_4 \mid X_3, X_2, X_1) \cdot \mathrm{P}(X_3, X_2, X_1) \\
&= \mathrm{P}(X_4 \mid X_3, X_2, X_1) \cdot \mathrm{P}(X_3 \mid X_2, X_1) \cdot \mathrm{P}(X_2, X_1) \\
&= \mathrm{P}(X_4 \mid X_3, X_2, X_1) \cdot \mathrm{P}(X_3 \mid X_2, X_1) \cdot \mathrm{P}(X_2 \mid X_1) \cdot \mathrm{P}(X_1)
\end{aligned}$$

Bayesian Decision Making

Making Decisions under Uncertainty based on Evidence

Probabilistic Inference

Suppose the agent has to guess the value of an unobserved *query variable* X given some observed *evidence* $E = e$ and we assume X probabilistically causes E.

Examples:

x ∈ {zebra, giraffe, hippo}, e = image features
x ∈ {spam, not spam}, e = email message

What is the best guess x^*?

Notation: We use here $\hat{x}$ for an estimate and x^* for the best estimate.

Bayes' Decision Theory

- Assumption: The agent has a ***loss function***, which is 0 if the value of X (x) is guessed correctly, and 1 otherwise.

$$L(x, \hat{x}) = \begin{cases} 1 \text{ if } \hat{x} \neq x, \text{ and} \\ 0 \text{ otherwise.} \end{cases}$$

- The value for X that minimizes the ***expected loss*** is the one that has the greatest posterior probability given the evidence.

$P(X = x \mid E = e)$ often written as $P(x \mid e)$.

- This is called the **MAP** (maximum a posteriori) decision.

MAP: Maximum A Posteriori Decision

Use the value x that has the highest (maximum) posterior probability given the evidence e

$$x^* = \text{argmax}_x \overbrace{P\,(x|e)}^{\text{Posterior Prob.}} = \text{argmax}_x \frac{P(e|x)\overbrace{P(x)}^{\text{Prior Prob.}}}{P(e)}$$

$$\propto \text{argmax}_x\, P(e|x)P(x)$$

$P(e)$ is fixed for a given example.

For comparison: the maximum likelihood decision ignores $P(x)$

$$x^* = \text{argmax}_x \underbrace{P\,(e|x)}_{\text{likelihood}}$$

MAP: Example

Value of x that has the highest (maximum) posterior probability given the evidence e.

$x \in$ {zebra, dog, cat}, e = stripes

$$x^* = \operatorname{argmax}_x \overbrace{P(x|e)}^{\text{Posterior Prob.}} = \operatorname{argmax}_x \frac{P(\text{stripes}|x)P(x)}{P(stripes)}$$

$$\propto \operatorname{argmax}_x \underbrace{P(\text{stripes}|x)}_{\text{likelihood}} \underbrace{P(x)}_{\text{Prior Prob.}}$$

The likelihood $P(\text{stripes} \mid \text{zebra})$ is the highest, but it also depends on the prior $P(\text{zebra})$, the chance that we see a zebra (which would be a lot higher in a zoo than in downtown Dallas).

Bayes Classifier

- Suppose we have many different types of observations (evidence, symptoms, features) $F_1, \ldots, Fn$ that we want to use to decide on an underlying hypothesis H.

- MAP decision involves estimating

$$P(H|F_1, \ldots,\ F_n) \propto P(F_1, \ldots,\ F_n|H)P(H)$$

- If each feature can take on k values, how many entries are in the joint probability table $P(F_1, \ldots,\ F_n|H)$?

Naïve Bayes model

- Suppose we have many different types of observations (evidence, symptoms, features) $F_1, \ldots, F_n$ that we want to use to obtain evidence about an underlying hypothesis H
- MAP decision involves estimating
$$P(H|F_1, \ldots,\ F_n) \propto P(F_1, \ldots,\ F_n|H)P(H)$$
- **Issue**: The likelihood table size grows exponentially with the number of features n.

- We can make the **simplifying assumption** that the different **features are conditionally independent *given the hypothesis*** reduces the joint probability distribution table to size $k \times n$:
$$P(F_1, \ldots,\ F_n|H) = \prod_{i=1}^{n} P(F_i|H)$$

Example: Naïve Bayes Spam Filter

- The hypothesis can be $spam$ or $\neg$spam and the evidence is the message.
- **MAP decision:** to minimize the probability of error, we should classify a message as spam if

$$P(H = \text{spam} \mid \text{message}) > P(H = \neg\text{spam} \mid \text{message})$$

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.

How do we represent the messages?

Natural Language Processing: Bag of Words

Represent a document features as binary vector $(w_1, \dots, w_n)$.
Each element represents the event that word w_i is present (w_i = 1) or not (w_i = 0) .

Simplifications:

- The order of the words in the message is ignored.
- How often a word is repeated is ignored.

Represent a document features as binary vector $(w_1, \dots, w_n)$.
Each element represents the event that word w_i is present (w_i = 1) or not (w_i = 0) .

Simplifications:

- The order of the words in the message is ignored.
- How often a word is repeated is ignored.

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 Spam Filter

- If we assume that each word is conditionally independent of the others given message class (spam or not spam), then we can use a naïve Bayes classifier.

$$P(\text{message}|\mathrm{H}) = P(w_1, \dots, w_n|\mathrm{H}) = \prod_{i=1}^{n} P(w_i|\mathrm{H})$$

$$\underbrace{P(\mathrm{H}|w_1, \dots, w_n)}_{\text{posterior}} \propto \underbrace{P(\mathrm{H})}_{\text{prior}} \underbrace{\prod_{i=1}^{n} P(w_i|\mathrm{H})}_{\substack{\text{Evidence} \\ \text{(presents and absence of words)}}}$$

Parameter estimation

In order to classify a message, we need to know

1. the prior $P(H)$ and
2. the likelihoods $P(\text{word} = 1 \mid H)$, $P(\text{word} = 0 \mid \text{spam})$

These are the *parameters* of the probabilistic model:

prior

spam:	0.33
¬spam:	0.67

$P(\text{word} = 1 \mid H = \text{spam})$

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

$P(\text{word} = 1 \mid H = \neg\text{spam})$

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

+ likelihoods for the absence of words:

$P(\text{word} = 0 \mid H = \text{spam}) = 1 - P(\text{word} = 1 \mid H = \text{spam})$

$P(\text{word} = 0 \mid H = \neg\text{spam}) = 1 - P(\text{word} = 1 \mid H = \neg\text{spam})$

Parameter estimation: Prior

How do we obtain the prior $P(H)$?

Empirically: use training data

$$P(H = spam) = \frac{\#\ of\ spam\ messages}{total\ \#\ of\ messages}$$

$P(H = \neg spam) = 1 - P(H = spam)$

Parameter estimation: Likelihoods

How do we obtain the likelihoods P(word = 1 | H = spam) and P(word = 1 | H = ¬spam)?

Empirically: use training data

$$P(word = 1 \mid H = spam) = \frac{\text{\# of spam messages that contain the word}}{\text{total \# of spam messages}}$$

Parameter estimation: Smoothing

- **Problem:** What happens with words that we have never seen or seen only a few times?

- **Laplacian smoothing:** add one to each count

$$P(\textit{word} = 1 \mid H = \textit{spam}) = \frac{\textit{\# of spam messages that contain the word} + 1}{\textit{total \# of spam messages} + \textit{\# of classes}}$$

Note: This is actually a Bayesian estimate with +1 and # of classes (2 for spam/not spam) representing an uniformed prior probability.

Summary of model and parameters

- **Naïve Bayes model:**

$$P(H = spam|message) \propto P(H = spam)\prod_{i=1}^{n} P(w_i|H = spam) = score_{spam}(message)$$

$$P(H = \neg spam|message) \propto P(H = \neg spam)\prod_{i=1}^{n} P(w_i|H = \neg spam) = score_{\neg spam}(message)$$

- **Model parameters:**

prior

$P(\mathrm{H} = spam)$

$P(\mathrm{H} = \neg spam)$

Likelihood of words in spam

$P(w_1 = 1 \mid \mathrm{H}=spam)$

$P(w_2 = 1 \mid \mathrm{H}=spam)$

$\dots$

$P(w_n = 1 \mid \mathrm{H}=spam)$

Likelihood of words in ¬spam

$P(w_1 = 1 \mid \mathrm{H}=\neg spam)$

$P(w_2 = 1 \mid \mathrm{H}=\neg spam)$

$\dots$

$P(w_n = 1 \mid \mathrm{H}=\neg spam)$

+ Laplacian Smoothing

+ likelihood of words not in spam (or $\neg spam$) can be calculated as $P(w_{\mathrm{i}} = 0 \mid \mathrm{H}=spam) = 1 - P(w_{\mathrm{i}} = 1 \mid \mathrm{H}=spam)$

- **Decision:** Spam if

$$P(H = spam \mid message) \;>\; P(H = \neg spam \mid message)$$

equivalent to

$$score_{spam}(message) > score_{\neg spam}(message)$$

Bayesian decision making: Summary

- Suppose the agent has to guess 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 the posterior probability $P(X \mid E = e)$?

  Use $P(x|e) = \frac{P(e|x)P(x)}{P(e)}$

- **Learning problem:** estimate the parameters of the probabilistic model needed for inference. Estimate the probability distributions $\boldsymbol{P}(E \mid X)$ and $\boldsymbol{P}(X)$ given a set of training samples $\{(x1, e1), \dots, (xn, en)\}$

- A general framework for learning from data is the goal of **Machine Learning**.