CS 5/7320
Artificial Intelligence

Quantifying Uncertainty: Probabilities

AIMA Chapter 12

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

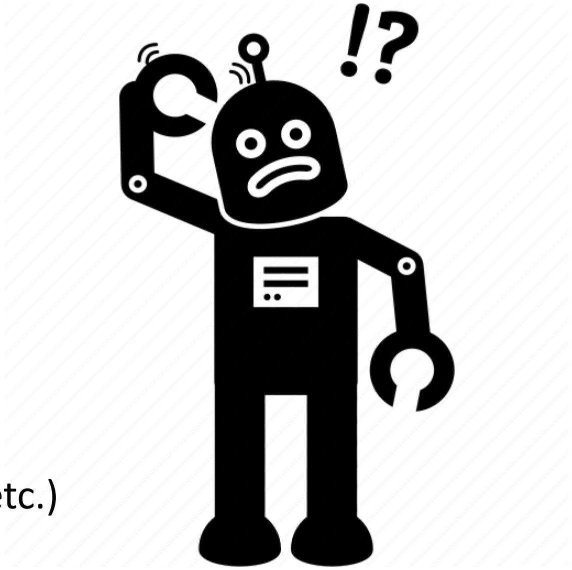
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

"Dice" by Steve A Johnson

Contents

1. Making Decisions under Uncertainty
2. Probability Theory
3. Bayes’ Rule
4. Independence between Events
5. Bayesian Decision Making
6. Example: Naïve Bayesian Classifier

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?

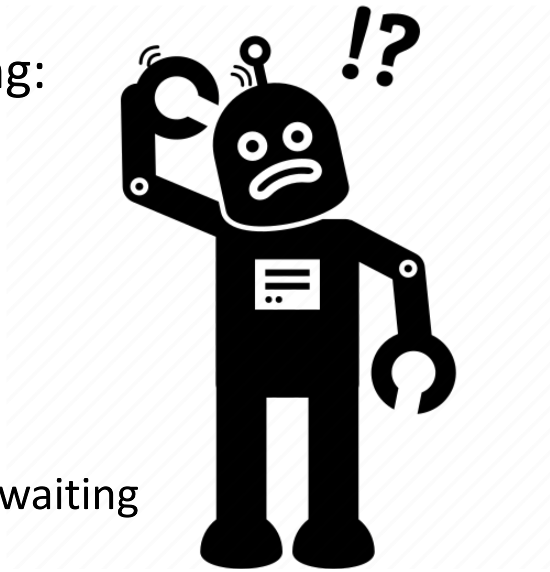
Problems:

- 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
- **Utility theory** represents preferences for actions using a utility function $U(action)$

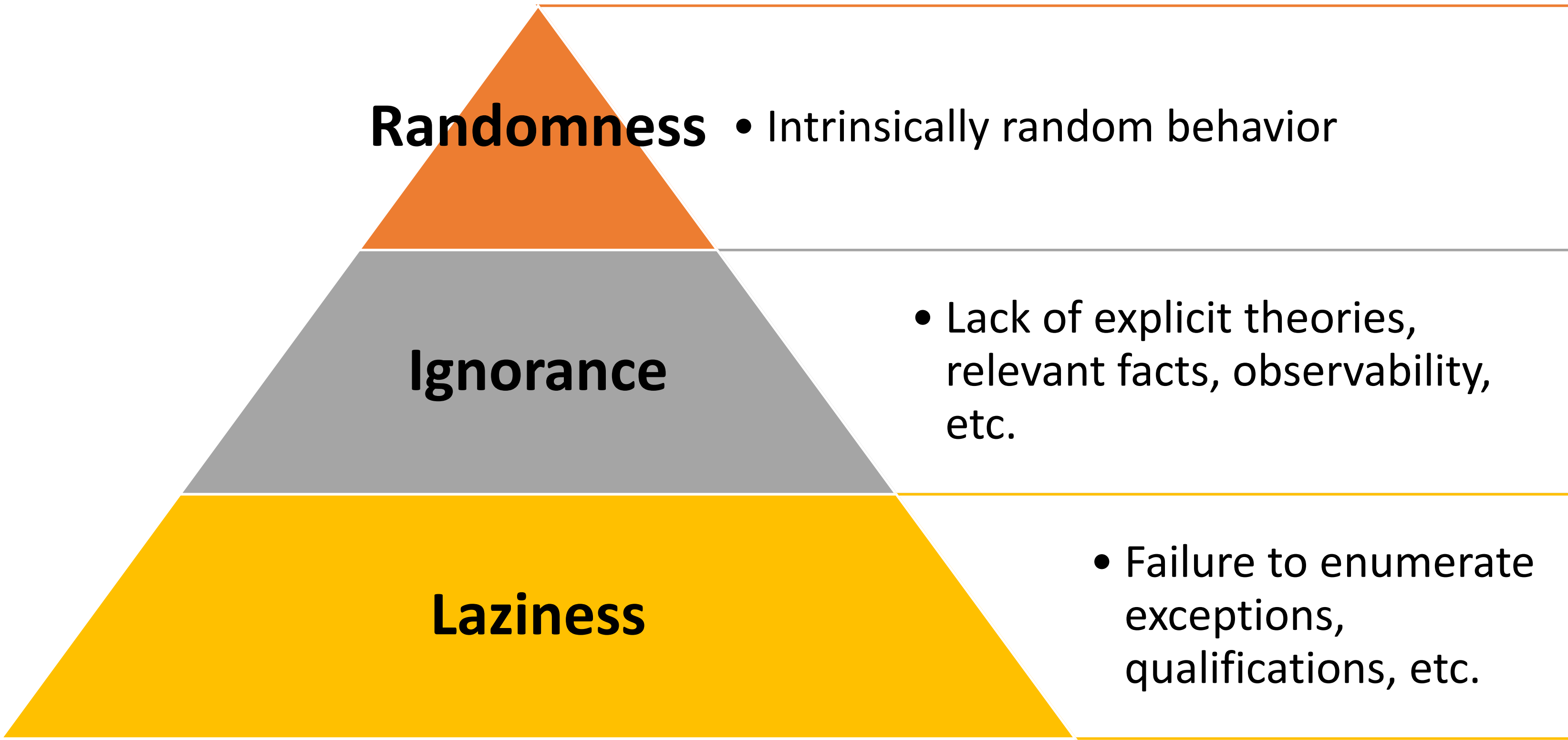
Decision theory = probability theory + utility theory

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

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

Sources of Uncertainty

Probabilistic assertions summarize effects of:



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

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**.

Good if we have lots of data

Bayesian Statistics (Subjective)

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

Provides tools to:

- Assign belief values to statements without evidence
- Update our degrees of belief given observations = **Learning**

Good for little data and learning

Probability Theory Recap

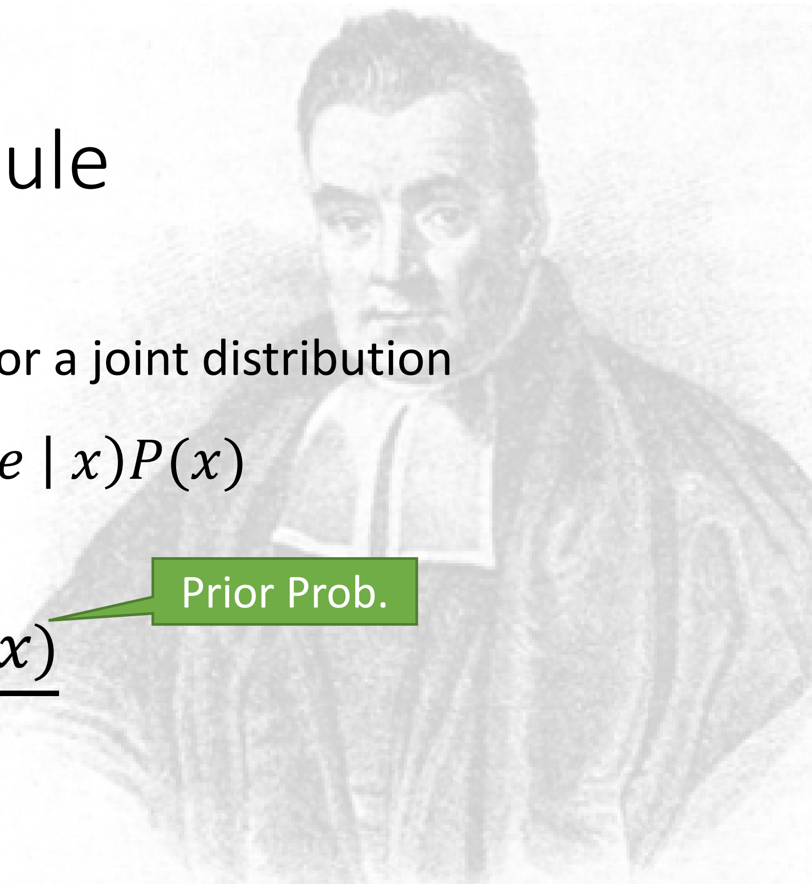
- Notation: Prob. of an event $P(X = x) = P(x)$
  Prob. distribution $\boldsymbol{P}(X) = \langle P(X = x_1), P(X = x_2), \dots, P(X = x_n)\rangle$

- Product rule $P(x, y) = P(x|y)P(y)$
- Chain rule $\boldsymbol{P}(\mathrm{X_1}, \mathrm{X_2}, \dots, \mathrm{X_n}) = \mathbf{P}(\mathrm{X_1})\mathbf{P}(\mathrm{X_2}|\mathrm{X_1})\mathbf{P}(\mathrm{X_3}|\mathrm{X_1}, \mathrm{X_2}) \dots$
  $= \prod_{i=1}^{n} \boldsymbol{P}(X\_i|X_1, \dots, X_{i-1})$
- Conditional probability $P(x|y) = \frac{P(x,y)}{P(y)} = \alpha P(x, y)$

- Independence
  - $X \perp\!\!\!\perp Y$: X, Y are independent (written as $X \perp\!\!\!\perp Y$) if and only if:
    $\forall x, y: P(x, y) = P(x)P(y)$

  - $X \perp\!\!\!\perp Y|Z$: X and Y are conditionally independent given Z if and only if:
    $\forall x, y, z: P(x, y|z) = P(x|z)P(y|z)$

Bayesian Update: Bayes' Rule



Rev. Thomas Bayes
(1702-1761)

The **product rule** gives us two ways to factor a joint distribution for events x and e:

$$P(x, e) = P(x \mid e)P(e) = P(e \mid x)P(x)$$

Posterior Prob.

Prior Prob.

Therefore, $$P(x \mid e) = \frac{P(e|x)P(x)}{P(e)}$$

Why is this useful?

- We can update our beliefs about an event $X = x$ based on new evidence ($E = e$).
- We can get *diagnostic probability* $P(Cavity \mid Toothache)$ from *causal probability* $P(Toothache \mid Cavity)$

Written as distributions $$\boldsymbol{P}(X \mid E) = \frac{\boldsymbol{P}(E|X)\boldsymbol{P}(X)}{\boldsymbol{P}(E)}$$

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 of rain

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?

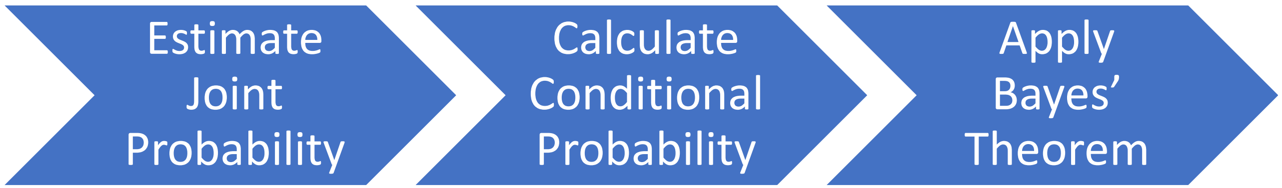
Posterior Probability?

$$P(x \mid e) = \frac{P(x \mid e)\, P(x)}{P(e)}$$

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

The weather forecast updates her belief from 0.014 to 0.111

Issue With Applying Bayes’ Theorem



Issue: The joint probability table is typically way 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:

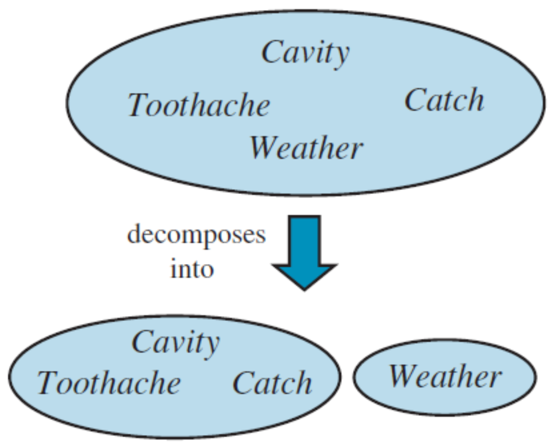
- Decomposition of joint probability distributions using **independence** and conditional independence between events.
- A large table can be broken into several much smaller tables.

Independence Between Events

- Two events A and B are **independent** ($A \perp\!\!\!\perp B$) if and only if

$$\boldsymbol{P}(A, B) = \boldsymbol{P}(A)\,\boldsymbol{P}(B)$$

- This is equivalent to $\boldsymbol{P}(A \mid B) = \boldsymbol{P}(A)$ and $\boldsymbol{P}(B \mid A) = \boldsymbol{P}(B)$
- 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)

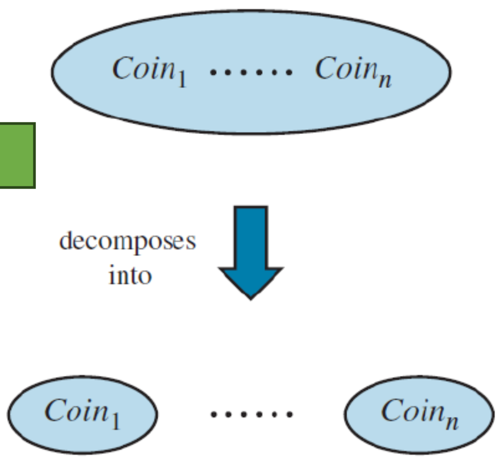
Decomposition of the Joint Probability Distribution With Independence

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

$2^n - 1$ entries

n entries

$$\boldsymbol{P}(Coin_1, \dots, Coin_n) =$$
$$\boldsymbol{P}(Coin_1) \times \cdots \times \boldsymbol{P}(Coin_n) = \prod_{i=1}^{n} \boldsymbol{P}(Coin_i)$$



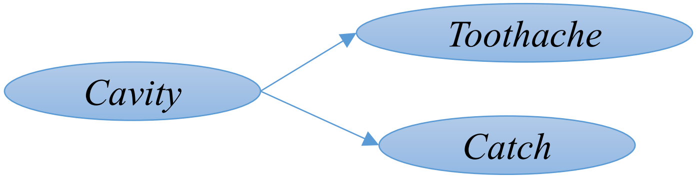
- We need for each coin one parameter for the chance of getting Head.
- Independence reduces the numbers needed to specify the joint distribution from $\mathbf{2^n - 1}$ **to** $\boldsymbol{n}$.
- If we have identical (iid) coins, then we even only need 2 numbers, the probability of H and the number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

Conditional Independence

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

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

Example:



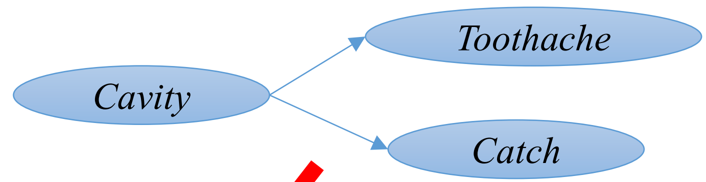
- If the patient has a cavity, the probability that the probe catches does not depend on whether he/she has a toothache

  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 With Conditional Independence



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

$2^3 - 1 =$
7 entries

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

$1 + 2 + 2 =$
5 entries

- In many practical applications, conditional independence reduces the space requirements significantly from $\boldsymbol{O(2^n)}$ **to** $\boldsymbol{O(n)}$.
- This makes Bayesian Networks (in the next chapter) so useful.

Bayesian Decision Making

Making Decisions Under Uncertainty Based on Evidence

Probabilistic Inference

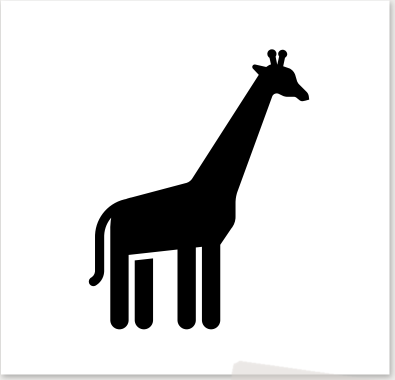
Suppose the agent must guess the value of an unobserved *query variable* X given some observed *evidence* $E = e$ and we assume X probabilistically causes E.

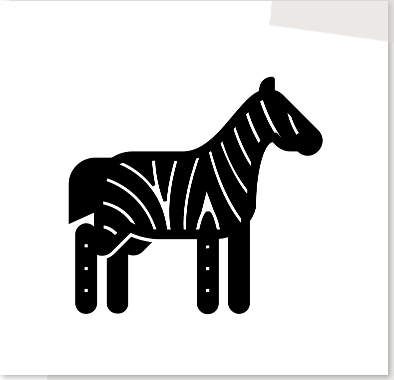
Example:

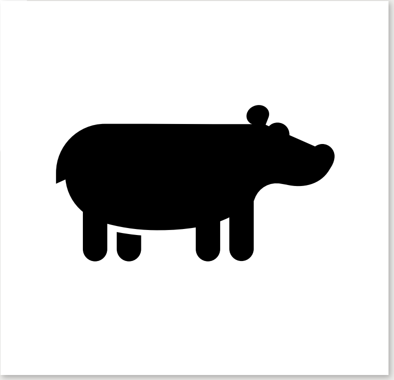
$x \in$ {zebra, giraffe, hippo}, e = image features

What is the best guess x^*?

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







The Optimal Bayes Decision Rule

- **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.

$$\operatorname{argmax}_x P(X = x \mid E = e)$$

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

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 frequentist 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. If the likelihood for cats is smaller, but the prior probability is much higher, cat may have a larger posterior probability!

Bayes Classifier

$F_1, F_2, \ldots, F_n, H$

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

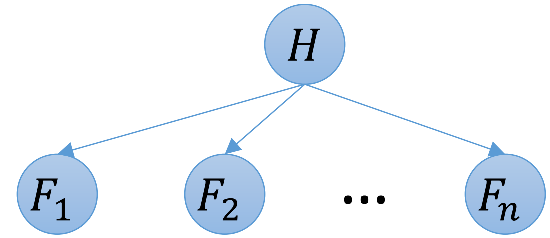
- MAP decision involves estimating

$$\mathrm{argmax}_{h \in H} \; P(f_1, \ldots, \; f_n | h) P(h)$$

- If each feature can take on k values, how many entries are in the joint probability table $P(f_1, \ldots, \; f_n, h)$?

- The table has $O(n^k)$ entries!
  What if we have 1000s of features?

Naïve Bayes Model



- Suppose we have many different types of observations (evidence, symptoms, features) $F_1, \ldots, Fn$ that we want to use to obtain evidence about an underlying hypothesis H
- MAP decision involves estimating

$$\text{argmax}_{\text{h}\in H}\ P(f_1, \ldots,\ f_n|h)P(h)$$

- **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.*** This reduces the joint probability distribution table from $\mathrm{O}(n^k)$ to size $\mathrm{O}(k \times n)$:

$$\text{argmax}_{\text{h}\in H}\ P(h) \prod_{i=1}^{n} P(f_i|h)$$

Example: Naïve Bayes Spam Filter

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.

We need the following:

- A hypothesis H: $spam$ or $\neg$spam
- Define features of the message.
- Estimate the parameters to make a MAP decision which minimizes the classification error (0-1 loss)

Message Features: Bag of Words from NLP

- Model a document as a vector of binary random variables $(W_1, \ldots, W_n)$.
- Each random variable represents if a specific word i is present ($W_i = w_i = 1$) or not ($W_i = w_i = 0$) in the message.
- Simplifications used by bag-of-words:
  - The order of the words in the message is ignored.
  - How often a word is repeated is ignored.

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 Spam Filter Using Words

- We use the simplifying assumption that each word is conditionally independent of the others given the message class (h = spam or not spam):

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

- Now we can calculate the a posteriori probability after the evidence of the message as

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

Note: It is only proportional since we do not divide by $P(w_1, \dots, w_n)$

Naïve Bayes Spam Filter: Model and Decision

- **Model**

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

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

- **Decision:** $argmax_h \;\; P(h|message)$

  that means predict spam if

$$score(spam) > score(\neg spam)$$

- **Needed Data**: $\boldsymbol{P}(H)$ and $\boldsymbol{P}(W_i|H)$

Naïve Bayes Spam Filter: Parameter Estimation

Count in training data:

$$P(H = spam) = \frac{\textit{\# of spam messages} + 1}{\textit{total \# of messages} + \textit{\# of classes}}$$

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

Smoothing for low counts.

Prior $\boldsymbol{P}(H)$

spam:	0.33
¬spam:	0.67

$P(W_i = 1 | H = spam)$

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

$P(W_i = 1 | H = \neg spam)$

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:

$$P(W_i = 0 | H = spam) = 1 - P(W_i = 1 | H = spam)$$
$$P(W_i = 0\ H = \neg spam) = 1 - P(W_i = 1 | H = \neg spam)$$

Summary

Decision theory
To make decisions under uncertainty requires:

1. Estimating probabilities of outcomes for different actions.
2. Assign utility to outcomes.
3. Choose the action with the larges expected utility.

Bayes decision rule
Choose the most likely outcome by minimizing expected 0-1 loss. Required steps:

1. Estimate prior probabilities of outcomes and the likelihood of seeing evidence given different outcomes.
2. Use the evidence to update the probability of the outcome.
3. Apply the MAP decision rule to determine the most likely outcome.

- A general framework for learning functions and decision rules from data is the goal of **Machine Learning**.
- Issue is that we need to define/learn the complete joint probability distribution! Much of ML is about overcoming this issue.

Appendix: A Quick Review of Probability Theory

Random variables

Events

Joint probabilities

Marginal probabilities

Conditional probabilities

Bayes' Rule

Independence

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:

- $P(X = x)$ 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.
- For propositions it means the probability of the set of possible worlds in which the proposition holds.

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* $\boldsymbol{P}(Y)$

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

- **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 $\wedge$ *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

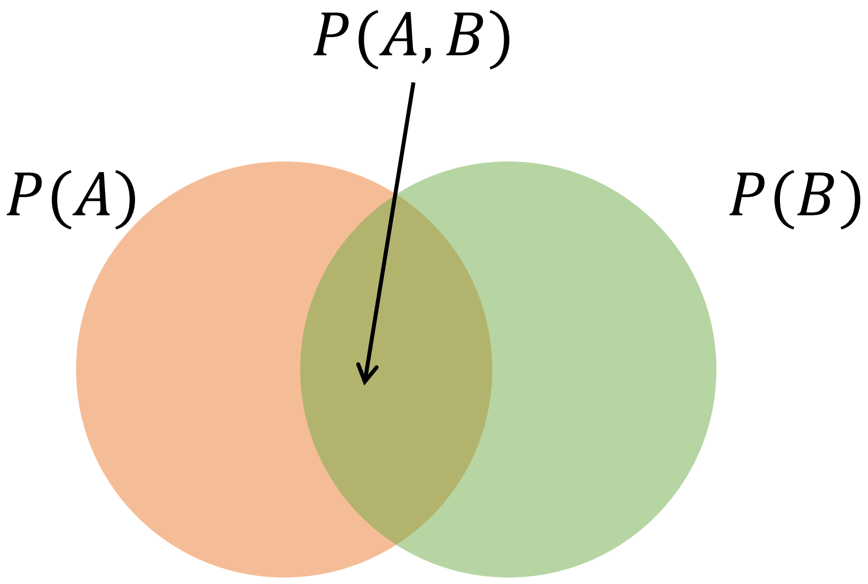
Marginal Prob. Distr.

P(Cavity)	
Cavity = false	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:
  $\mathrm{P(Cavity = true \mid Toothache = true)}$
- 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 $\wedge$ *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	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 $\boldsymbol{P}(X \mid Y = y)$ at once, select all entries in the joint distribution matching $Y = y$ and renormalize them to sum to one.

P(Cavity, Toothache)	
Cavity = false $\wedge$ *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

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