



MACHINE LEARNING

Vasant Honavar

Artificial Intelligence Research Laboratory

Department of Computer Science

Bioinformatics and Computational Biology Program

Center for Computational Intelligence, Learning, & Discovery

Iowa State University

honavar@cs.iastate.edu

www.cs.iastate.edu/~honavar/

www.cild.iastate.edu/

Computer Science

The science of information processing

- The **language** of computation is the best language we have so far for describing how information is encoded, stored, manipulated and used by natural as well as synthetic systems
- Algorithmic or information processing models provide for biological, cognitive, and social sciences what calculus provided for classical physics

Algorithmic explanations of mind

- **Computation: Cognition :: Calculus : Physics**
(Artificial Intelligence, Cognitive Science)
- What are the information requirements of learning?
- What is the algorithmic basis of learning?
- What is the algorithmic basis of rational decision making?
- Can we automate scientific discovery?
- Can we automate creativity?

Algorithms as theories

We will have a theory of

- **Learning** when we have precise information processing models of learning (computer programs that learn from experience)
- **Rational decision** making ...
- **Communication** ...

Conceptual impact of Computer Science

Pre-Turing

- Focus on **physical basis** of the universe with the objective of **explaining all natural phenomena in terms of physical processes**

Post-Turing

- Focus on **informational and algorithmic basis** of the universe with the objective of **explaining natural phenomena in terms of processes that acquire, store, process, manipulate, and use information**

Conceptual impact of Computer Science

We **understand a phenomenon when we can write a computer program that models it at the desired level of detail**

- When theories and explanations in science take the form of algorithms, all sciences become computer science!

Why should Machines Learn?

- Some tasks are best specified by example (e.g., medical diagnosis)
- Buried in large volume of data are useful predictive relationships (data mining)
- The operating environment of certain types of software (user characteristics, distribution of problem instances) may not be completely known at design time
- Environment changes over time – ability of software to adapt to changes would enhance usability

Why study machine learning?

Practical

- Intelligent behavior requires knowledge
- Explicitly specifying the knowledge needed for specific tasks is hard, and often infeasible

Machine Learning is most useful when

- the structure of the task is not well understood but a representative dataset is available
- task (or its parameters) change dynamically

Why study machine learning?

If we can program computers to learn from experience, we can

- Dramatically enhance the usability of software
- Dramatically reduce the cost of software development
- Automate aspects of scientific discovery in emerging data rich domains
 - Bioinformatics
 - Security informatics
 - Social informatics
 - Ecological informatics
 - Enterprise informatics
 - Engineering informatics

Why study machine learning?

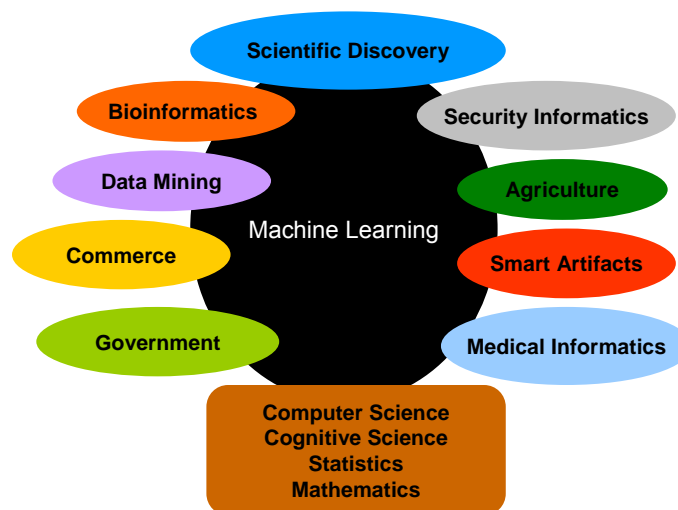
Applications

- Medical diagnosis/image analysis (e.g. pneumonia, pap smears)
- Scientific discovery Spam Filtering, fraud detection (e.g. credit cards, phone calls)
- Search and recommendation (e.g. google, amazon)
- Automatic speech recognition & speaker verification
- Locating/tracking/identifying objects in images & video (e.g. faces)

Why study machine learning?

- Scientific
Information processing models can provide useful insights into
 - How humans and animals learn
 - Information requirements of learning tasks
 - The precise conditions under which certain learning goals are achievable
 - Inherent difficulty of learning tasks
 - How to improve learning – e.g. value of active versus passive learning
 - Computational architectures for learning

Machine Learning in Context



Machine Learning: Contributing Disciplines

- **Computer Science** – Artificial Intelligence, Algorithms and Complexity, Databases, Data Mining
- **Statistics** – Statistical Inference, Experiment Design, Exploratory Data Analysis
- **Mathematics** – Abstract Algebra, Logic, Information Theory, Probability Theory
- **Psychology and Neuroscience** – Behavior, Perception, Learning, Memory, Problem solving
- **Philosophy** – Ontology, Epistemology, Philosophy of Mind, Philosophy of Science

Machine Learning – related disciplines

- Data mining – emphasis on large data sets, computational and memory considerations
- Applied Statistics – applied almost always to small data sets, manually by a statistician sometimes assisted by a computer
- Machine learning – emphasis on automating the discovery of regularities from data, characterizing what can be learned and under what conditions, obtaining guarantees regarding quality of learned models

Machine Learning = (Statistical) Inference + Data Structures + Algorithms

What is learning?

Learning = Inference + Memorization

Inference

$$\forall x \text{ At}(\text{Smoke}, x) \Rightarrow \text{At}(\text{Fire}, x)$$

Deduction



$$\frac{\text{At}(\text{Smoke}, \text{Room 1})}{\text{At}(\text{Fire}, \text{Room 1})}$$

Induction



$$\frac{\begin{array}{l} \text{At}(\text{Smoke}, \text{Room 2}) \wedge \text{At}(\text{Fire}, \text{Room 2}) \\ \text{At}(\text{Smoke}, \text{Room 1}) \wedge \text{At}(\text{Fire}, \text{Room 1}) \\ \text{At}(\text{Ice}, \text{Room 3}) \wedge \neg \text{At}(\text{Fire}, \text{Room 3}) \end{array}}{\forall x \text{ At}(\text{Smoke}, x) \Rightarrow \text{At}(\text{Fire}, x)?}$$

Abduction



$$\frac{\forall x \text{ At}(\text{Smoke}, x) \Rightarrow \text{At}(\text{Fire}, x) \quad \text{At}(\text{Fire}, \text{Room 1})}{\text{At}(\text{Smoke}, \text{Room 1})?}$$

Why study machine learning?

- Applications
- Spam Filtering, fraud detection (e.g. credit cards, phone calls)
- Search and recommendation (e.g. google, amazon)
- Automatic speech recognition & speaker verification
- Printed and handwritten text parsing
- Locating/tracking/identifying objects in images & video (e.g. faces)
- Financial prediction, pricing, volatility analysis
- Medical diagnosis/image analysis (e.g. pneumonia, pap smears)
- Driving computer players in games
- Scientific discovery

Examples of Applications

- Using historical data to improve decisions
 - credit risk assessment, diagnosis, electric power usage prediction
- Using scientific data to acquire knowledge
 - in computational molecular biology
- Software applications that are hard to program
 - autonomous driving
 - face recognition,
 - speech recognition
- Self-customizing programs
 - newsreader that learns user interests

Why study machine learning?

- Scientific
- Information processing models can provide useful insights into
 - How humans and animals learn
 - Information requirements of learning tasks
 - The precise conditions under which certain learning goals are achievable
 - Inherent difficulty of learning tasks
 - How to improve learning – e.g. value of active versus passive learning
 - Computational architectures for learning

What is Machine Learning?

A program M is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance as measured by P on tasks in T in an environment Z improves with experience E .

Example 1

T – cancer diagnosis

E – a set of diagnosed cases

P – accuracy of diagnosis on new cases

Z – noisy measurements, occasionally misdiagnosed training cases

M – a program that runs on a general purpose computer

What is Machine Learning?

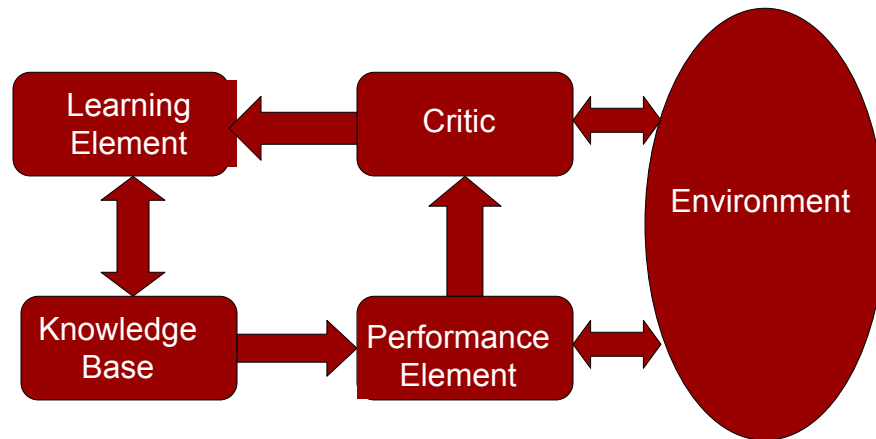
Example 2

T – annotating protein sequences with function labels

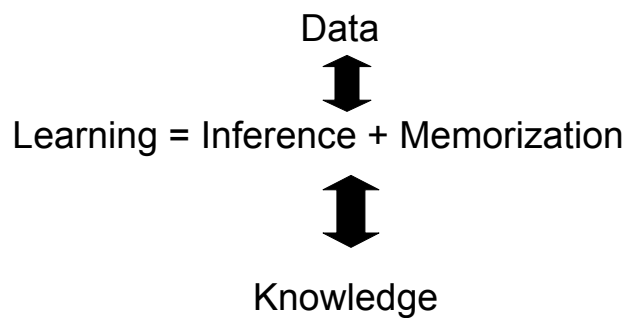
E – a data set of annotated protein sequences

P – score on a test set not seen during training (e.g., accuracy of annotations)

A general framework for learning



Learning



Canonical Learning Problems

Supervised Learning: given examples of inputs and corresponding desired outputs, predict outputs on future inputs.

- Classification
- Regression
- Time series prediction

Unsupervised Learning: given only inputs, automatically discover representations, features, structure, etc.

- Clustering
- Outlier detection
- Compression

Reinforcement Learning

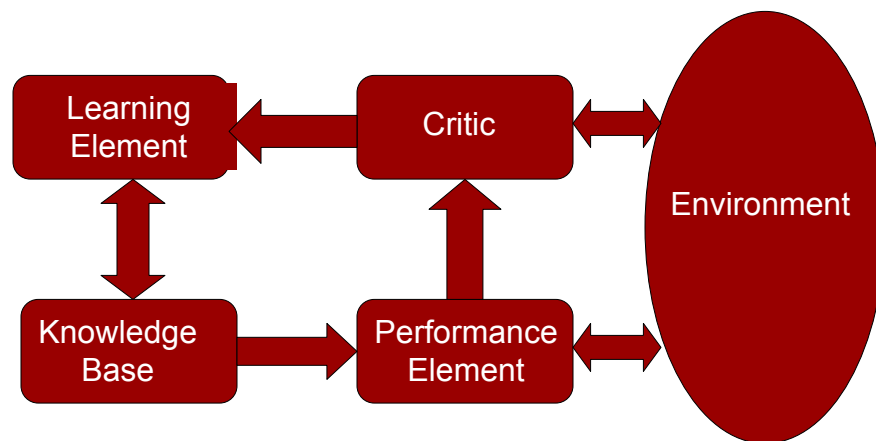
Computational Model of Learning

- **Model of the Learner:** Computational capabilities, sensors, effectors, knowledge representation, inference mechanisms, prior knowledge, etc.
- **Model of the Environment:** Tasks to be learned, information sources (teacher, queries, experiments), performance measures
- **Key questions:** Can a learner with a certain structure learn a specified task in a particular environment? Can the learner do so efficiently? If so, how? If not, why not?

Models of Learning: What are they good for?

- To make explicit relevant aspects of the learner and the environment
- To identify easy and hard learning problems (and the precise conditions under which they are easy or hard)
- To guide the design of learning systems
- To shed light on natural learning systems
- To help analyze the performance of learning systems

A general framework for learning



Designing a learning program for a task

Experience – What experiences are available?

Data – in medical diagnosis, expert diagnosed cases,
feedback

How representative is the experience?

Critic – Can the learner ask questions?

- What type of questions?
- How am I doing? – performance query
- How would you diagnose X? – example based query
- Why was I wrong? – explanation

Designing a learning program for a task

Experience – What experiences are available?

Data – in medical diagnosis, expert diagnosed cases,
feedback

How representative is the experience?

Critic – can the learner ask questions?

What type of questions?

How am I doing? – performance query

How would you diagnose X? – example based query

Why was I wrong? – explanation

Designing a learning program

Performance element –

How is the learned knowledge encoded?

- rules, probabilistic model, programs

How is the learned knowledge used?

- e.g. matching rules, inferring probabilities?

What is the performance measure?

How is performance measured?

- online? batch?

Designing a learning program

Learning element

What is the learning algorithm?

- search for a set of classification rules that are likely to perform well on novel cases (how?)
- estimate a class conditional probability distribution (how?)

Environment

Deterministic or stochastic?

Noisy or noise free?

...

Machine Learning

Learning involves synthesis or adaptation of computational structures e.g., classifiers, grammars, action policies...

**Machine Learning = (Statistical) Inference +
Data Structures + Algorithms**

Learning input – output functions

Target function f – unknown to the learner – $f \in F$

Learner's **hypothesis** about what f might be – $h \in H$

H – hypothesis space

Instance space – X – domain of f, h

Output space – Y – range of f, h

Example – an ordered pair (x, y) where

$$x \in X \quad \text{and} \quad f(x) = y \in Y$$

F and H may or may not be the same!

Training set E – a multi set of examples

Learning algorithm L – a procedure which given some E ,
outputs an $h \in H$

Learning input – output functions

Must choose

Hypothesis language

Instance language

Semantics associated with both

Machines can learn only functions that have *finite descriptions or representations* if we require learning programs to be halting programs

Examples: “Tom likes science fiction horror films”
“F = ma”

Learning from Examples

Premise – A hypothesis (e.g., a classifier) that is consistent with a sufficiently large number of representative training examples is likely to accurately classify novel instances drawn from the same universe

We can prove that this is an optimal approach (under appropriate assumptions)

When the number of examples is limited, the learner needs to be smarter (e.g., find a concise hypothesis that is consistent with the data)

A Simple Learning Scenario

Example – Learning Conjunctive Concepts

Given an arbitrary, noise-free sequence of labeled examples $(X_1, C(X_1)), (X_2, C(X_2)) \dots (X_m, C(X_m))$ of an unknown binary conjunctive concept C over $\{0,1\}^N$, the learner's task is to predict $C(X)$ for a given X .

Online learning of conjunctive concepts

Algorithm A.1

Initialize $L = \{X_1, \sim X_1, \dots, X_N, \sim X_N\}$

Predict according to match between an instance and the conjunction of literals in L

Whenever a mistake is made on a positive example, drop the offending literals from L

Example

$(0111, 1)$ will result in $L = \{\sim X_1, X_2, X_3, X_4\}$

$(1110, 1)$ will yield $L = \{X_2, X_3\}$

Mistake bound analysis of conjunctive concept learning

Theorem: Exact online learning of conjunctive concepts can be accomplished with at most $(N+1)$ prediction mistakes.

Proof Sketch

No literal in C is ever eliminated from L

Each mistake eliminates at least one literal from L

The first mistake eliminates N of the $2N$ literals

Conjunctive concepts can be learned with at most $(N+1)$ mistakes

Conclusion: Conjunctive concepts are *easy* to learn

Bayesian Reasoning, Classification, and Learning Classifiers from Data

Probability is the logic of Science (Jaynes)

Bayesian (subjective) probability provides a basis for updating beliefs based on evidence

By updating beliefs about hypotheses based on data, we can learn about the world.

Bayesian framework provides a sound probabilistic basis for understanding many learning algorithms and designing new algorithms

Bayesian framework provides several practical learning algorithms

Course objectives

- Understand, implement, and use machine learning algorithms to solve practical problems
- Make intelligent choices among learning algorithms for specific applications
- Formulate and solve new machine learning problems combining or adapting elements of existing algorithms
- Analyze learning algorithms (e.g., performance guarantees) and distinguish between easy and hard learning problems
- Gain adequate background to understand current literature
- Gain an understanding of the current state of the art in machine learning
- Learn to conduct original research in machine learning

Course materials

- Text – Pattern Recognition/ Several recommended texts available.
- Assigned readings (~50% of the material) from journals, conference proceedings, lecture notes, etc.
- Lecture outlines and weekly study guide posted on the course web page: <http://www.cs.iastate.edu/~cs573x/>
- Programming language – Java – Must learn on your own if you do not know it
- Software environment – WEKA (Open Source tool for developing machine learning algorithms)

Course mechanics

- Assignments
 - Problem sets, reading and writing assignments
 - Laboratory (implementation) exercises
 - Examinations (2)
 - Term project
- Course staff
 - Vasant Honavar, Professor of Computer Science
 - Oksana Yakhnenko, Ph.D. student in Computer Science
- Office hours etc. See web page for information

Probability. Statistics, and Information Theory

Basic probability theory

Bayes theorem

Random variables

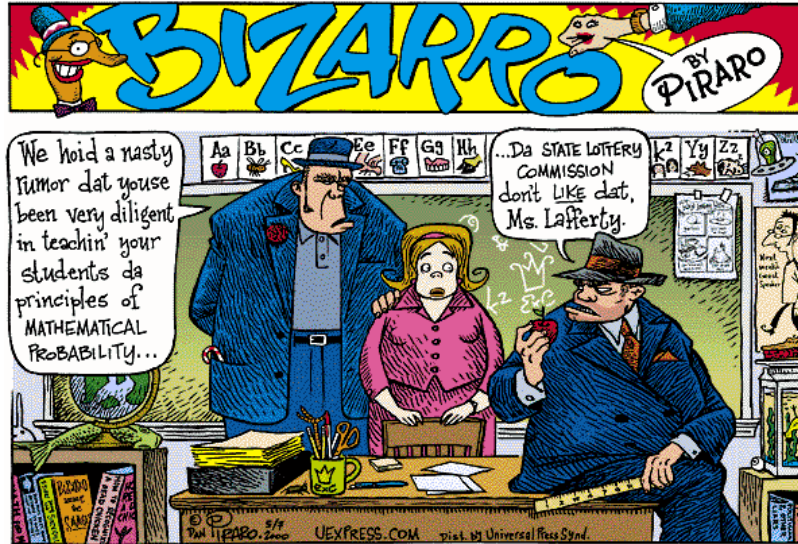
Distributions over random variables

Independence and conditional independence

Statistics and estimation

Basics of Information Theory

A Brief Detour – Probability



Copyright Vasant Honavar, 2006.

Representing and Reasoning under Uncertainty

- Example of reasoning under uncertainty
 - Beliefs:
 - If Oksana studies, there is an 60% chance that she will pass the test; and a 40 percent chance that she will not.
 - If she does not study, there is 20% percent chance that she will pass the test and 80% chance that she will not.
 - Observation: Oksana did not study.
 - Inference task: What is the chance that she will pass the test? What is the chance that she will fail?
- Probability theory generalizes propositional logic
 - Probability theory associates probabilities that lie in the interval $[0, 1]$ as opposed to 0 or 1 (exclusively)

Copyright Vasant Honavar, 2006.

Probability Theory as a Knowledge Representation

- **Ontological commitments** (what do we want to talk about?)
 - Propositions that represent the agent's beliefs about the world
- **Epistemological Commitments** (what can we believe?)
 - What is the *probability* that a given proposition true (given the beliefs and observations)?
- **Syntax**
 - Much like propositional logic
- **Semantics**
 - Relative frequency interpretation
 - Bayesian interpretation
- **Proof Theory**
 - Based on laws of probability

Sources of uncertainty

Uncertainty modeled by Probabilistic assertions may be due to

- In a deterministic world
 - **Laziness**: failure to enumerate exceptions, qualifications, etc. that may be too numerous to state explicitly
 - Sensory limitations
 - **Ignorance**: lack of relevant facts etc.
- In a stochastic world
 - Inherent uncertainty (as in quantum physics)

The framework is agnostic about the source of uncertainty

The world according to Agent Bob

An **atomic event** or **world state** is a **complete specification** of the state of the agent's world.

Event set is a set of mutually exclusive and exhaustive possible world states (relative to an agent's representational commitments and sensing abilities)

From the point of view of an agent Bob who can sense only 3 colors and 2 shapes, the world can be in only one of 6 states

Atomic events (world states) are

- mutually exclusive
- exhaustive

Semantics: Probability as a subjective measure of belief

- Suppose there are 3 agents – Oksana, Cornelia, Jun, in a world where a fair dice has been tossed.
- Oksana observes that the outcome is a “6” and whispers to Cornelia that the outcome is “even” but
- Jun knows nothing about the outcome.

Set of possible mutually exclusive and exhaustive world states = {1, 2, 3, 4, 5, 6}

Set of possible states of the world based on what Cornelia knows = {2, 4, 6}

Probability as a subjective measure of belief

Probability is a measure over all of the world states that are possible, or simply, possible worlds, given what an agent knows

$$\text{Possibleworlds}_{Oksana} = \{6\}, \text{Possibleworlds}_{Cornelia} = \{2,4,6\}$$

$$\text{Possibleworlds}_{Jun} = \{1,2,3,4,5,6\}$$

$$\text{Pr}_{Oksana}(\text{worldstate} = 6) = 1$$

$$\text{Pr}_{Cornelia}(6) = \frac{1}{3}$$

$$\text{Pr}_{Jun}(6) = \frac{1}{6}$$

Oksana, Cornelia, and Jun assign different beliefs to the same world state because of differences in their knowledge

Random variables

- The “domain” of a random variable is the set of values it can take. The values are mutually exclusive and exhaustive.
- The domain of a Boolean random variable X is {true, false} or {1, 0}
- Discrete random variables take values from a countable domain.
 - The domain of the random variable Color may be {Red, Green}.
 - If $E = \{(\text{Red}, \text{Square}), (\text{Green}, \text{Circle}), (\text{Red}, \text{Circle}), (\text{Green}, \text{Square})\}$, the proposition (Color = Red) is True in the world states $\{(\text{Red}, \text{Square}), (\text{Red}, \text{Circle})\}$.
 - Each state of a discrete random variable corresponds to a proposition e.g., (Color = Red)

Syntax

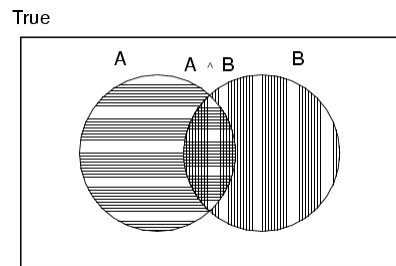
- Basic element: **random variable**
 - Similar to propositional logic: possible worlds defined by assignment of values to random variables.
 - *Cavity* (do I have a cavity?) \square
 - *Weather* is one of $\langle \text{sunny, rainy, cloudy, snow} \rangle$
 - Values must be **exhaustive** and **mutually exclusive**
- Elementary proposition constructed by assignment of a value to a \square random variable
 - *Weather = sunny, Cavity = false* \square (abbreviated as $\neg \text{cavity}$) \square
- Complex propositions formed from elementary propositions and standard logical connectives
 - *Weather = sunny \vee Cavity = false* \square

Syntax and Semantics

- **Atomic event**: A **complete** specification of the state of the world about which the agent is uncertain
- Atomic events correspond to a possible worlds (much like in the case of propositional logic)
 - E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events or 4 possible worlds: \square
 - $Cavity = false \wedge Toothache = false$
 - $Cavity = false \wedge Toothache = true$
 - $Cavity = true \wedge Toothache = false$
 - $Cavity = true \wedge Toothache = true$
- Atomic events are mutually exclusive and exhaustive \square

Axioms of probability

- For any propositions A, B
 - $0 \leq P(A) \leq 1$
 - $P(\text{true}) = 1$ and $P(\text{false}) = 0$
 - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Prior probability

- **Prior or unconditional probabilities** of propositions
 - $P(\text{Cavity} = \text{true}) = 0.1$ and $P(\text{Weather} = \text{sunny}) = 0.72$ correspond to belief prior to arrival of any (new) evidence
- **Probability distribution** gives values for all possible assignments:
 - $\mathbf{P}(\text{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$
 - Note that the probabilities sum to 1
- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables
 - $\mathbf{P}(\text{Cavity}, \text{Play})$ = a 4×2 matrix of values

Prior probability

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables □

– $P(\text{Weather}, \text{Cavity}) =$ a 4×2 matrix of values:

\square		sunny	rainy	cloudy	snow
$\text{Weather} =$					
$\text{Cavity} = \text{true}$		0.144	0.02	0.016	0.02
$\text{Cavity} = \text{false}$		0.576	0.08	0.064	0.08 □

- Every question about a domain can be answered by the joint distribution □

Inference using the joint distribution

	Toothache	\neg Toothache
Cavity	0.4	0.1
\neg Cavity	0.1	0.4

$$P(\text{cavity}) = P(\text{cavity}, \text{ache}) + P(\text{cavity}, \neg \text{ache})$$

Conditional probability

- **Conditional or posterior probabilities** □
 - $P(\text{Cavity} \mid \text{Toothache}) = 0.8$ □
(note *Cavity* is shorthand for *Cavity = True*)
Probability of *Cavity* **given** that *Toothache* □
- Notation for conditional distributions: □
 $\mathbf{P}(\text{Cavity} \mid \text{Toothache}) = 2\text{-element vector of 2-element vectors}$ □
 $P(\text{Cavity} \mid \text{Toothache}, \text{Cavity}) = 1$ □
- New evidence may be irrelevant (Probability of Cavity given Toothache is independent of Weather) □
 $P(\text{Cavity} \mid \text{Toothache}, \text{Sunny}) = P(\text{Cavity} \mid \text{Toothache}) = 0.8$

Conditional probability

- **Definition of conditional probability:** □
 $P(a \mid b) = P(a \wedge b) / P(b)$ if $P(b) > 0$ □
- **Product rule** gives an alternative formulation: □
 $P(a \wedge b) = P(a \mid b) P(b) = P(b \mid a) P(a)$ □

Example:

- Suppose I have two coins – one a normal fair coin, and the other with 2 heads. I pick a coin at *random* and tell you that the side I am looking at is a head. What is the probability that I am looking at a normal coin?

Conditional probability

- A general version holds for whole distributions, e.g., \square

$$\mathbf{P}(\text{Weather}, \text{Cavity}) = \mathbf{P}(\text{Weather} \mid \text{Cavity}) \mathbf{P}(\text{Cavity})$$
- View as a compact notation for a set of 4×2 equations, **not** matrix multiplication \square

- **Chain rule** is derived by successive application of product rule: \square

$$\begin{aligned} \mathbf{P}(X_1, \dots, X_n) &= \mathbf{P}(X_1, \dots, X_{n-1}) \mathbf{P}(X_n \mid X_1, \dots, X_{n-1}) \\ &= \mathbf{P}(X_1, \dots, X_{n-2}) \mathbf{P}(X_{n-1} \mid X_1, \dots, X_{n-2}) \mathbf{P}(X_n \mid X_1, \dots, X_{n-1}) \\ &= \dots \\ &= \prod_i \mathbf{P}(X_i \mid X_1, \dots, X_{i-1}) \quad (i \text{ ranges from } 1 \text{ to } n) \square \end{aligned}$$

Possible worlds semantics

- A possible world is an assignment of Truth values to every simple proposition about the world. Let Ω be a set of possible worlds. Let $\omega \in \Omega$ and let p, q be propositions (atomic sentences or syntactically well formed logical formulae). Then p is True in ω (written $\omega \models p$) where

$\omega \models p$ if ω assigns value *True* to p

$\omega \models p \wedge q$ if $\omega \models p$ and $\omega \models q$

$\omega \models p \vee q$ if $\omega \models p$ or $\omega \models q$ (or both)

$\omega \models \neg p$ if $\omega \not\models p$

Possible Worlds and Random Variables

- A possible world is an assignment of exactly one value to every random variable. Let Ω be a set of possible worlds. Let $\omega \in \Omega$ and let f be a (logical) formula. Then f is True in ω (written $\omega \models f$) where

$\omega \models X = v$ if ω assigns value v to X

$\omega \models f \wedge g$ if $\omega \models f$ and $\omega \models g$

$\omega \models f \vee g$ if $\omega \models f$ or $\omega \models g$ (or both)

$\omega \models \neg f$ if $\omega \not\models f$

Probability as a Measure over Possible worlds

- Associated with each possible world is a measure. When there are only a finite number of possible worlds, the measure of the world ω , denoted by $\mu(\omega)$ has the following properties:

$$\forall \omega \in \Omega, 0 \leq \mu(\omega)$$

$$\sum_{\omega \in \Omega} \mu(\omega) = 1$$

The probability of a formula f , written as $P(f)$, is the sum of the measures of the possible words in which f is True. That is,

$$P(f) = \sum_{\omega \models f} \mu(\omega)$$

Conditional probability as a Measure over Possible worlds not ruled out by evidence

- A given piece of evidence e rules out all possible worlds that are incompatible with e or selects the possible worlds in which e is *True*.

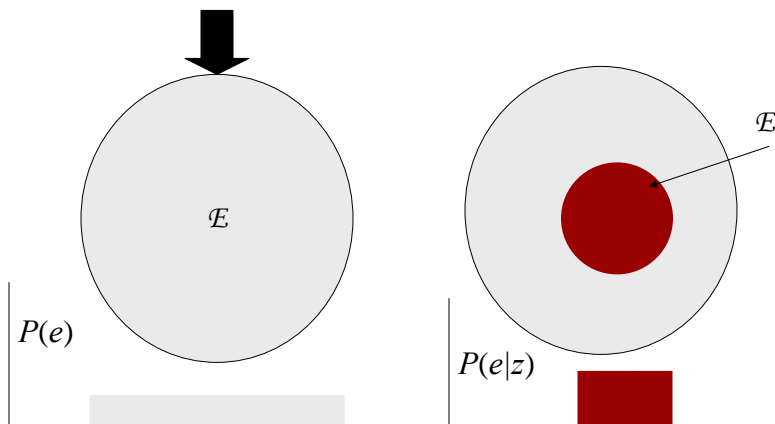
Evidence e induces a new measure μ_e .

$$\mu_e(\omega) = \begin{cases} \frac{1}{P(e)} \mu(\omega) & \text{if } \omega \models e \\ 0 & \text{if } \omega \not\models e \end{cases}$$

$$P(h|e) = \sum_{\omega \models h} \mu_e(\omega) = \frac{1}{P(e)} \sum_{\omega \models h \wedge e} \mu(\omega) = \frac{P(h \wedge e)}{P(e)}$$

Effect of Evidence on Possible worlds

Evidence z e.g., (color = red) rules out some assignments of values to some of the random variables



Evidence redistributes probability mass over possible worlds

- A given piece of evidence z rules out all possible worlds that are incompatible with z or selects the possible worlds in which z is *True*.

Evidence z *induces* a distribution P_z

$$P_z(e) = \begin{cases} \frac{1}{P(z)} P(e) & \text{if } e \models z \\ 0 & \text{if } e \not\models z \end{cases}$$

$$P(h|z) = \sum_{e \models h} P_z(e) = \frac{1}{P(z)} \sum_{e \models h \wedge z} P(e) = \frac{P(h \wedge z)}{P(z)}$$

Defining probability as a Measure over Possible worlds – infinite sets of variables, continuous random variables

$$\forall \omega \in \Omega, 0 \leq \mu(\omega), \int_{\omega} \mu(\omega) d\omega = 1, \quad P(f) = \int_{\omega \models f} \mu(\omega) d\omega$$

When a random variable takes on real values the measure corresponds to a probability density function p . The probability that a random variable X takes values between a and b is given by

$$P(a \leq x \leq b) = \int_a^b p(x) dx$$

This definition can be generalized to handle vector valued random variables

Example:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Note: we now have an infinite set of models

Inference by enumeration

- Start with the joint probability distribution: □

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

□

- For any proposition ϕ , sum the atomic events where it is true: $P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$ □

Inference by enumeration

- Start with the joint probability distribution: □

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

□

- For any proposition ϕ , sum the atomic events where it is true: $P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$ □
- $P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$ □

Inference by enumeration

- Start with the joint probability distribution: □

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

- Can also compute conditional probabilities: □

$$\begin{aligned}
 P(\neg \text{cavity} \mid \text{toothache}) &= \frac{P(\neg \text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\
 &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} \\
 &= 0.4 \square
 \end{aligned}$$

Normalization

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

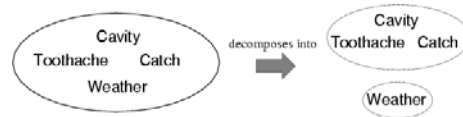
Denominator can be viewed as a **normalization constant** α □

$$\begin{aligned}
 P(\text{Cavity} \mid \text{toothache}) &= \alpha, P(\text{Cavity}, \text{toothache}) \\
 &= \alpha, [P(\text{Cavity}, \text{toothache}, \text{catch}) + P(\text{Cavity}, \text{toothache}, \neg \text{catch})] \\
 &= \alpha, [<0.108, 0.016> + <0.012, 0.064>] \\
 &= \alpha, <0.12, 0.08> = <0.6, 0.4> \square
 \end{aligned}$$

General idea: compute distribution on query variable by fixing **evidence variables** and summing over **unobserved variables**

Independence

- A and B are independent iff
 $P(A/B) = P(A)$ or $P(B/A) = P(B)$ or $P(A, B) = P(A) P(B)$ □



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

- 32 entries reduced to 12; for n independent biased coins, $O(2^n) \rightarrow O(n)$ □
- Absolute independence powerful but rare □
- How can we manage a large numbers of variables? □

Conditional independence

- $P(\textit{Toothache}, \textit{Cavity}, \textit{Catch})$ has $2^3 - 1 = 7$ independent entries □
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache: □
 - $P(\textit{catch} \mid \textit{toothache}, \textit{cavity}) = P(\textit{catch} \mid \textit{cavity})$
- The same independence holds if I haven't got a cavity: □
 - $P(\textit{catch} \mid \textit{toothache}, \neg \textit{cavity}) = P(\textit{catch} \mid \neg \textit{cavity})$ □
- **Catch** is **conditionally independent** of **Toothache** given **Cavity**: □
 - $P(\textit{Catch} \mid \textit{Toothache}, \textit{Cavity}) = P(\textit{Catch} \mid \textit{Cavity})$ □

Conditional independence

- *Catch* is **conditionally independent** of *Toothache* given *Cavity*: \square
 - $P(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = P(\text{Catch} \mid \text{Cavity})$ \square
- Equivalent statements:
 - $P(\text{Toothache} \mid \text{Catch}, \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity})$ \square
 - $P(\text{Toothache}, \text{Catch} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) P(\text{Catch} \mid \text{Cavity})$ \square

Conditional independence

- Write out full joint distribution using chain rule: \square

$$P(\text{Toothache}, \text{Catch}, \text{Cavity})$$

$$= P(\text{Toothache} \mid \text{Catch}, \text{Cavity}) P(\text{Catch}, \text{Cavity})$$

$$= P(\text{Toothache} \mid \text{Catch}, \text{Cavity}) P(\text{Catch} \mid \text{Cavity}) P(\text{Cavity})$$

$$= P(\text{Toothache} \mid \text{Cavity}) P(\text{Catch} \mid \text{Cavity}) P(\text{Cavity})$$

I.e., $2 + 2 + 1 = 5$ independent numbers \square
- **Conditional independence**
 - often reduces the size of the representation of the joint distribution from exponential in n to linear in n \square
 - Is one of the most basic and robust form of knowledge about uncertain environments \square

Conditional Independence

X is **conditionally independent** of Y given Z if the probability distribution governing X is independent of the value of Y given the value of Z :

$P(X|Y, Z) = P(X|Z)$ that is, if

$$(\forall x_i, y_j, z_k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Independence and Conditional Independence

Let Z_1, \dots, Z_n and W be pairwise disjoint sets of random variables on a given event space.

Z_1, \dots, Z_n are mutually independent given W if

$$P(Z_1 \cup \dots \cup Z_n | W) = \prod_{i=1}^n P(Z_i | W)$$

$P(Z_1 | Z_2 \cup W) = P(Z_1 | W)$ if Z_1 and Z_2 are independent.

Note that these represent sets of equations, for all possible value assignments to random variables

Independence Properties of Random Variables

Let W, X, Y, Z be pairwise disjoint sets of random variables on a given event space.

Let $I(X, Y, Z)$ denote that X and Z are *independent* given Y .

That is, $P(X \cup Z | Y) = P(X | Y)P(Z | Y)$, or $P(X | Y \cup Z) = P(X | Y)$. Then :

- $I(X, Z, Y) \Rightarrow I(Y, Z, X)$
- $I(X, Z, Y \cup W) \Rightarrow I(X, Z, Y)$
- $I(X, Z, Y \cup W) \Rightarrow I(X, Z \cup W, Y)$
- $I(X, Z, Y) \wedge I(X, Z \cup Y, W) \Rightarrow I(X, Z, Y \cup W)$

Proof : Follows from definition of *independence*.

Bayes Rule

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

$$\begin{array}{ll}
 P(\text{cancer}) = & P(\neg \text{cancer}) = \\
 P(+ | \text{cancer}) = & P(- | \text{cancer}) = \\
 P(+ | \neg \text{cancer}) = & P(- | \neg \text{cancer}) =
 \end{array}$$

Bayes Rule

Does patient have cancer or not?

$$P(\text{cancer}) = 0.008 \quad P(\neg\text{cancer}) = 0.992$$

$$P(+ | \text{cancer}) = 0.98 \quad P(- | \text{cancer}) = 0.02$$

$$P(+ | \neg\text{cancer}) = 0.03 \quad P(- | \neg\text{cancer}) = 0.97$$

$$P(\text{cancer} | +) = \frac{P(+ | \text{cancer})P(\text{cancer})}{P(+)}; \quad P(\neg\text{cancer} | +) = \frac{P(+ | \neg\text{cancer})P(\neg\text{cancer})}{P(+)}$$

$$P(\text{cancer} | +)P(+) = 0.98 \times 0.008 = 0.0078; \quad P(\neg\text{cancer} | +)P(+) = 0.03 \times 0.992 = 0.0298$$

$$P(+) = 0.0078 + 0.0298$$

$$P(\text{cancer} | +) = 0.21; \quad P(\neg\text{cancer} | +) = 0.79$$

The patient, more likely than not, does not have cancer

Bayes Rule

- Product rule
 - $P(a \wedge b) = P(a | b) P(b) = P(b | a) P(a)$ □
 - Bayes' rule: $P(a | b) = P(b | a) P(a) / P(b)$ □
- In distribution form □

$$P(Y|X) = P(X|Y) P(Y) / P(X) = \alpha P(X|Y) P(Y)$$

Bayes' Rule and conditional independence

$$P(\text{Cavity} \mid \text{toothache} \wedge \text{catch})$$

$$= \alpha P(\text{toothache} \wedge \text{catch} \mid \text{Cavity}) P(\text{Cavity})$$

$$= \alpha P(\text{toothache} \mid \text{Cavity}) P(\text{catch} \mid \text{Cavity}) P(\text{Cavity})$$

- This is an example of a **naïve Bayes (idiot Bayes)** model: \square
 - $P(\text{Cause}, \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i \mid \text{Cause}) \square \square$



- Total number of parameters is **linear** in $n \square$

Summary

- Probability is a rigorous formalism for uncertain knowledge \square
- **Joint probability distribution** specifies probability of every **atomic event**
- Queries can be answered by summing over atomic events \square
- For nontrivial domains, we must find a way to reduce the joint size \square
- **Independence** and **conditional independence** provide the tools \square