"The mind is a neural computer, fitted by natural selection with combinatorial algorithms for causal and probabilistic reasoning about plants, animals, objects, and people."

. . .

"In a universe with any regularities at all, decisions informed about the past are better than decisions made at random. That has always been true, and we would expect organisms, especially informavores such as humans, to have evolved acute intuitions about probability. The founders of probability, like the founders of logic, assumed they were just formalizing common sense."

Steven Pinker, How the Mind Works, 1997, pp. 524, 343.

Admin

- Please do Assignment 0
- Assignment 1A due Wednesday
- I will post readings by Wednesday. Please participate in readings/presentations even if only sitting in.

Today

- Background
- Machine Learning
- Probability, conditioning
- Graphical Models

Learning

Learning Overview

Supervised Learning

Probability

Semantics of Probability

Graphical Models

Learning

Learning is the ability to improve one's behavior based on experience.

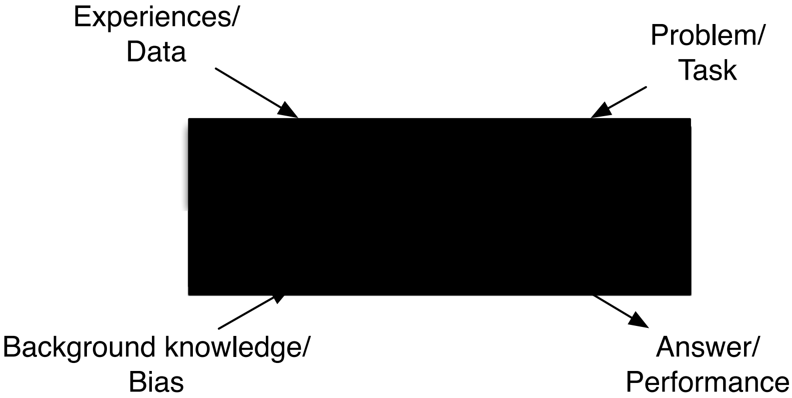
- The range of behaviors is expanded: the agent can do more.
- The accuracy on tasks is improved: the agent can do things better.
- The speed is improved: the agent can do things faster.

Components of a learning problem

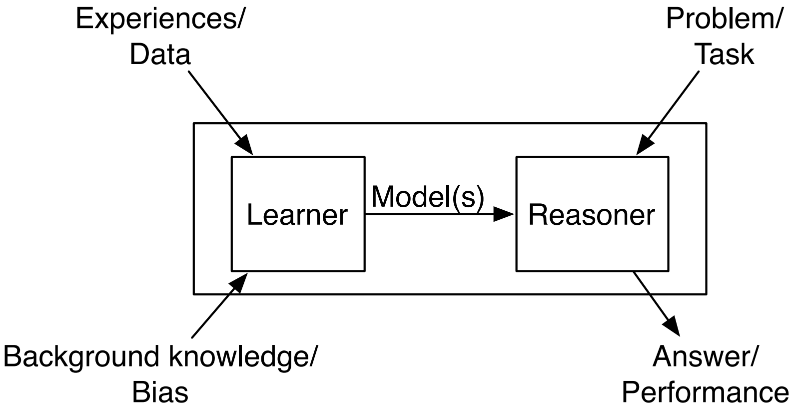
The following components are part of any learning problem:

- task The behavior or task that's being improved.
  For example: classification, acting in an environment
- data The experiences that are being used to improve performance in the task.
- measure of improvement How can the improvement be measured?
  For example: increasing accuracy in prediction, new skills that were not present initially, improved speed.

Black-box Learner



Learning architecture



Common Learning Tasks

- Supervised classification Given a set of pre-classified training examples, classify a new instance.
- Unsupervised learning Find natural classes for examples.
- Reinforcement learning Determine what to do based on rewards and punishments.
- Analytic learning Reason faster using experience.
- Inductive logic programming Build richer models in terms of logic programs.
- Statistical relational learning learning relational representations that also deal with uncertainty.

Example Classification Data

Training Examples:

	Action	Author	Thread	Length	Where
e1	skips	known	new	long	home
e2	reads	unknown	new	short	work
e3	skips	unknown	old	long	work
e4	skips	known	old	long	home
e5	reads	known	new	short	home
e6	skips	known	old	long	work

New Examples:

e7	???	known	new	short	work
e8	???	unknown	new	short	work

We want to classify new examples on feature *Action* based on the examples' *Author*, *Thread*, *Length*, and *Where*.

Feedback

Learning tasks can be characterized by the feedback given to the learner.

- Supervised learning What has to be learned is specified for each example.
- Unsupervised learning No classifications are given; the learner has to discover categories and regularities in the data.
- Reinforcement learning Feedback occurs after a sequence of actions.

Measuring Success

- The measure of success is not how well the agent performs on the training examples, but how well the agent performs for new examples.

Measuring Success

- The measure of success is not how well the agent performs on the training examples, but how well the agent performs for new examples.
- Consider two agents:
  - P claims the negative examples seen are the only negative examples. Every other instance is positive.
  - N claims the positive examples seen are the only positive examples. Every other instance is negative.

Measuring Success

- The measure of success is not how well the agent performs on the training examples, but how well the agent performs for new examples.
- Consider two agents:
  - P claims the negative examples seen are the only negative examples. Every other instance is positive.
  - N claims the positive examples seen are the only positive examples. Every other instance is negative.
- Both agents correctly classify every training example, but disagree on every other example.

Bias

- The tendency to prefer one hypothesis over another is called a bias.
- Saying a hypothesis is better than N's or P's hypothesis isn't something that's obtained from the data.

Bias

- The tendency to prefer one hypothesis over another is called a bias.
- Saying a hypothesis is better than N's or P's hypothesis isn't something that's obtained from the data.
- To have any inductive process make predictions on unseen data, an agent needs a bias.

Bias

- The tendency to prefer one hypothesis over another is called a bias.
- Saying a hypothesis is better than N's or P's hypothesis isn't something that's obtained from the data.
- To have any inductive process make predictions on unseen data, an agent needs a bias.
- What constitutes a good bias is an empirical question about which biases work best in practice.

Learning as search

- Given a representation, data, and a bias, the problem of learning can be reduced to one of search.

Learning as search

- Given a representation, data, and a bias, the problem of learning can be reduced to one of search.
- Learning is search through the space of possible representations looking for the representation or representations that best fits the data, given the bias.

Learning as search

- Given a representation, data, and a bias, the problem of learning can be reduced to one of search.
- Learning is search through the space of possible representations looking for the representation or representations that best fits the data, given the bias.
- These search spaces are typically prohibitively large for systematic search. E.g., use gradient descent.

Learning as search

- Given a representation, data, and a bias, the problem of learning can be reduced to one of search.
- Learning is search through the space of possible representations looking for the representation or representations that best fits the data, given the bias.
- These search spaces are typically prohibitively large for systematic search. E.g., use gradient descent.
- A learning algorithm is made of a search space, an evaluation function, and a search method.

Data

- Data isn't perfect:
  - the features given are inadequate to predict the classification
  - there are examples with missing features
  - some of the features are assigned the wrong value
  - there isn't enough data to determine the correct hypothesis

Data

- Data isn't perfect:
  - the features given are inadequate to predict the classification
  - there are examples with missing features
  - some of the features are assigned the wrong value
  - there isn't enough data to determine the correct hypothesis
- overfitting occurs when distinctions appear in the training data, but not in the unseen examples.

Errors in learning

Errors in learning are caused by:

- Limited representation (representation bias)

Errors in learning

Errors in learning are caused by:

- Limited representation (representation bias)
- Limited search (search bias)

Errors in learning

Errors in learning are caused by:

- Limited representation (representation bias)
- Limited search (search bias)
- Limited data (variance)

Errors in learning

Errors in learning are caused by:

- Limited representation (representation bias)
- Limited search (search bias)
- Limited data (variance)
- Limited features (noise)

Choosing a representation for models

- The richer the representation, the more useful it is for subsequent problem solving.
- The richer the representation, the more difficult it is to learn.

"bias-variance tradeoff"

Characterizations of Learning

- Find the best representation given the data.
- Delineate the class of consistent representations given the data.
- Find a probability distribution of the representations given the data.

Supervised Learning

Given:

- a set of inputs features $X_1, \ldots, X_n$
- a set of target features $Y_1, \ldots, Y_k$
- a set of training examples where the values for the input features and the target features are given for each example
- a new example, where only the values for the input features are given

predict the values for the target features for the new example.

Supervised Learning

Given:

- a set of inputs features $X_1, \ldots, X_n$
- a set of target features $Y_1, \ldots, Y_k$
- a set of training examples where the values for the input features and the target features are given for each example
- a new example, where only the values for the input features are given

predict the values for the target features for the new example.

- classification when the Y_i are discrete
- regression when the Y_i are continuous

Example Data Representations

A travel agent wants to predict the preferred length of a trip, which can be from 1 to 6 days. (No input features).

Example Data Representations

A travel agent wants to predict the preferred length of a trip, which can be from 1 to 6 days. (No input features).

Two representations of the same data:

— Y is the length of trip chosen.

— Each Y_i is an indicator variable that has value 1 if the chosen length is i, and is 0 otherwise.

Example	Y
e_1	1
e_2	6
e_3	6
e_4	2
e_5	1

Example Data Representations

A travel agent wants to predict the preferred length of a trip, which can be from 1 to 6 days. (No input features).

Two representations of the same data:

— Y is the length of trip chosen.

— Each Y_i is an indicator variable that has value 1 if the chosen length is i, and is 0 otherwise.

Example	Y
e_1	1
e_2	6
e_3	6
e_4	2
e_5	1

Example	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
e_1	1	0	0	0	0	0
e_2	0	0	0	0	0	1
e_3	0	0	0	0	0	1
e_4	0	1	0	0	0	0
e_5	1	0	0	0	0	0

What is a prediction?

Evaluating Predictions

Suppose we want to make a prediction of a value for a target feature on example e:

- o_e is the observed value of target feature on example e.
- p_e is the predicted value of target feature on example e.
- The error of the prediction is a measure of how close p_e is to o_e.
- There are many possible errors that could be measured.

Sometimes p_e can be a real number even though o_e can only have a few values.

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$
- sum of squares error $L_2^2(E) = \sum_{e \in E} (o_e - p_e)^2$

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$
- sum of squares error $L_2^2(E) = \sum_{e \in E} (o_e - p_e)^2$
- worst-case error: $L_\infty(E) = \max_{e \in E} |o_e - p_e|$

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$
- sum of squares error $L_2^2(E) = \sum_{e \in E} (o_e - p_e)^2$
- worst-case error: $L_\infty(E) = \max_{e \in E} |o_e - p_e|$
- number wrong: $L_0(E) = \#\{e : o_e \neq p_e\}$

Measures of error

E is the set of examples, with single target feature. For $e \in E$, o_e is observed value and p_e is predicted value:

- absolute error $L_1(E) = \sum_{e \in E} |o_e - p_e|$
- sum of squares error $L_2^2(E) = \sum_{e \in E} (o_e - p_e)^2$
- worst-case error: $L_\infty(E) = \max_{e \in E} |o_e - p_e|$
- number wrong: $L_0(E) = \#\{e : o_e \neq p_e\}$
- A cost-based error takes into account costs of errors.

Measures of error (cont.)

With binary feature: $o_e \in \{0, 1\}$:

- likelihood of the data

$$\prod_{e \in E} p_e^{o_e} (1 - p_e)^{(1 - o_e)}$$

Measures of error (cont.)

With binary feature: $o_e \in \{0, 1\}$:

- likelihood of the data

$$\prod_{e \in E} p_e^{o_e}(1 - p_e)^{(1-o_e)}$$

- log likelihood

$$\sum_{e \in E} (o_e \log p_e + (1 - o_e) \log(1 - p_e))$$

log loss is the negative of log likelihood.

Measures of error (cont.)

With binary feature: $o_e \in \{0, 1\}$:

- likelihood of the data

$$\prod_{e \in E} p_e^{o_e} (1 - p_e)^{(1 - o_e)}$$

- log likelihood

$$\sum_{e \in E} (o_e \log p_e + (1 - o_e) \log(1 - p_e))$$

log loss is the negative of log likelihood.

in terms of bits:

Measures of error (cont.)

With binary feature: $o_e \in \{0, 1\}$:

- likelihood of the data

$$\prod_{e \in E} p_e^{o_e}(1 - p_e)^{(1-o_e)}$$

- log likelihood

$$\sum_{e \in E} (o_e \log p_e + (1 - o_e) \log(1 - p_e))$$

  log loss is the negative of log likelihood.
  in terms of bits: negative of number of bits to encode the data given a code based on p_e.

Information theory overview

- A bit is a binary digit.
- 1 bit can distinguish

Information theory overview

- A bit is a binary digit.
- 1 bit can distinguish 2 items

Information theory overview

- A bit is a binary digit.
- 1 bit can distinguish 2 items
- k bits can distinguish

Information theory overview

- A bit is a binary digit.
- 1 bit can distinguish 2 items
- k bits can distinguish 2^k items

Information theory overview

- A bit is a binary digit.
- 1 bit can distinguish 2 items
- k bits can distinguish 2^k items
- n items can be distinguished using

Information theory overview

- A bit is a binary digit.
- 1 bit can distinguish 2 items
- k bits can distinguish 2^k items
- n items can be distinguished using $\log_2 n$ bits
- Can we do better?

Information and Probability

Consider a code to distinguish elements of $\{a, b, c, d\}$ with

$$P(a) = \frac{1}{2}, P(b) = \frac{1}{4}, P(c) = \frac{1}{8}, P(d) = \frac{1}{8}$$

Consider the code:

a	0	b	10	c	110	d	111

The string *aacabbda* has code

Information and Probability

Consider a code to distinguish elements of $\{a, b, c, d\}$ with

$$P(a) = \frac{1}{2}, P(b) = \frac{1}{4}, P(c) = \frac{1}{8}, P(d) = \frac{1}{8}$$

Consider the code:

a	0	b	10	c	110	d	111

The string *aacabbda* has code 00110010101110.
The code 0111110010100 represents string

Information and Probability

Consider a code to distinguish elements of $\{a, b, c, d\}$ with

$$P(a) = \frac{1}{2}, P(b) = \frac{1}{4}, P(c) = \frac{1}{8}, P(d) = \frac{1}{8}$$

Consider the code:

a	0	b	10	c	110	d	111

The string *aacabbda* has code 00110010101110.
The code 0111110010100 represents string *adcabba*

Information and Probability

Consider a code to distinguish elements of $\{a, b, c, d\}$ with

$$P(a) = \frac{1}{2}, P(b) = \frac{1}{4}, P(c) = \frac{1}{8}, P(d) = \frac{1}{8}$$

Consider the code:

a	0	b	10	c	110	d	111

The string *aacabbda* has code 00110010101110.
The code 0111110010100 represents string *adcabba*
This code uses 1 to 3 bits. On average, it uses

Information and Probability

Consider a code to distinguish elements of $\{a, b, c, d\}$ with

$$P(a) = \frac{1}{2}, P(b) = \frac{1}{4}, P(c) = \frac{1}{8}, P(d) = \frac{1}{8}$$

Consider the code:

a	0	b	10	c	110	d	111

The string *aacabbda* has code 00110010101110.
The code 0111110010100 represents string *adcabba*
This code uses 1 to 3 bits. On average, it uses

$$\begin{aligned} & P(a) \times 1 + P(b) \times 2 + P(c) \times 3 + P(d) \times 3 \\ = \quad & \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = 1\frac{3}{4} \text{ bits.} \end{aligned}$$

Information Content

- To identify x, we need $-\log_2 P(x)$ bits.
- Give a distribution over a set, to a identify a member, the expected number of bits

$$\sum_x -P(x) \times \log_2 P(x).$$

is the information content or entropy of the distribution.

Information Content

- To identify x, we need $-\log_2 P(x)$ bits.
- Give a distribution over a set, to a identify a member, the expected number of bits

$$\sum_x -P(x) \times \log_2 P(x).$$

is the information content or entropy of the distribution.

- The expected number of bits it takes to describe a distribution given evidence e:

$$I(e) = \sum_x -P(x|e) \times \log_2 P(x|e).$$

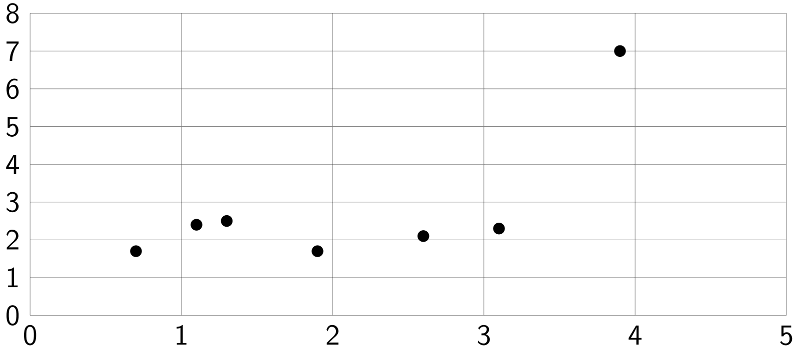
Information Gain

Given a test that can distinguish the cases where α is true from the cases where α is false, the information gain from this test is:

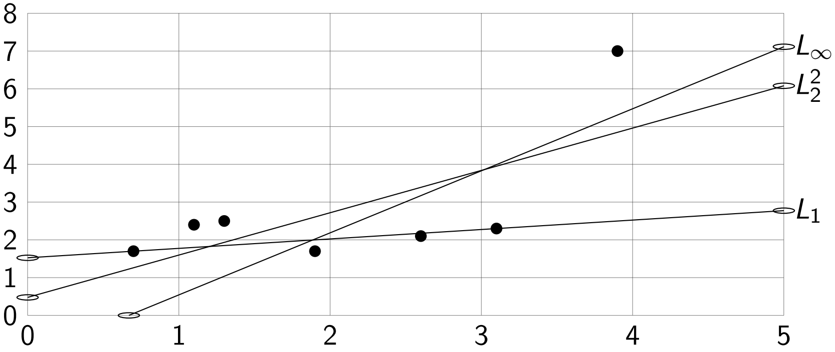
$$I(true) - (P(\alpha) \times I(\alpha) + P(\neg\alpha) \times I(\neg\alpha)).$$

- $I(true)$ is the expected number of bits needed before the test
- $P(\alpha) \times I(\alpha) + P(\neg\alpha) \times I(\neg\alpha)$ is the expected number of bits after the test.

Linear Predictions



Linear Predictions



Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is the mode of Y.

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is the mode of Y.
- The prediction that minimizes the worst-case error on E is

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is the mode of Y.
- The prediction that minimizes the worst-case error on E is $(maximum + minimum)/2$

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is the mode of Y.
- The prediction that minimizes the worst-case error on E is $(maximum + minimum)/2$
- When Y has values $\{0, 1\}$, the prediction that maximizes the likelihood on E is

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is the mode of Y.
- The prediction that minimizes the worst-case error on E is $(maximum + minimum)/2$
- When Y has values $\{0, 1\}$, the prediction that maximizes the likelihood on E is the empirical probability.

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is the mode of Y.
- The prediction that minimizes the worst-case error on E is $(maximum + minimum)/2$
- When Y has values $\{0, 1\}$, the prediction that maximizes the likelihood on E is the empirical probability.
- When Y has values $\{0, 1\}$, the prediction that minimizes the entropy on E is

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is the mode of Y.
- The prediction that minimizes the worst-case error on E is $(maximum + minimum)/2$
- When Y has values $\{0, 1\}$, the prediction that maximizes the likelihood on E is the empirical probability.
- When Y has values $\{0, 1\}$, the prediction that minimizes the entropy on E is the empirical probability.

Point Estimates

To make a single prediction for feature Y, with examples E.

- The prediction that minimizes the sum of squares error on E is the mean (average) value of Y.
- The prediction that minimizes the absolute error on E is the median value of Y.
- The prediction that minimizes the number wrong on E is the mode of Y.
- The prediction that minimizes the worst-case error on E is $(maximum + minimum)/2$
- When Y has values $\{0, 1\}$, the prediction that maximizes the likelihood on E is the empirical probability.
- When Y has values $\{0, 1\}$, the prediction that minimizes the entropy on E is the empirical probability.

But that doesn't mean that these predictions minimize the error for future predictions....

Training and Test Sets

To evaluate how well a learner will work on future predictions, we divide the examples into:

- training examples that are used to train the learner
- test examples that are used to evaluate the learner

...these must be kept separate.

Using Uncertain Knowledge

- Agents don't have complete knowledge about the world.
- Agents need to make decisions based on their uncertainty.
- It isn't enough to assume what the world is like.
  Example: wearing a seat belt.

Using Uncertain Knowledge

- Agents don't have complete knowledge about the world.
- Agents need to make decisions based on their uncertainty.
- It isn't enough to assume what the world is like.
  Example: wearing a seat belt.
- An agent needs to reason about its uncertainty.
- When an agent makes an action under uncertainty, it is gambling $\Longrightarrow$ probability.

Probability

- Probability is an agent's measure of belief in some proposition — subjective probability.

Probability

- Probability is an agent's measure of belief in some proposition — subjective probability.
- An agent's belief depends on its prior belief and what it observes.
- Example: An agent's probability of a particular bird flying
  - Other agents may have different probabilities
  - An agent's belief in a bird's flying ability is affected by what the agent knows about that bird.

Learning

Learning Overview

Supervised Learning

Probability

Semantics of Probability

Graphical Models

Random Variables

- A random variable starts with upper case.
- The range of a variable X, written $range(X)$, is the set of values X can take. (Sometimes use "domain", "frame", "possible values").

Random Variables

- A random variable starts with upper case.
- The range of a variable X, written $range(X)$, is the set of values X can take. (Sometimes use "domain", "frame", "possible values").
- A tuple of random variables $\langle X_1, \ldots, X_n \rangle$ is a complex random variable with range
  $range(X_1) \times \cdots \times range(X_n)$.
  Often the tuple is written as $X_1, \ldots, X_n$.

Random Variables

- A random variable starts with upper case.
- The range of a variable X, written $range(X)$, is the set of values X can take. (Sometimes use "domain", "frame", "possible values").
- A tuple of random variables $\langle X_1, \ldots, X_n \rangle$ is a complex random variable with range
  $range(X_1) \times \cdots \times range(X_n)$.
  Often the tuple is written as $X_1, \ldots, X_n$.
- Assignment $X = x$ means variable X has value x.

Random Variables

- A random variable starts with upper case.
- The range of a variable X, written $range(X)$, is the set of values X can take. (Sometimes use "domain", "frame", "possible values").
- A tuple of random variables $\langle X_1, \ldots, X_n \rangle$ is a complex random variable with range
  $range(X_1) \times \cdots \times range(X_n)$.
  Often the tuple is written as $X_1, \ldots, X_n$.
- Assignment $X = x$ means variable X has value x.
- When ranges are ordered: Inequality $X \leq Y$ means value of X is less than or equal to value of Y.

Random Variables

- A random variable starts with upper case.
- The range of a variable X, written $range(X)$, is the set of values X can take. (Sometimes use "domain", "frame", "possible values").
- A tuple of random variables $\langle X_1, \ldots, X_n \rangle$ is a complex random variable with range
  $range(X_1) \times \cdots \times range(X_n)$.
  Often the tuple is written as $X_1, \ldots, X_n$.
- Assignment $X = x$ means variable X has value x.
- When ranges are ordered: Inequality $X \leq Y$ means value of X is less than or equal to value of Y.
- A proposition is a Boolean formula made from assignments and inequalities.

Possible World Semantics

- A possible world specifies an assignment of one value to each random variable.
- A random variable is a function from possible worlds into the range of the random variable.

Possible World Semantics

- A possible world specifies an assignment of one value to each random variable.
- A random variable is a function from possible worlds into the range of the random variable.
- $\omega \models X = x$
  means variable X is assigned value x in world ω.

Possible World Semantics

- A possible world specifies an assignment of one value to each random variable.
- A random variable is a function from possible worlds into the range of the random variable.
- $\omega \models X = x$
  means variable X is assigned value x in world ω.
- Logical connectives have their standard meaning:

  $\omega \models \alpha \wedge \beta$ if $\omega \models \alpha$ and $\omega \models \beta$

  $\omega \models \alpha \vee \beta$ if $\omega \models \alpha$ or $\omega \models \beta$

  $\omega \models \neg\alpha$ if $\omega \not\models \alpha$
- Let Ω be the set of all possible worlds.

Semantics of Probability

Probability defines a measure on sets of possible worlds.
A probability measure is a function μ from sets of worlds into the non-negative real numbers such that:

- $\mu(\Omega) = 1$
- $\mu(S_1 \cup S_2) = \mu(S_1) + \mu(S_2)$
  if $S_1 \cap S_2 = \{\}$.

Semantics of Probability

Probability defines a measure on sets of possible worlds.

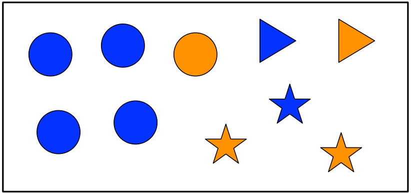
A probability measure is a function μ from sets of worlds into the non-negative real numbers such that:

- $\mu(\Omega) = 1$
- $\mu(S_1 \cup S_2) = \mu(S_1) + \mu(S_2)$
  if $S_1 \cap S_2 = \{\}$.

Then $P(\alpha) = \mu(\{\omega \mid \omega \models \alpha\})$.

Semantics

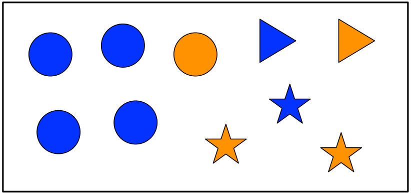
Possible Worlds:



Suppose the measure of each singleton world is 0.1.

Semantics

Possible Worlds:

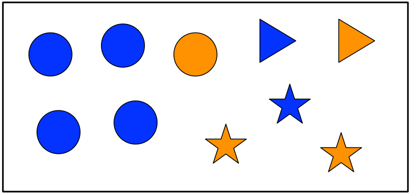


Suppose the measure of each singleton world is 0.1.

- What is the probability of circle?

Semantics

Possible Worlds:

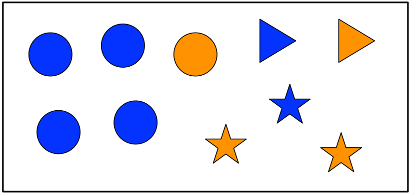


Suppose the measure of each singleton world is 0.1.

- What is the probability of circle?
- What us the probability of star?

Semantics

Possible Worlds:

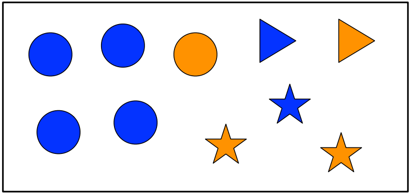


Suppose the measure of each singleton world is 0.1.

- What is the probability of circle?
- What us the probability of star?
- What is the probability of triangle?

Semantics

Possible Worlds:

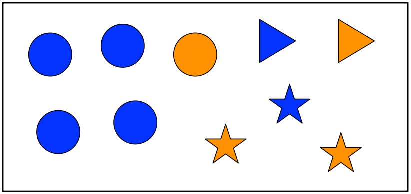


Suppose the measure of each singleton world is 0.1.

- What is the probability of circle?
- What us the probability of star?
- What is the probability of triangle?
- What is the probability of orange?

Semantics

Possible Worlds:

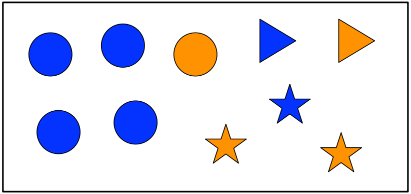


Suppose the measure of each singleton world is 0.1.

- What is the probability of circle?
- What us the probability of star?
- What is the probability of triangle?
- What is the probability of orange?
- What is the probability of blue?

Semantics

Possible Worlds:



Suppose the measure of each singleton world is 0.1.

- What is the probability of circle?
- What us the probability of star?
- What is the probability of triangle?
- What is the probability of orange?
- What is the probability of blue?
- What are the random variables?

Axioms of Probability

Three axioms define what follows from a set of probabilities:

Axiom 1 $0 \leq P(a)$ for any proposition a.

Axiom 2 $P(true) = 1$

Axiom 3 $P(a \vee b) = P(a) + P(b)$ if a and b cannot both be true.

- These axioms are sound and complete with respect to the semantics.

Conditioning

- Probabilistic conditioning specifies how to revise beliefs based on new information.

Conditioning

- Probabilistic conditioning specifies how to revise beliefs based on new information.
- An agent builds a probabilistic model taking all background information into account.
  This gives a prior probability.
- All other information must be conditioned on.
- If evidence e is the all of the information obtained subsequently, the conditional probability $P(h \mid e)$ of h given e is the posterior probability of h.

Semantics of Conditional Probability

- Evidence e rules out possible worlds incompatible with e.

Semantics of Conditional Probability

- Evidence e rules out possible worlds incompatible with e.
- Evidence e induces a new measure, μ_e, over possible worlds

$$\mu_e(S) = \begin{cases} c \times \mu(S) & \text{if } \omega \models e \text{ for all } \omega \in S \\ 0 & \text{if } \omega \not\models e \text{ for all } \omega \in S \end{cases}$$

We can show $c =$

Semantics of Conditional Probability

- Evidence e rules out possible worlds incompatible with e.
- Evidence e induces a new measure, μ_e, over possible worlds

$$\mu_e(S) = \begin{cases} c \times \mu(S) & \text{if } \omega \models e \text{ for all } \omega \in S \\ 0 & \text{if } \omega \not\models e \text{ for all } \omega \in S \end{cases}$$

  We can show $c = \frac{1}{P(e)}$.
- The conditional probability of formula h given evidence e is

$$\begin{aligned} P(h \mid e) &= \mu_e(\{\omega : \omega \models h\}) \\ &= \end{aligned}$$

Semantics of Conditional Probability

- Evidence e rules out possible worlds incompatible with e.
- Evidence e induces a new measure, μ_e, over possible worlds

$$\mu_e(S) = \begin{cases} c \times \mu(S) & \text{if } \omega \models e \text{ for all } \omega \in S \\ 0 & \text{if } \omega \not\models e \text{ for all } \omega \in S \end{cases}$$

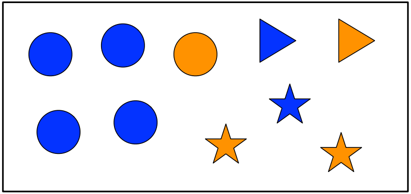
  We can show $c = \frac{1}{P(e)}$.

- The conditional probability of formula h given evidence e is

$$\begin{aligned} P(h \mid e) &= \mu_e(\{\omega : \omega \models h\}) \\ &= \frac{P(h \wedge e)}{P(e)} \end{aligned}$$

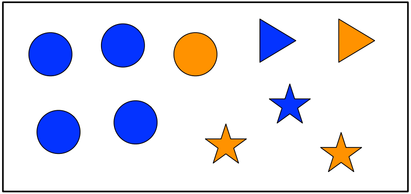
Conditioning

Possible Worlds:



Conditioning

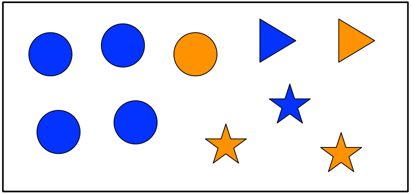
Possible Worlds:



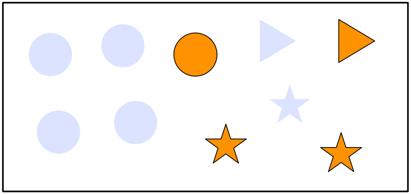
Observe *Color* = *orange*:

Conditioning

Possible Worlds:



Observe *Color* = *orange*:



Exercise

Flu	*Sneeze*	*Snore*	μ
true	true	true	0.064
true	true	false	0.096
true	false	true	0.016
true	false	false	0.024
false	true	true	0.096
false	true	false	0.144
false	false	true	0.224
false	false	false	0.336

What is:

(a) $P(flu \wedge sneeze)$
(b) $P(flu \wedge \neg sneeze)$
(c) $P(flu)$
(d) $P(sneeze \mid flu)$
(e) $P(\neg flu \wedge sneeze)$
(f) $P(flu \mid sneeze)$
(g) $P(sneeze \mid flu \wedge snore)$
(h) $P(flu \mid sneeze \wedge snore)$

Chain Rule

$$P(f_1 \wedge f_2 \wedge \ldots \wedge f_n)$$
$$=$$

Chain Rule

$$
\begin{aligned}
& P(f_1 \wedge f_2 \wedge \ldots \wedge f_n) \\
& \quad = \quad P(f_n \mid f_1 \wedge \cdots \wedge f_{n-1}) \times \\
& \qquad\quad P(f_1 \wedge \cdots \wedge f_{n-1}) \\
& \quad =
\end{aligned}
$$

Chain Rule

$$
\begin{aligned}
P(f_1 \wedge f_2 \wedge \ldots \wedge f_n) \\
&= P(f_n \mid f_1 \wedge \cdots \wedge f_{n-1}) \times \\
&\quad P(f_1 \wedge \cdots \wedge f_{n-1}) \\
&= P(f_n \mid f_1 \wedge \cdots \wedge f_{n-1}) \times \\
&\quad P(f_{n-1} \mid f_1 \wedge \cdots \wedge f_{n-2}) \times \\
&\quad P(f_1 \wedge \cdots \wedge f_{n-2}) \\
&= P(f_n \mid f_1 \wedge \cdots \wedge f_{n-1}) \times \\
&\quad P(f_{n-1} \mid f_1 \wedge \cdots \wedge f_{n-2}) \\
&\quad \times \cdots \times P(f_3 \mid f_1 \wedge f_2) \times P(f_2 \mid f_1) \times P(f_1) \\
&= \prod_{i=1}^{n} P(f_i \mid f_1 \wedge \cdots \wedge f_{i-1})
\end{aligned}
$$

Bayes' theorem

The chain rule and commutativity of conjunction ($h \wedge e$ is equivalent to $e \wedge h$) gives us:

$$P(h \wedge e) \quad =$$

Bayes' theorem

The chain rule and commutativity of conjunction ($h \wedge e$ is equivalent to $e \wedge h$) gives us:

$$P(h \wedge e) = P(h \mid e) \times P(e)$$

Bayes' theorem

The chain rule and commutativity of conjunction ($h \wedge e$ is equivalent to $e \wedge h$) gives us:

$$\begin{aligned} P(h \wedge e) &= P(h \mid e) \times P(e) \\ &= P(e \mid h) \times P(h). \end{aligned}$$

Bayes' theorem

The chain rule and commutativity of conjunction ($h \wedge e$ is equivalent to $e \wedge h$) gives us:

$$\begin{aligned} P(h \wedge e) &= P(h \mid e) \times P(e) \\ &= P(e \mid h) \times P(h). \end{aligned}$$

If $P(e) \neq 0$, divide the right hand sides by $P(e)$:

$$P(h \mid e) =$$

Bayes' theorem

The chain rule and commutativity of conjunction ($h \wedge e$ is equivalent to $e \wedge h$) gives us:

$$\begin{aligned} P(h \wedge e) &= P(h \mid e) \times P(e) \\ &= P(e \mid h) \times P(h). \end{aligned}$$

If $P(e) \neq 0$, divide the right hand sides by $P(e)$:

$$P(h \mid e) = \frac{P(e \mid h) \times P(h)}{P(e)}.$$

This is Bayes' theorem.

Why is Bayes' theorem interesting?

- Often you have causal knowledge:
  $P(\textit{symptom} \mid \textit{disease})$
  $P(\textit{light is off} \mid \textit{status of switches and switch positions})$
  $P(\textit{alarm} \mid \textit{fire})$
  $P(\textit{image looks like}$  $\mid \textit{a tree is in front of a car})$
- and want to do evidential reasoning:
  $P(\textit{disease} \mid \textit{symptom})$
  $P(\textit{status of switches} \mid \textit{light is off and switch positions})$
  $P(\textit{fire} \mid \textit{alarm})$.
  $P(\textit{a tree is in front of a car} \mid \textit{image looks like}$  $)$

Exercise

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as Blue. The court tested the reliability of the witness in the circumstances that existed on the night of the accident and concluded that the witness correctly identifies each one of the two colours 80% of the time and failed 20% of the time.

What is the probability that the cab involved in the accident was Blue?

[From D. Kahneman, Thinking Fast and Slow, 2011, p. 166.]

Conditional independence

Random variable X is independent of random variable Y given random variable(s) Z if,

$$P(X \mid YZ) = P(X \mid Z)$$

Conditional independence

Random variable X is independent of random variable Y given random variable(s) Z if,

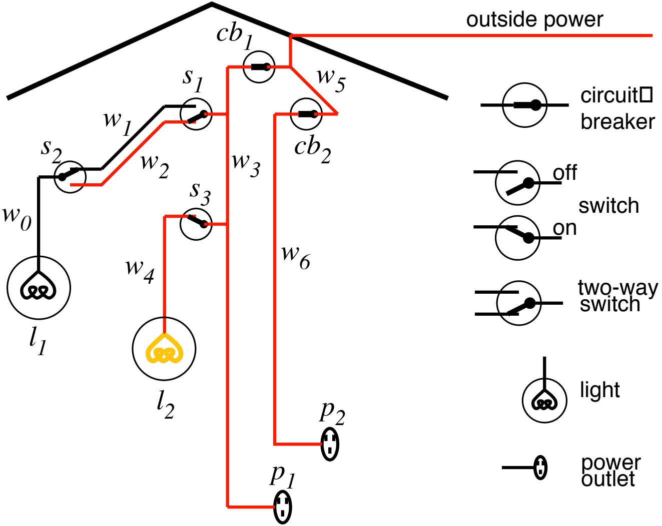
$$P(X \mid YZ) = P(X \mid Z)$$

i.e. for all $x \in dom(X)$, $y, y' \in dom(Y)$, and $z \in dom(Z)$,

$$\begin{aligned} P(X = x \mid Y = y \wedge Z = z) \\ &= P(X = x \mid Y = y' \wedge Z = z) \\ &= P(X = x \mid Z = z). \end{aligned}$$

That is, knowledge of Y's value doesn't affect the belief in the value of X, given a value of Z.

Example domain (diagnostic assistant)



Examples of conditional independence?

- Soppose you know whether there was power in w_1 and whether there was power in w_2 what information is relevant to whether light l_1 is lit? What is independent?

Examples of conditional independence?

- Soppose you know whether there was power in w_1 and whether there was power in w_2 what information is relevant to whether light l_1 is lit? What is independent?
- Whether light $l1$ is lit is independent of the position of light switch $s2$ given what?

Examples of conditional independence?

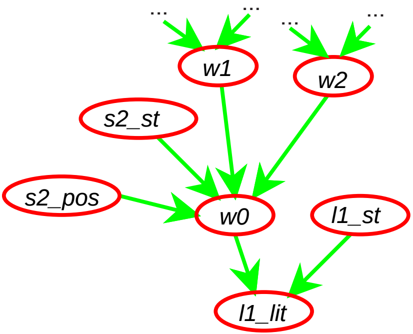
- Soppose you know whether there was power in w_1 and whether there was power in w_2 what information is relevant to whether light l_1 is lit? What is independent?
- Whether light $l1$ is lit is independent of the position of light switch $s2$ given what?
- Every other variable may be independent of whether light $l1$ is lit given

Examples of conditional independence?

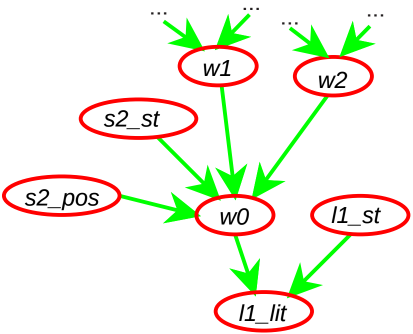
- Soppose you know whether there was power in w_1 and whether there was power in w_2 what information is relevant to whether light l_1 is lit? What is independent?
- Whether light $l1$ is lit is independent of the position of light switch $s2$ given what?
- Every other variable may be independent of whether light $l1$ is lit given whether there is power in wire w_0 and the status of light $l1$ (if it's ok, or if not, how it's broken).

Idea of belief networks

- $l1$ is lit ($L1\_lit$) depends only on the status of the light ($L1\_st$) and whether there is power in wire $w0$.
- In a belief network, $W0$ and $L1\_st$ are parents of $L1\_lit$.
- $W0$ depends only on



Idea of belief networks



- $l1$ is lit ($L1\_lit$) depends only on the status of the light ($L1\_st$) and whether there is power in wire $w0$.
- In a belief network, $W0$ and $L1\_st$ are parents of $L1\_lit$.
- $W0$ depends only on whether there is power in $w1$, whether there is power in $w2$, the position of switch $s2$ ($S2\_pos$), and the status of switch $s2$ ($S2\_st$).

Belief networks

- Totally order the variables of interest: $X_1, \ldots, X_n$
- Theorem of probability theory (chain rule):
  $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$
- The parents *parents*(X_i) of X_i are those predecessors of X_i that render X_i independent of the other predecessors. That is,

Belief networks

- Totally order the variables of interest: $X_1, \ldots, X_n$
- Theorem of probability theory (chain rule):
  $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$
- The parents $parents(X_i)$ of X_i are those predecessors of X_i that render X_i independent of the other predecessors. That is,
  $parents(X_i) \subseteq X_1, \ldots, X_{i-1}$ and
  $P(X_i \mid parents(X_i)) = P(X_i \mid X_1, \ldots, X_{i-1})$
- So $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$
- A belief network is a graph: the nodes are random variables; there is an arc from the parents of each node into that node.

Example: fire alarm belief network

Variables:

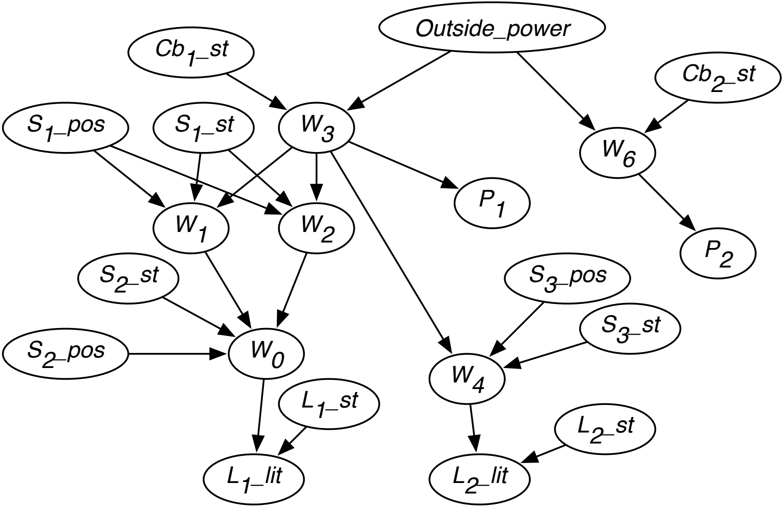
- Fire: there is a fire in the building
- Tampering: someone has been tampering with the fire alarm
- Smoke: what appears to be smoke is coming from an upstairs window
- Alarm: the fire alarm goes off
- Leaving: people are leaving the building *en masse*.
- Report: a colleague says that people are leaving the building *en masse*. (A noisy sensor for leaving.)

Components of a belief network

A belief network consists of:

- a directed acyclic graph with nodes labeled with random variables
- a range for each random variable
- a set of conditional probability tables for each variable given its parents (including prior probabilities for nodes with no parents).

Example belief network



Example belief network (continued)

The belief network also specifies:

- The range of the variables:
  $W_0, \ldots, W_6$ have range $\{live, dead\}$
  S_1_pos, S_2_pos, and S_3_pos have range $\{up, down\}$
  S_1_st has $\{ok, upside\_down, short, intermittent, broken\}$.
- Conditional probabilities, including:
  $P(W_1 = live \mid s_1\_pos = up,\ S_1\_st = ok,\ W_3 = live)$
  $P(W_1 = live \mid s_1\_pos = up,\ S_1\_st = ok,\ W_3 = dead)$
  $P(S_1\_pos = up)$
  $P(S_1\_st = upside\_down)$

Belief network summary

- A belief network is a directed acyclic graph (DAG) where nodes are random variables.
- The parents of a node n are those variables on which n directly depends.
- A belief network is automatically acyclic by construction.
- A belief network is a graphical representation of dependence and independence:
  - A variable is independent of its non-descendants given its parents.

Constructing belief networks

To represent a domain in a belief network, you need to consider:

- What are the relevant variables?
  - What will you observe?
  - What would you like to find out (query)?
  - What other features make the model simpler?
- What values should these variables take?
- What is the relationship between them? This should be expressed in terms of a directed graph, representing how each variable is generated from its predecessors.
- How does the value of each variable depend on its parents? This is expressed in terms of the conditional probabilities.