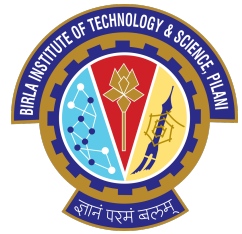
IS-ZC444: ARTIFICIAL INTELLIGENCE

Lecture-13: Machine Learning



Dr. Kamlesh Tiwari
Assistant Professor
Department of Computer Science and Information Systems,
BITS Pilani, Pilani, Jhunjhunu-333031, Rajasthan, INDIA

October 26, 2018 (WILP @ BITS-Pilani Jul-Nov 2018)

Computational Problems

We have problems (the computational ones)

- Sorting: Arranging numbers in ascending/descending order
- Searching: Finding whether an item has specified key
- Determining the existence of Hamiltonian circuit (traversing every vertex once) or Euler walk (through every edge) in a graph

Computational Problems

We have problems (the computational ones)

- Sorting: Arranging numbers in ascending/descending order
- Searching: Finding whether an item has specified key
- Determining the existence of Hamiltonian circuit (traversing every vertex once) or Euler walk (through every edge) in a graph

If we know how to solve the problem, then we could write a program

Computational Problems

We have problems (the computational ones)

- Sorting: Arranging numbers in ascending/descending order
- Searching: Finding whether an item has specified key
- Determining the existence of Hamiltonian circuit (traversing every vertex once) or Euler walk (through every edge) in a graph

If we know how to solve the problem, then we could write a program

But, for some problems we don't precisely know

- Is there a cat in figure? which cat?
- What is written on the board? Which language it is in?
- How to ask for a help from foreigner *etc*.

Either 1) we don't know how to solve, or 2) difficult to specify solution procedure

Computational Problems

We have problems (the computational ones)

- Sorting: Arranging numbers in ascending/descending order
- Searching: Finding whether an item has specified key
- Determining the existence of Hamiltonian circuit (traversing every vertex once) or Euler walk (through every edge) in a graph

If we know how to solve the problem, then we could write a program

But, for some problems we don't precisely know

- Is there a cat in figure? which cat?
- What is written on the board? Which language it is in?
- How to ask for a help from foreigner *etc*.

Either 1) we don't know how to solve, or 2) difficult to specify solution procedure

Then we go for **Machine Learning** (ML)

Machine Learning: Tasks

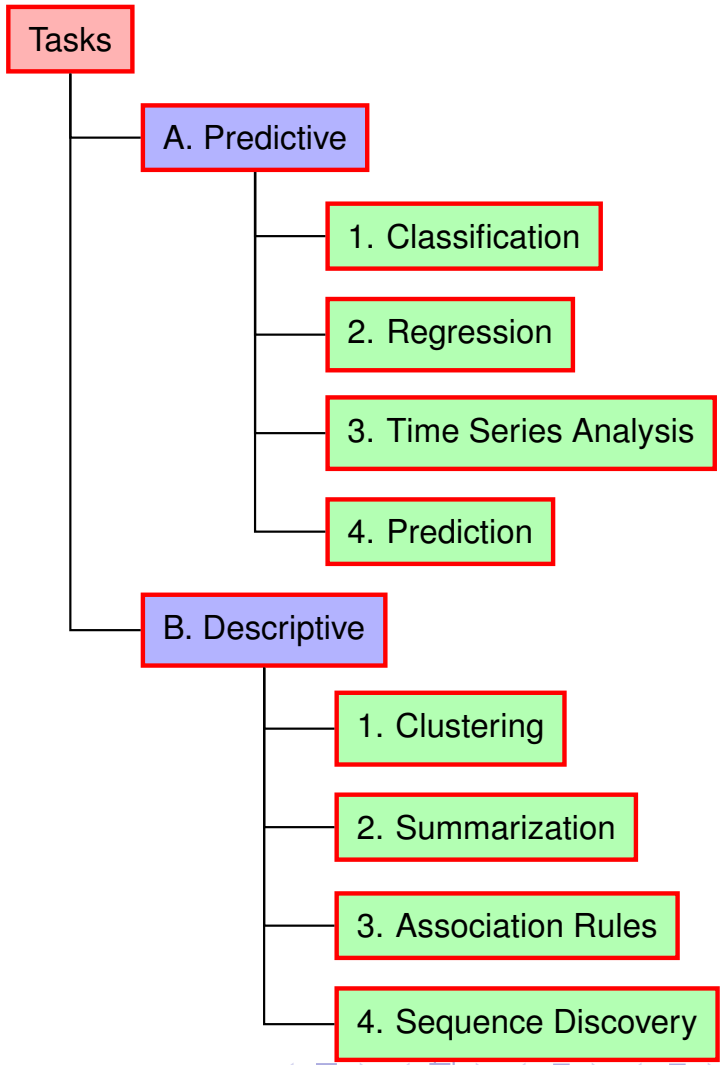
Two broad categories of machine learning models are *Predictive* and *Descriptive*. Some of the related tasks are

Predictive

focuses towards new data item or expanding beyond known facts

Descriptive

focus to understand the available data



Types of Learning

- **Supervised:** “right answers” are provided for sufficient training examples. Computer tells “right answers” for new input. Performance measure. (Classification and regression)
- **Unsupervised:** “right answers” are NOT provided and the computer tries to make sense of the data. How good the spread of items is. (clustering and association rule)
- **Semi-supervised:** “right answers” are provided for few training examples only
- **Active:** computer can ask questions. Needs less training. Opposite is passive learning
- **Lazy:** learner do not consolidate the findings.
- **Reinforced:** hit and trial method to minimize cost. (game playing)
- **Transfer:** Learning a task B to do A. (cycle riding for bike riding)
- **Deep:** processing like human brain

Challenges

Model

Come from philosophy, and we fit its parameters by training and tune while validation. However, actual performance is found during testing.

- How do I choose a model
- How good is the model
- Do I have enough data
- Whether the data is of sufficient quality (there could be error in data, noise in data, missing value)
- How confidence the result is
- Am I describing data correctly (whether features are correct)

Applications of ML

In many domains including finance, robotics, bioinformatics, vision, natural language, *etc*.

- Spam filtering
- Speech/handwriting recognition
- Object detection/recognition
- Weather prediction
- Stock market analysis
- Search engines (e.g, Google)
- Ad placement on websites
- Adaptive website design
- Credit-card fraud detection
- Webpage clustering (e.g.,Google News)
- Machine Translation (e.g., Google Translate)
- Recommendation systems (e.g., Netflix, Amazon)
- Classifying DNA sequences
- Automatic vehicle navigation
- Performance tuning of computer systems
- Predicting good compilation flags for programs
- .. and many more

Building Blocks

- Input: x
- Output: y
- Training data: $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$, ..., $(x^{(m)}, y^{(m)})$
- $x^{(i)}$ could be a multivariate say $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)})$
- Target function: true function

$$f : x \rightarrow y$$

- Hypothesis

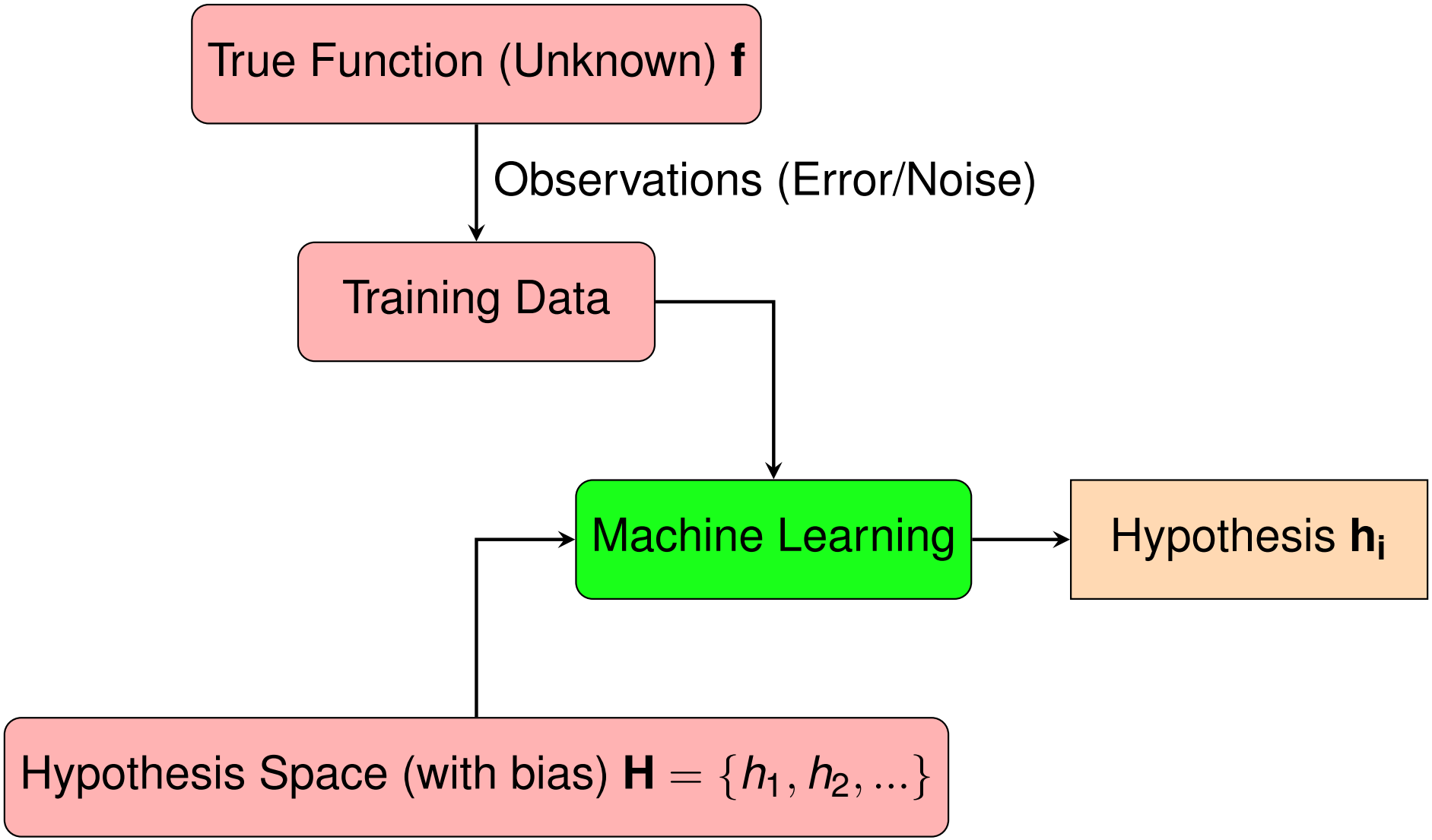
$$h : x \rightarrow y$$

- Accuracy: agreement b/w f and h

Issue is

True function is not known.

The Flow of ML



A Toy model

- **The Problem:** credit approval.
- Input: $x = (x_1, x_2, ..., x_n)$

A Toy model

- **The Problem:** credit approval.
- Input: $x = (x_1, x_2, ..., x_n)$
- Let x_1=accountBal, x_2=Salary, x_3=age ...

A Toy model

- **The Problem:** credit approval.
- Input: $x = (x_1, x_2, ..., x_n)$
- Let x_1=accountBal, x_2=Salary, x_3=age ...
- What weight we should give w_1=0.6, x_2=0.3, x_3=-0.1 ...

A Toy model

- **The Problem:** credit approval.
- Input: $x = (x_1, x_2, ..., x_n)$
- Let x_1=accountBal, x_2=Salary, x_3=age ...
- What weight we should give w_1=0.6, x_2=0.3, x_3=-0.1 ...
- The Model

$$\sum_{i=1}^{n} w_i \times x_i = \begin{cases} > \textit{Threshold} & \text{Then APPROVE} \\ \textit{otherwise} & \text{DENY/REJECT} \end{cases}$$

A Toy model

- **The Problem:** credit approval.
- Input: $x = (x_1, x_2, ..., x_n)$
- Let x_1=accountBal, x_2=Salary, x_3=age ...
- What weight we should give w_1=0.6, x_2=0.3, x_3=-0.1 ...
- The Model

$$\sum_{i=1}^{n} w_i \times x_i = \begin{cases} > Threshold & \text{Then APPROVE} \\ otherwise & \text{DENY/REJECT} \end{cases}$$

- Simplified:

$$h(x) = sign(\sum_{i=1}^{n} w_i \times x_i - Threshold)$$

- Add an extra term x_0 then

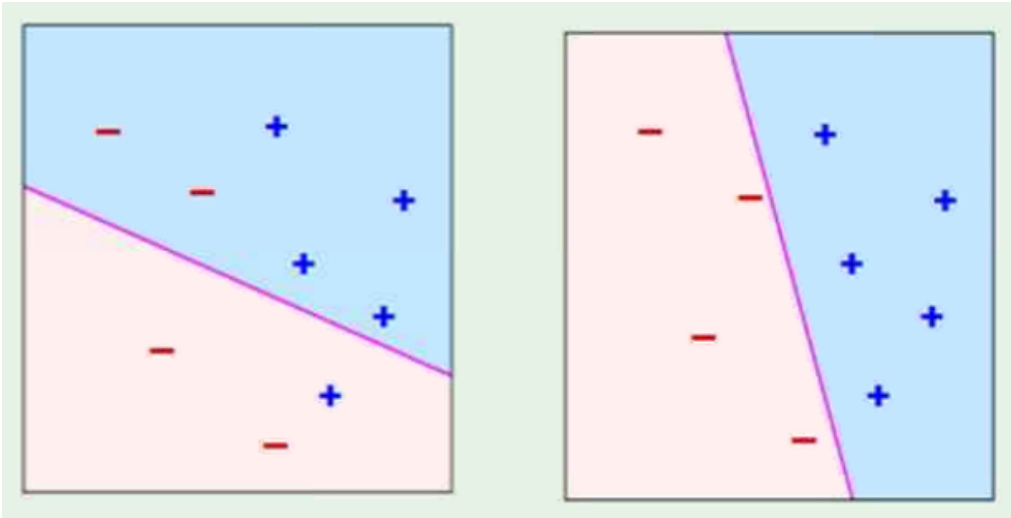
$$h(x) = sign(\sum_{i=0}^{n} w_i \times x_i)$$

A Toy model (Contd..)

- Can you recognize $h(x) = sign(\sum_{i=0}^{n} w_i \times x_i)$

A Toy model (Contd..)

- Can you recognize $h(x) = sign(\sum_{i=0}^{n} w_i \times x_i)$
- It is a linear equation (in two dimension) or hyper plane



- Vector $(w_1, w_2, ..., w_m)$ would be normal on the plane. (why? because dot product is $\cos\theta$)
- What could change this plane? w_i's
- Learning: Use misclassified examples to update $w_i = w_i + \alpha y_i x_i$

Loss function

- Performance is the closeness of hypothesis function with target function
- For example
  - Classification

$$loss(y, h(x)) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{otherwise} \end{cases}$$

  - Regression

$$loss(y, h(x)) = \begin{cases} (h(x) - y)^2 & \text{if } h(x) \neq y \\ 0 & \text{otherwise} \end{cases}$$

Let's change our focus a bit

- Input: x Output: y
- Training data: $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$, ..., $(x^{(m)}, y^{(m)})$
- Target function $f : x \rightarrow y$
- Hypothesis $h : x \rightarrow y$

Let's change our focus a bit

- Input: x Output: y
- Training data: $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$, ..., $(x^{(m)}, y^{(m)})$
- Target function $f : x \rightarrow y$
- Hypothesis $h : x \rightarrow y$

Instead of reporting y let's report $P(y)$; probability of being y

Let's change our focus a bit

- Input: x Output: y
- Training data: $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$, ..., $(x^{(m)}, y^{(m)})$
- Target function $f : x \rightarrow y$
- Hypothesis $h : x \rightarrow y$

Instead of reporting y let's report $P(y)$; probability of being y

- For a given input x, output is not True/False
- But,

$$P(True) = 0.3$$

$$P(False) = 0.7$$

Probability of observing a dataset

Assume you are flipping a biased coin where $p(H) = 0.4$. What is the probability that you see this dataset $D =< H, H, T, T, H, H >$

Probability of observing a dataset

Assume you are flipping a biased coin where $p(H) = 0.4$. What is the probability that you see this dataset $D =< H, H, T, T, H, H >$

- $p(H) = 0.4$
- $p(T) = 1 - p(H) = 1 - 0.4 = 0.6$

Probability of observing a dataset

Assume you are flipping a biased coin where $p(H) = 0.4$. What is the probability that you see this dataset $D =< H, H, T, T, H, H >$

- $p(H) = 0.4$
- $p(T) = 1 - p(H) = 1 - 0.4 = 0.6$
- If **all the trails are independent** then $p(D|\theta)$

Probability of observing a dataset

Assume you are flipping a biased coin where $p(H) = 0.4$. What is the probability that you see this dataset $D =< H, H, T, T, H, H >$

- $p(H) = 0.4$
- $p(T) = 1 - p(H) = 1 - 0.4 = 0.6$
- If **all the trails are independent** then $p(D|\theta)$

$$= p(H) \times p(H) \times p(T) \times p(T) \times p(H) \times p(H)$$

Probability of observing a dataset

Assume you are flipping a biased coin where $p(H) = 0.4$. What is the probability that you see this dataset $D =< H, H, T, T, H, H >$

- $p(H) = 0.4$
- $p(T) = 1 - p(H) = 1 - 0.4 = 0.6$
- If **all the trails are independent** then $p(D|\theta)$

$$= p(H) \times p(H) \times p(T) \times p(T) \times p(H) \times p(H)$$

$$= 0.4^4 \times 0.6^2 = 0.009216$$

Probability of observing a dataset

Assume you are flipping a biased coin where $p(H) = 0.4$. What is the probability that you see this dataset $D =< H, H, T, T, H, H >$

- $p(H) = 0.4$
- $p(T) = 1 - p(H) = 1 - 0.4 = 0.6$
- If **all the trails are independent** then $p(D|\theta)$

$$= p(H) \times p(H) \times p(T) \times p(T) \times p(H) \times p(H)$$

$$= 0.4^4 \times 0.6^2 = 0.009216$$

Note: Order of elements in the data set do not matter in the trial. So $p(< H, H, H, H, T, T >)$ is same (in fact any other permutation)

What is θ

It is the parameter. For our case it represents $p(H) = 0.4$

Hypothesis

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	0	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

Hypothesis

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	0	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- In this example h_1, h_2, ... are hypothesis.
- **Hypothesis** is a function that aims to provide value of the Y

Hypothesis

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	0	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- In this example h_1, h_2, ... are hypothesis.
- **Hypothesis** is a function that aims to provide value of the Y
- Can you identify h_1 and h_2

Hypothesis

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	0	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- In this example h_1, h_2, ... are hypothesis.
- **Hypothesis** is a function that aims to provide value of the Y
- Can you identify h_1 and h_2
- Represent H as candidate set of hypothesis, $i.e. h_i \in H$
- Size of H is at least 2^m

Bayesian Learning

It is based on assumption that quantities of interest are governed by probability distribution

Bayesian Learning

It is based on assumption that quantities of interest are governed by probability distribution

- Notation
  - $P(h)$: initial probability that hypothesis h holds

Bayesian Learning

It is based on assumption that quantities of interest are governed by probability distribution

- Notation
  - $P(h)$: initial probability that hypothesis h holds
  - $P(D)$: probability that data D will be observed

Bayesian Learning

It is based on assumption that quantities of interest are governed by probability distribution

- Notation
  - $P(h)$: initial probability that hypothesis h holds
  - $P(D)$: probability that data D will be observed
  - $P(D|h)$: probability of observing data D given some world in which hypothesis h holds

Bayesian Learning

It is based on assumption that quantities of interest are governed by probability distribution

- Notation
  - $P(h)$: initial probability that hypothesis h holds
  - $P(D)$: probability that data D will be observed
  - $P(D|h)$: probability of observing data D given some world in which hypothesis h holds
  - $P(h|D)$: probability of holding hypothesis h when data D is observed

Bayesian Learning

It is based on assumption that quantities of interest are governed by probability distribution

- Notation
  - $P(h)$: initial probability that hypothesis h holds
  - $P(D)$: probability that data D will be observed
  - $P(D|h)$: probability of observing data D given some world in which hypothesis h holds
  - $P(h|D)$: probability of holding hypothesis h when data D is observed

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Maximum a posteriori (MAP)

- Choose a hypothesis that maximizes $P(h|D)$

$$h_{MAP} = \underset{h \in H}{\text{argmax}}\, P(h|D)$$

Maximum a posteriori (MAP)

- Choose a hypothesis that maximizes $P(h|D)$

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{\text{argmax}}\, P(h|D) \\ &= \underset{h \in H}{\text{argmax}}\, \frac{P(D|h)P(h)}{P(D)} \end{aligned}$$

Maximum a posteriori (MAP)

- Choose a hypothesis that maximizes $P(h|D)$

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{\text{argmax}}\, P(h|D) \\ &= \underset{h \in H}{\text{argmax}}\, \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{\text{argmax}}\, P(D|h)P(h) \end{aligned} \tag{1}$$

Maximum a posteriori (MAP)

- Choose a hypothesis that maximizes $P(h|D)$

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{\text{argmax}}\, P(h|D) \\ &= \underset{h \in H}{\text{argmax}}\, \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{\text{argmax}}\, P(D|h)P(h) \end{aligned} \tag{1}$$

- Because $P(D)$ is independent of h

Maximum a posteriori (MAP)

- Choose a hypothesis that maximizes $P(h|D)$

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{\operatorname{argmax}} P(h|D) \\ &= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h) \end{aligned} \tag{1}$$

- Because $P(D)$ is independent of h
- If all the hypothesis are equally probable, we may further simplify called *maximum likelihood (ML)*

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} P(D|h) \tag{2}$$

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively
- Posterior is (7/9)*(2/50) and (3/9)*(6/50)

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively
- Posterior is (7/9)*(2/50) and (3/9)*(6/50)
- Normalized probabilities are 0.4375 and 0.5625 respectively

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively
- Posterior is (7/9)*(2/50) and (3/9)*(6/50)
- Normalized probabilities are 0.4375 and 0.5625 respectively
- So MAP hypothesis corresponds to h_2

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively
- Posterior is (7/9)*(2/50) and (3/9)*(6/50)
- Normalized probabilities are 0.4375 and 0.5625 respectively
- So MAP hypothesis corresponds to h_2
- Can you guess ML hypothesis?

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively
- Posterior is (7/9)*(2/50) and (3/9)*(6/50)
- Normalized probabilities are 0.4375 and 0.5625 respectively
- So MAP hypothesis corresponds to h_2
- Can you guess ML hypothesis? it is h_1

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively
- Posterior is (7/9)*(2/50) and (3/9)*(6/50)
- Normalized probabilities are 0.4375 and 0.5625 respectively
- So MAP hypothesis corresponds to h_2
- Can you guess ML hypothesis? it is h_1

- **Brute-force MAP learning algorithm:** Evaluates posterior probability for all and returns the one with maximum

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively
- Posterior is (7/9)*(2/50) and (3/9)*(6/50)
- Normalized probabilities are 0.4375 and 0.5625 respectively
- So MAP hypothesis corresponds to h_2
- Can you guess ML hypothesis? it is h_1

- **Brute-force MAP learning algorithm:** Evaluates posterior probability for all and returns the one with maximum
- **Consistent Learner:** learning algorithm is consistent learner if it provides a hypothesis that commits zero error

Bayes Optimal Classifier

Switching the question, from "which is most probable hypothesis?" to "what is the most probable classification of the new instance?"

Bayes Optimal Classifier

Switching the question, from "which is most probable hypothesis?" to "what is the most probable classification of the new instance?"

Is it possible to do better then MAP?

Bayes Optimal Classifier

Switching the question, from "which is most probable hypothesis?" to "what is the most probable classification of the new instance?"

Is it possible to do better then MAP?

Example: Let posterior probabilities of three hypotheses h_1, h_2, h_3 given the training data are 0.4, 0.3, and 0.3 (obviously h_1 is MAP)

Bayes Optimal Classifier

Switching the question, from "which is most probable hypothesis?" to "what is the most probable classification of the new instance?"

Is it possible to do better then MAP?

Example: Let posterior probabilities of three hypotheses h_1, h_2, h_3 given the training data are 0.4, 0.3, and 0.3 (obviously h_1 is MAP)

- Let classification of a new instance x is positive by h_1 and negative by h_2 and h_3

Bayes Optimal Classifier

Switching the question, from "which is most probable hypothesis?" to "what is the most probable classification of the new instance?"

Is it possible to do better then MAP?

Example: Let posterior probabilities of three hypotheses h_1, h_2, h_3 given the training data are 0.4, 0.3, and 0.3 (obviously h_1 is MAP)

- Let classification of a new instance x is positive by h_1 and negative by h_2 and h_3
- By taking all hypotheses into account, the probability that x is positive is 0.4, and negative is 0.6

Bayes Optimal Classifier

Switching the question, from "which is most probable hypothesis?" to "what is the most probable classification of the new instance?"

Is it possible to do better then MAP?

Example: Let posterior probabilities of three hypotheses h_1, h_2, h_3 given the training data are 0.4, 0.3, and 0.3 (obviously h_1 is MAP)

- Let classification of a new instance x is positive by h_1 and negative by h_2 and h_3
- By taking all hypotheses into account, the probability that x is positive is 0.4, and negative is 0.6
  - *Most probable classification is negative and it differs from MAP*

Bayes Optimal Classifier

Switching the question, from "which is most probable hypothesis?" to "what is the most probable classification of the new instance?"

Is it possible to do better then MAP?

Example: Let posterior probabilities of three hypotheses h_1, h_2, h_3 given the training data are 0.4, 0.3, and 0.3 (obviously h_1 is MAP)

- Let classification of a new instance x is positive by h_1 and negative by h_2 and h_3
- By taking all hypotheses into account, the probability that x is positive is 0.4, and negative is 0.6
  - *Most probable classification is negative and it differs from MAP*

Bayes optimal classification:

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

where classification v_j is from V and $P(v_j|D)$ is the correct classification

Bayes Optimal Classifier

$$\boxed{\text{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)}$$

$V = \{\oplus, \ominus\}$

$P(h_1\|D)$=0.4	$P(\ominus\|h_1)$=0	$P(\oplus\|h_1)$=1
$P(h_2\|D)$=0.3	$P(\ominus\|h_2)$=1	$P(\oplus\|h_2)$=0
$P(h_3\|D)$=0.3	$P(\ominus\|h_3)$=1	$P(\oplus\|h_3)$=0

Bayes Optimal Classifier

$$\boxed{\text{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)}$$

$V = \{\oplus, \ominus\}$

$P(h_1\|D)$=0.4	$P(\ominus\|h_1)$=0	$P(\oplus\|h_1)$=1
$P(h_2\|D)$=0.3	$P(\ominus\|h_2)$=1	$P(\oplus\|h_2)$=0
$P(h_3\|D)$=0.3	$P(\ominus\|h_3)$=1	$P(\oplus\|h_3)$=0

Therefore,

$$\sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = 0.4$$

Bayes Optimal Classifier

$$\boxed{\text{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)}$$

$V = \{\oplus, \ominus\}$

$P(h_1\|D)$=0.4	$P(\ominus\|h_1)$=0	$P(\oplus\|h_1)$=1
$P(h_2\|D)$=0.3	$P(\ominus\|h_2)$=1	$P(\oplus\|h_2)$=0
$P(h_3\|D)$=0.3	$P(\ominus\|h_3)$=1	$P(\oplus\|h_3)$=0

Therefore,

$$\sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = 0.4 \qquad \sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = 0.6$$

Bayes Optimal Classifier

$$\boxed{\text{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)}$$

$V = \{\oplus, \ominus\}$

$P(h_1\|D)$=0.4	$P(\ominus\|h_1)$=0	$P(\oplus\|h_1)$=1
$P(h_2\|D)$=0.3	$P(\ominus\|h_2)$=1	$P(\oplus\|h_2)$=0
$P(h_3\|D)$=0.3	$P(\ominus\|h_3)$=1	$P(\oplus\|h_3)$=0

Therefore,

$$\sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = 0.4 \qquad \sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = 0.6$$

and

$$\text{argmax}_{v_j \in \{\oplus, \ominus\}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \ominus$$

Bayes Optimal Classifier

$$\boxed{\text{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)}$$

$V = \{\oplus, \ominus\}$

$P(h_1\|D)$=0.4	$P(\ominus\|h_1)$=0	$P(\oplus\|h_1)$=1
$P(h_2\|D)$=0.3	$P(\ominus\|h_2)$=1	$P(\oplus\|h_2)$=0
$P(h_3\|D)$=0.3	$P(\ominus\|h_3)$=1	$P(\oplus\|h_3)$=0

Therefore,

$$\sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = 0.4 \qquad \sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = 0.6$$

and

$$\text{argmax}_{v_j \in \{\oplus,\ominus\}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \ominus$$

This type of classifier is called a Bayes optimal classifier, or Bayes optimal learner.

Bayes Optimal Classifier

- No other classification method using the same hypothesis space and same prior knowledge can outperform this method on an average.

[1] we could have infinitely many hypothesis

Bayes Optimal Classifier

- No other classification method using the same hypothesis space and same prior knowledge can outperform this method on an average.
- This method maximizes the probability that the new instance is classified correctly, given the available data, hypothesis space, and prior probabilities over the hypotheses

[1] we could have infinitely many hypothesis

Bayes Optimal Classifier

- No other classification method using the same hypothesis space and same prior knowledge can outperform this method on an average.
- This method maximizes the probability that the new instance is classified correctly, given the available data, hypothesis space, and prior probabilities over the hypotheses
- Note that the predictions made by Bayes optimal classifier may not be contained in H

[1] we could have infinitely many hypothesis

Bayes Optimal Classifier

- No other classification method using the same hypothesis space and same prior knowledge can outperform this method on an average.
- This method maximizes the probability that the new instance is classified correctly, given the available data, hypothesis space, and prior probabilities over the hypotheses
- Note that the predictions made by Bayes optimal classifier may not be contained in H

Issue: Although the Bayes optimal classifier obtains the best performance that can be achieved from the given training data, it can be quite costly to apply[1]

[1] we could have infinitely many hypothesis

GIBBS Algorithm [2]

A less optimal method is the Gibbs algorithm

[2] Opper, Manfred and Haussler, David, "Generalization performance of Bayes optimal classification algorithm for learning a perceptron", In Physical Review Letters, 66(20), pp-2677, APS-1991

GIBBS Algorithm [2]

A less optimal method is the Gibbs algorithm

- For a new instance x

[2] Opper, Manfred and Haussler, David, "Generalization performance of Bayes optimal classification algorithm for learning a perceptron", In Physical Review Letters, 66(20), pp-2677, APS-1991

GIBBS Algorithm [2]

A less optimal method is the Gibbs algorithm

- For a new instance x
  1. Choose a hypothesis $h \in H$ at random, according to the posterior probability distribution over H

[2] Opper, Manfred and Haussler, David, "Generalization performance of Bayes optimal classification algorithm for learning a perceptron", In Physical Review Letters, 66(20), pp-2677, APS-1991

GIBBS Algorithm [2]

A less optimal method is the Gibbs algorithm

- For a new instance x
  1. Choose a hypothesis $h \in H$ at random, according to the posterior probability distribution over H
  2. Use h to predict the classification of the instance x

[2] Opper, Manfred and Haussler, David, "Generalization performance of Bayes optimal classification algorithm for learning a perceptron", In Physical Review Letters, 66(20), pp-2677, APS-1991

GIBBS Algorithm [2]

A less optimal method is the Gibbs algorithm

- For a new instance x
  1. Choose a hypothesis $h \in H$ at random, according to the posterior probability distribution over H
  2. Use h to predict the classification of the instance x

Importance

Under certain conditions the expected misclassification error for the Gibbs algorithm is at most twice the expected error of the Bayes optimal classifier

[2] Opper, Manfred and Haussler, David, "Generalization performance of Bayes optimal classification algorithm for learning a perceptron", In Physical Review Letters, 66(20), pp-2677, APS-1991

Naive Bayes Classifier

Bayes classifier is a highly practical Bayesian learning method

Naive Bayes Classifier

Bayes classifier is a highly practical Bayesian learning method

- In some domains, its performance found to be comparable to neural network and decision tree

Naive Bayes Classifier

Bayes classifier is a highly practical Bayesian learning method

- In some domains, its performance found to be comparable to neural network and decision tree
- The Bayesian approach to classify a new instance is to assign the most probable target value describing the instance
  $v_{MAP} = \text{argmax}_{v_j \in V} P(v_j | a_1, a_2, ..., a_n)$

Naive Bayes Classifier

Bayes classifier is a highly practical Bayesian learning method

- In some domains, its performance found to be comparable to neural network and decision tree
- The Bayesian approach to classify a new instance is to assign the most probable target value describing the instance
  $v_{MAP} = \text{argmax}_{v_j \in V} P(v_j | a_1, a_2, ..., a_n)$
- We can use Bayes theorem to rewrite this expression as

$$v_{MAP} = \underset{v_j \in V}{\text{argmax}} \frac{P(a_1, a_2, ..., a_n | v_j) P(v_j)}{P(a_1, a_2, ..., a_n)}$$

Naive Bayes Classifier

Bayes classifier is a highly practical Bayesian learning method

- In some domains, its performance found to be comparable to neural network and decision tree
- The Bayesian approach to classify a new instance is to assign the most probable target value describing the instance
  $v_{MAP} = \text{argmax}_{v_j \in V} P(v_j | a_1, a_2, ..., a_n)$
- We can use Bayes theorem to rewrite this expression as

$$\begin{aligned} v_{MAP} &= \underset{v_j \in V}{\text{argmax}} \frac{P(a_1, a_2, ..., a_n | v_j) P(v_j)}{P(a_1, a_2, ..., a_n)} \\ &= \underset{v_j \in V}{\text{argmax}} P(a_1, a_2, ..., a_n | v_j) P(v_j) \end{aligned} \tag{3}$$

Naive Bayes Classifier

Bayes classifier is a highly practical Bayesian learning method

- In some domains, its performance found to be comparable to neural network and decision tree
- The Bayesian approach to classify a new instance is to assign the most probable target value describing the instance
  $v_{MAP} = \text{argmax}_{v_j \in V} P(v_j | a_1, a_2, ..., a_n)$
- We can use Bayes theorem to rewrite this expression as

$$\begin{aligned} v_{MAP} &= \underset{v_j \in V}{\text{argmax}} \frac{P(a_1, a_2, ..., a_n | v_j) P(v_j)}{P(a_1, a_2, ..., a_n)} \\ &= \underset{v_j \in V}{\text{argmax}} P(a_1, a_2, ..., a_n | v_j) P(v_j) \end{aligned} \tag{3}$$

Naive Bayes has assumption is that the attribute values are conditionally independent given the target value

Naive Bayes Classifier

If attribute values are conditionally independent given the target value

- Under this assumption,
- Given a target value, the probability of observing the conjunction $< a_1, a_2, ..., a_n >$ is just the product of the probabilities.

$$P(a_1, a_2, ..., a_n | v_j) = \Pi_i P(a_i | v_j)$$

Naive Bayes Classifier

If attribute values are conditionally independent given the target value

- Under this assumption,
- Given a target value, the probability of observing the conjunction $< a_1, a_2, ..., a_n >$ is just the product of the probabilities.

$$P(a_1, a_2, ..., a_n | v_j) = \Pi_i P(a_i | v_j)$$

Naive Bayes classifier

is the one which

$$\underset{v_j \in V}{\text{argmax}} P(v_j) \Pi_i P(a_i | v_j)$$

Example: Naive Bayes Classification

Given the data

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rainy	Mild	High	Strong	No

Example: Naive Bayes Classification

Given the data

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rainy	Mild	High	Strong	No

Determine classification for $< \textit{Rainy}, \textit{Hot}, \textit{High}, \textit{Strong} >$

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

P(No) = 5/14

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

P(No) = 5/14

Outlook

	Yes	No
Sunny		

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	

Example: Naive Bayes Classification

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D6	Rainy	Cool	Normal	Strong	No
D8	Sunny	Mild	High	Weak	No
D14	Rainy	Mild	High	Strong	No

Day	Outlook	Temperature	Humidity	Wind	Play
D3	Overcast	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

P(Yes) = 9/14

P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5

Example: Naive Bayes Classification

- P(Yes) = 9/14 P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Example: Naive Bayes Classification

- P(Yes) = 9/14 P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Humidity

	Yes	No
High	3/9	4/5
Low	6/9	1/5

Example: Naive Bayes Classification

- P(Yes) = 9/14 P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Humidity

	Yes	No
High	3/9	4/5
Low	6/9	1/5

Wind

	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

Example: Naive Bayes Classification

- P(Yes) = 9/14 P(No) = 5/14

Outlook

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Humidity

	Yes	No
High	3/9	4/5
Low	6/9	1/5

Wind

	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

Temperature

	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times$
  $P(Hot|Yes) \times P(High|Yes) \times$
  $P(Strong|Yes) \times P(Yes)$

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times$
  $P(Hot|Yes) \times P(High|Yes) \times$
  $P(Strong|Yes) \times P(Yes)$
- $3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14$

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times$
  $P(Hot|Yes) \times P(High|Yes) \times$
  $P(Strong|Yes) \times P(Yes)$
- $3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14$
- 0.005291...

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times$
  $P(Hot|Yes) \times P(High|Yes) \times$
  $P(Strong|Yes) \times P(Yes)$
- $3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14$
- 0.005291...

P(No)

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times P(Hot|Yes) \times P(High|Yes) \times P(Strong|Yes) \times P(Yes)$
- $3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14$
- 0.005291...

P(No)

- $P(x|No) \times P(No)$

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times P(Hot|Yes) \times P(High|Yes) \times P(Strong|Yes) \times P(Yes)$
- $3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14$
- 0.005291...

P(No)

- $P(x|No) \times P(No)$
- $P(Rainy|No) \times P(Hot|No) \times P(High|No) \times P(Strong|No) \times P(No)$

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times P(Hot|Yes) \times P(High|Yes) \times P(Strong|Yes) \times P(Yes)$
- $3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14$
- 0.005291...

P(No)

- $P(x|No) \times P(No)$
- $P(Rainy|No) \times P(Hot|No) \times P(High|No) \times P(Strong|No) \times P(No)$
- $2/5 \times 2/5 \times 4/5 \times 3/5 \times 5/14$

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times P(Hot|Yes) \times P(High|Yes) \times P(Strong|Yes) \times P(Yes)$
- $3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14$
- 0.005291...

P(No)

- $P(x|No) \times P(No)$
- $P(Rainy|No) \times P(Hot|No) \times P(High|No) \times P(Strong|No) \times P(No)$
- $2/5 \times 2/5 \times 4/5 \times 3/5 \times 5/14$
- 0.027428...

Example: Naive Bayes Classification

For $x =< Rainy, Hot, High, Strong >$

P(Yes)

- $P(x|Yes) \times P(Yes)$
- $P(Rainy|Yes) \times P(Hot|Yes) \times P(High|Yes) \times P(Strong|Yes) \times P(Yes)$
- $3/9 \times 2/9 \times 3/9 \times 3/9 \times 9/14$
- 0.005291...

P(No)

- $P(x|No) \times P(No)$
- $P(Rainy|No) \times P(Hot|No) \times P(High|No) \times P(Strong|No) \times P(No)$
- $2/5 \times 2/5 \times 4/5 \times 3/5 \times 5/14$
- 0.027428...

So the classification of x is **No**

Probability

- $P(x, y) = P(x) \times P(y|x)$

Probability

- $P(x, y) = P(x) \times P(y|x)$
- **Independence** of x and y implies $P(y|x) = P(y)$

Probability

- $P(x, y) = P(x) \times P(y|x)$
- **Independence** of x and y implies $P(y|x) = P(y)$
- Then $P(x, y) = P(x) \times P(y)$

Probability

- $P(x, y) = P(x) \times P(y|x)$
- **Independence** of x and y implies $P(y|x) = P(y)$
- Then $P(x, y) = P(x) \times P(y)$
- Bayes Rule

$$P(x|y) = \frac{P(x, y)}{P(y)} = \frac{P(y|x) \times P(x)}{P(y)}$$

Probability

- $P(x, y) = P(x) \times P(y|x)$
- **Independence** of x and y implies $P(y|x) = P(y)$
- Then $P(x, y) = P(x) \times P(y)$
- Bayes Rule

$$P(x|y) = \frac{P(x, y)}{P(y)} = \frac{P(y|x) \times P(x)}{P(y)}$$

- **Marginal:** distribution of a single variable x can be obtained from a given joint distribution $p(x, y)$ by

$$p(x) = \sum_{y} p(x, y)$$

Probability

- $P(x, y) = P(x) \times P(y|x)$
- **Independence** of x and y implies $P(y|x) = P(y)$
- Then $P(x, y) = P(x) \times P(y)$
- Bayes Rule

$$P(x|y) = \frac{P(x, y)}{P(y)} = \frac{P(y|x) \times P(x)}{P(y)}$$

- **Marginal:** distribution of a single variable x can be obtained from a given joint distribution $p(x, y)$ by

$$p(x) = \sum_{y} p(x, y)$$

- The process of computing a marginal from a joint distribution is called marginalisation.

$$p(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = \sum_{x_i} p(x_1, x_2, ..., x_n)$$

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: Soft XOR

A	B	$p(C{=}1|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: Soft XOR

A	B	$p(C{=}1|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

If $p(A{=}1) = 0.65$, $p(B{=}1) = 0.77$
Determine $p(A{=}1|C{=}0)$

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: Soft XOR

A	B	$p(C{=}1|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

If $p(A{=}1) = 0.65$, $p(B{=}1) = 0.77$
Determine $p(A{=}1|C{=}0)$

- $p(A{=}1, C{=}0)$

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: Soft XOR

A	B	$p(C{=}1|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

If $p(A{=}1) = 0.65$, $p(B{=}1) = 0.77$
Determine $p(A{=}1|C{=}0)$

- $p(A{=}1, C{=}0) = \sum_B p(A = 1, B, C = 0)$

$$\begin{aligned} &= \sum_B p(C = 0|A = 1, B)p(A = 1)p(B) \\ &= p(C = 0|A = 1, B = 0)p(A = 1)p(B = 0) \\ &\quad + p(C = 0|A = 1, B = 1)p(A = 1)p(B = 1) \\ &= 0.2 \times 0.65 \times 0.23 + 0.75 \times 0.65 \times 0.77 \\ &= \boxed{0.405} \end{aligned}$$

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: Soft XOR

A	B	$p(C{=}1|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

If $p(A{=}1) = 0.65$, $p(B{=}1) = 0.77$
Determine $p(A{=}1|C{=}0)$

- $p(A{=}1, C{=}0) = \sum_B p(A = 1, B, C = 0)$

$$\begin{aligned} &= \sum_B p(C = 0|A = 1, B)p(A = 1)p(B) \\ &= p(C = 0|A = 1, B = 0)p(A = 1)p(B = 0) \\ &\quad + p(C = 0|A = 1, B = 1)p(A = 1)p(B = 1) \\ &= 0.2 \times 0.65 \times 0.23 + 0.75 \times 0.65 \times 0.77 \\ &= \boxed{0.405} \end{aligned}$$

- Similarly $p(A{=}0, C{=}0)$

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: Soft XOR

A	B	$p(C{=}1|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

If $p(A{=}1) = 0.65$, $p(B{=}1) = 0.77$
Determine $p(A{=}1|C{=}0)$

- $p(A{=}1, C{=}0) = \sum_B p(A = 1, B, C = 0)$

$$\begin{aligned} &= \sum_B p(C = 0|A = 1, B)p(A = 1)p(B) \\ &= p(C = 0|A = 1, B = 0)p(A = 1)p(B = 0) \\ &\quad + p(C = 0|A = 1, B = 1)p(A = 1)p(B = 1) \\ &= 0.2 \times 0.65 \times 0.23 + 0.75 \times 0.65 \times 0.77 \\ &= \boxed{0.405} \end{aligned}$$

- Similarly $p(A{=}0, C{=}0) = 0.075$

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: Soft XOR

A	B	$p(C{=}1|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

If $p(A{=}1) = 0.65$, $p(B{=}1) = 0.77$
Determine $p(A{=}1|C{=}0)$

- $p(A{=}1, C{=}0) = \sum_B p(A = 1, B, C = 0)$

$$\begin{aligned} &= \sum_B p(C = 0|A = 1, B)p(A = 1)p(B) \\ &= p(C = 0|A = 1, B = 0)p(A = 1)p(B = 0) \\ &\quad + p(C = 0|A = 1, B = 1)p(A = 1)p(B = 1) \\ &= 0.2 \times 0.65 \times 0.23 + 0.75 \times 0.65 \times 0.77 \\ &= \boxed{0.405} \end{aligned}$$

- Similarly $p(A{=}0, C{=}0) = 0.075$
- $p(A{=}1|C{=}0) = \frac{p(A=1, C=0)}{p(C=0)}$

Marginalisation

Conditional Independence when two variable are independent of each other provided we know state of some other variable

$$P(x, y|z) = P(x|z) \times P(y|z)$$

Consider: Soft XOR

A	B	$p(C{=}1|A, B)$
0	0	0.10
0	1	0.99
1	0	0.80
1	1	0.25

If $p(A{=}1) = 0.65$, $p(B{=}1) = 0.77$
Determine $p(A{=}1|C{=}0)$

- $p(A{=}1, C{=}0) = \sum_B p(A = 1, B, C = 0)$

$$\begin{aligned} &= \sum_B p(C = 0|A = 1, B)p(A = 1)p(B) \\ &= p(C = 0|A = 1, B = 0)p(A = 1)p(B = 0) \\ &\quad + p(C = 0|A = 1, B = 1)p(A = 1)p(B = 1) \\ &= 0.2 \times 0.65 \times 0.23 + 0.75 \times 0.65 \times 0.77 \\ &= \boxed{0.405} \end{aligned}$$

- Similarly $p(A{=}0, C{=}0) = 0.075$
- $p(A{=}1|C{=}0) = \frac{p(A=1,C=0)}{p(C=0)} = \frac{p(A=1,C=0)}{p(A=1,C=0)+p(A=0,C=0)}$
  $= \boxed{0.843}$

Let's see this

$$\sum_{J} \left(p(J|R) \times f(R) \right) =$$

Let's see this

$$\sum_J \left(p(J|R) \times f(R)\right) = f(R)$$

Proof:

$$\begin{aligned}
\sum_J \left(p(J|R) \times f(R)\right) &= \sum_J \left(\frac{p(J,R)}{p(R)} \times f(R)\right) \\
&= \frac{\sum_J \left(p(J,R) \times f(R)\right)}{p(R)} \\
&= \frac{f(R) \times \sum_J p(J,R)}{p(R)} \\
&= \frac{f(R) \times p(R)}{p(R)} \\
&= f(R)
\end{aligned}$$

Belief network (a graphical model)

- A **belief network** introduces structure into a probabilistic model by using graphs to represent independence assumptions among the variables

Belief network (a graphical model)

- A **belief network** introduces structure into a probabilistic model by using graphs to represent independence assumptions among the variables
- Independently specifying all the attributed is overkill

Belief network (a graphical model)

- A **belief network** introduces structure into a probabilistic model by using graphs to represent independence assumptions among the variables
- Independently specifying all the attributed is overkill
- With distribution of n attributes, marginal for one takes $O(2^{n-1})$

Belief network (a graphical model)

- A **belief network** introduces structure into a probabilistic model by using graphs to represent independence assumptions among the variables
- Independently specifying all the attributed is overkill
- With distribution of n attributes, marginal for one takes $O(2^{n-1})$
- By constraining variable interaction (specifying independence) one can get the form like

$$p(x_1, x_2, ..., x_{100}) = \Pi_{i=1}^{99} \phi(x_i, x_{i+1})$$

Belief network (a graphical model)

- A **belief network** introduces structure into a probabilistic model by using graphs to represent independence assumptions among the variables
- Independently specifying all the attributed is overkill
- With distribution of n attributes, marginal for one takes $O(2^{n-1})$
- By constraining variable interaction (specifying independence) one can get the form like

$$p(x_1, x_2, ..., x_{100}) = \Pi_{i=1}^{99} \phi(x_i, x_{i+1})$$

- Belief networks are a convenient framework for representing such independence assumptions

Belief network (a graphical model)

- A **belief network** introduces structure into a probabilistic model by using graphs to represent independence assumptions among the variables
- Independently specifying all the attributed is overkill
- With distribution of n attributes, marginal for one takes $O(2^{n-1})$
- By constraining variable interaction (specifying independence) one can get the form like

$$p(x_1, x_2, ..., x_{100}) = \Pi_{i=1}^{99} \phi(x_i, x_{i+1})$$

- Belief networks are a convenient framework for representing such independence assumptions
- Belief networks are also called as *Bayes' Networks* or *Bayesian Belief Networks*

Modeling Independencies

One morning Tracey leaves her house and realities that her grass is wet. Is it due to overnight rain or did she forget to turn off the sprinkler last night? Next she notices that the grass of her neighbor, Jack, is also wet.

Modeling Independencies

One morning Tracey leaves her house and realities that her grass is wet. Is it due to overnight rain or did she forget to turn off the sprinkler last night? Next she notices that the grass of her neighbor, Jack, is also wet.

(R=1) $\rightarrow$ rain last night,
(S=1) $\rightarrow$ sprinkler on last night,
(J=1) $\rightarrow$ Jack's grass is wet,
(T=1) $\rightarrow$ Traceya's Grass is wet

Modeling Independencies

One morning Tracey leaves her house and realities that her grass is wet. Is it due to overnight rain or did she forget to turn off the sprinkler last night? Next she notices that the grass of her neighbor, Jack, is also wet.

(R=1) $\rightarrow$ rain last night,
(S=1) $\rightarrow$ sprinkler on last night,
(J=1) $\rightarrow$ Jack's grass is wet,
(T=1) $\rightarrow$ Traceya's Grass is wet

- Model of Traceya's world involves probability distribution on T, J, R, S that has $2^4 = 16$ states

Modeling Independencies (contd..)

- However we know

$$p(T, J, R, S) = p(T|J, R, S)p(J, R, S)$$

Modeling Independencies (contd..)

- However we know

$$\begin{aligned} p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\ &= p(T|J, R, S)p(J|R, S)p(R, S) \end{aligned}$$

Modeling Independencies (contd..)

- However we know

$$
\begin{aligned}
p(T, J, R, S) &= p(T|J, R, S)p(J, R, S) \\
&= p(T|J, R, S)p(J|R, S)p(R, S) \\
&= p(T|J, R, S)p(J|R, S)p(R|S)p(S)
\end{aligned}
$$

Modeling Independencies (contd..)

- However we know

$$\begin{aligned} p(T,J,R,S) &= p(T|J,R,S)p(J,R,S) \\ &= p(T|J,R,S)p(J|R,S)p(R,S) \\ &= p(T|J,R,S)p(J|R,S)p(R|S)p(S) \end{aligned}$$

- Computation of $p(T|J,R,S)$ requires us to specify $2^3 = 8$ values

Modeling Independencies (contd..)

- However we know

$$\begin{aligned} p(T,J,R,S) &= p(T|J,R,S)p(J,R,S) \\ &= p(T|J,R,S)p(J|R,S)p(R,S) \\ &= p(T|J,R,S)p(J|R,S)p(R|S)p(S) \end{aligned}$$

- Computation of $p(T|J,R,S)$ requires us to specify $2^3 = 8$ values
- With $p(T=1|J,R,S)$, one can use normalization to compute $p(T=0|J,R,S)$ as $1 - p(T=1|J,R,S)$

Modeling Independencies (contd..)

- However we know

$$\begin{aligned} p(T,J,R,S) &= p(T|J,R,S)p(J,R,S) \\ &= p(T|J,R,S)p(J|R,S)p(R,S) \\ &= p(T|J,R,S)p(J|R,S)p(R|S)p(S) \end{aligned}$$

- Computation of $p(T|J,R,S)$ requires us to specify $2^3 = 8$ values
- With $p(T=1|J,R,S)$, one can use normalization to compute $p(T=0|J,R,S)$ as $1 - p(T=1|J,R,S)$
- Computation of other factors would also need 4+2+1 values
- Total we need 8+4+2+1=15 values

Conditional Independence

- We may assume that Traceya's grass is wet depends only directly on whether or not it has been raining and whether or not her sprinkler was on so $p(T|J, R, S) = p(T|R, S)$

Conditional Independence

- We may assume that Traceya's grass is wet depends only directly on whether or not it has been raining and whether or not her sprinkler was on so $p(T|J,R,S) = p(T|R,S)$
- Assume that Jack's grass is wet is influenced only directly by whether or not it has been raining $p(J|R,S) = p(J|R)$

Conditional Independence

- We may assume that Traceya's grass is wet depends only directly on whether or not it has been raining and whether or not her sprinkler was on so $p(T|J,R,S) = p(T|R,S)$
- Assume that Jack's grass is wet is influenced only directly by whether or not it has been raining $p(J|R,S) = p(J|R)$
- Furthermore, we assume the rain is not directly influenced by the sprinkler $p(R|S) = p(R)$

Conditional Independence

- We may assume that Traceya's grass is wet depends only directly on whether or not it has been raining and whether or not her sprinkler was on so $p(T|J,R,S) = p(T|R,S)$
- Assume that Jack's grass is wet is influenced only directly by whether or not it has been raining $p(J|R,S) = p(J|R)$
- Furthermore, we assume the rain is not directly influenced by the sprinkler $p(R|S) = p(R)$
- Therefore, our model becomes

$$\begin{aligned} p(T,J,R,S) &= p(T|J,R,S)p(J|R,S)p(R|S)p(S) \\ &= p(T|R,S)p(J|R)p(R)p(S) \end{aligned}$$

Conditional Independence

- We may assume that Traceya's grass is wet depends only directly on whether or not it has been raining and whether or not her sprinkler was on so $p(T|J,R,S) = p(T|R,S)$
- Assume that Jack's grass is wet is influenced only directly by whether or not it has been raining $p(J|R,S) = p(J|R)$
- Furthermore, we assume the rain is not directly influenced by the sprinkler $p(R|S) = p(R)$
- Therefore, our model becomes

$$\begin{aligned} p(T,J,R,S) &= p(T|J,R,S)p(J|R,S)p(R|S)p(S) \\ &= p(T|R,S)p(J|R)p(R)p(S) \end{aligned}$$

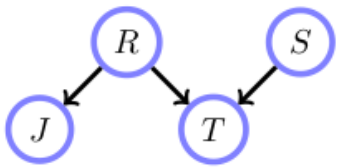
- Number of values we need to specify is 4+2+1+1=8

Conditional Independence

- We may assume that Traceya's grass is wet depends only directly on whether or not it has been raining and whether or not her sprinkler was on so $p(T|J,R,S) = p(T|R,S)$
- Assume that Jack's grass is wet is influenced only directly by whether or not it has been raining $p(J|R,S) = p(J|R)$
- Furthermore, we assume the rain is not directly influenced by the sprinkler $p(R|S) = p(R)$
- Therefore, our model becomes

$$\begin{aligned} p(T,J,R,S) &= p(T|J,R,S)p(J|R,S)p(R|S)p(S) \\ &= p(T|R,S)p(J|R)p(R)p(S) \end{aligned}$$

- Number of values we need to specify is 4+2+1+1=8
- We can represent these conditional independencies as



Belief network

How to represent these conditional independencies?

Belief network

How to represent these conditional independencies?

- **Belief network** is a distribution of the form

$$p(x_1, x_2, ..., x_n) = \Pi_{i=1}^{n} p(x_i | pa(x_i))$$

where $pa(x_i)$ represent the parental variables of variable x_i

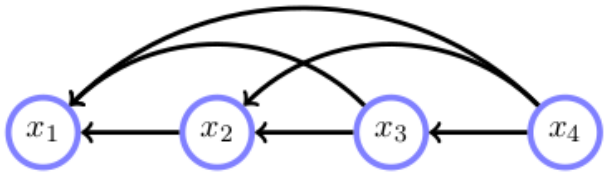
Belief network

How to represent these conditional independencies?

- **Belief network** is a distribution of the form

$$p(x_1, x_2, ..., x_n) = \Pi_{i=1}^{n} p(x_i | pa(x_i))$$

  where $pa(x_i)$ represent the parental variables of variable x_i
- Represented as a directed graph, with an arrow pointing from a parent variable to child variable, a belief network corresponds to a Directed Acyclic Graph (DAG)



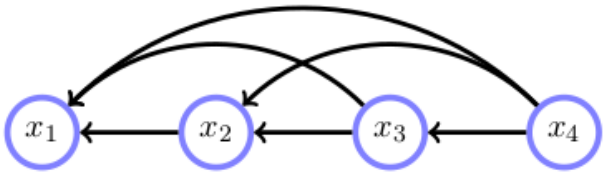
Belief network

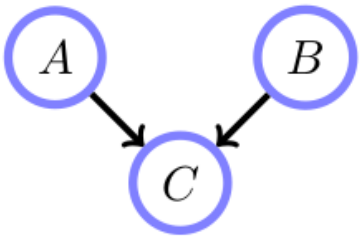
How to represent these conditional independencies?

- **Belief network** is a distribution of the form

$$p(x_1, x_2, ..., x_n) = \Pi_{i=1}^{n} p(x_i|pa(x_i))$$

  where $pa(x_i)$ represent the parental variables of variable x_i
- Represented as a directed graph, with an arrow pointing from a parent variable to child variable, a belief network corresponds to a Directed Acyclic Graph (DAG)





$$p(A, B, C) = p(C|A, B)p(A)p(B)$$

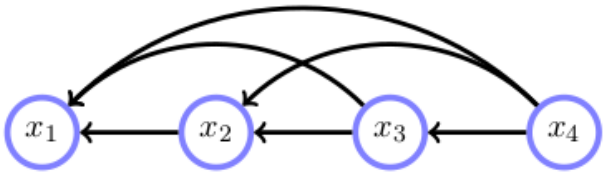
Belief network

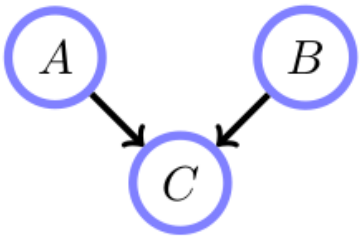
How to represent these conditional independencies?

- **Belief network** is a distribution of the form

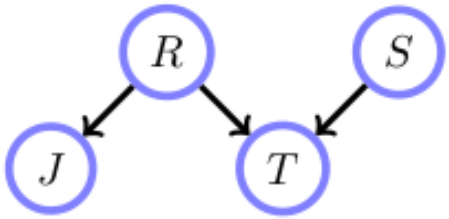
$$p(x_1, x_2, ..., x_n) = \Pi_{i=1}^{n} p(x_i|pa(x_i))$$

  where $pa(x_i)$ represent the parental variables of variable x_i
- Represented as a directed graph, with an arrow pointing from a parent variable to child variable, a belief network corresponds to a Directed Acyclic Graph (DAG)





$$p(A, B, C) = p(C|A, B)p(A)p(B)$$



Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$.

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$,

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$,

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$,

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$, $p(T{=}1|R{=}1, S{=}1) = 1$,

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$, $p(T{=}1|R{=}1, S{=}1) = 1$, $p(T{=}1|R{=}0, S{=}1) = 0.9$,

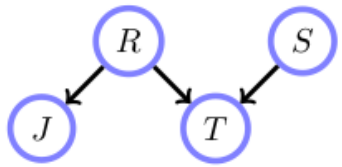
Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$, $p(T{=}1|R{=}1, S{=}1) = 1$, $p(T{=}1|R{=}0, S{=}1) = 0.9$, $p(T{=}1|R{=}0, S{=}0) = 0$

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$, $p(T{=}1|R{=}1, S{=}1) = 1$, $p(T{=}1|R{=}0, S{=}1) = 0.9$, $p(T{=}1|R{=}0, S{=}0) = 0$

Using following Beleif Network; calculate

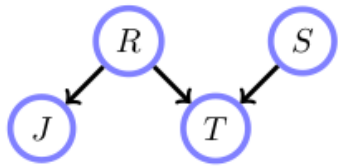


1. Probability that the sprinkler was *on* overnight, given that Traceya's grass is wet.

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$, $p(T{=}1|R{=}1, S{=}1) = 1$, $p(T{=}1|R{=}0, S{=}1) = 0.9$, $p(T{=}1|R{=}0, S{=}0) = 0$

Using following Beleif Network; calculate

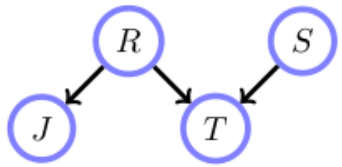


1. Probability that the sprinkler was *on* overnight, given that Traceya's grass is wet. $p(S{=}1|T{=}1) = 0.3382$

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$, $p(T{=}1|R{=}1, S{=}1) = 1$, $p(T{=}1|R{=}0, S{=}1) = 0.9$, $p(T{=}1|R{=}0, S{=}0) = 0$

Using following Beleif Network; calculate

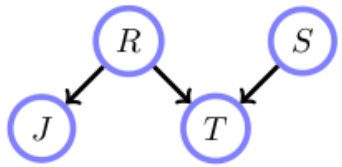


1. Probability that the sprinkler was *on* overnight, given that Traceya's grass is wet. $p(S{=}1|T{=}1) = 0.3382$ How? on next slide
2. Probability that Traceya's sprinkler was *on* overnight, given that her grass is wet and that Jack's grass is also wet.

Example

One morning Tracey realises that her grass is wet and the grass of her neighbour, Jack, is also wet. Let the prior probabilities be $p(R{=}1) = 0.2$ and $p(S{=}1) = 0.1$. We set $p(J{=}1|R{=}1) = 1$, $p(J{=}1|R{=}0) = 0.2$, $p(T{=}1|R{=}1, S{=}0) = 1$, $p(T{=}1|R{=}1, S{=}1) = 1$, $p(T{=}1|R{=}0, S{=}1) = 0.9$, $p(T{=}1|R{=}0, S{=}0) = 0$

Using following Beleif Network; calculate



1. Probability that the sprinkler was *on* overnight, given that Traceya's grass is wet. $p(S{=}1|T{=}1) = 0.3382$ How? on next slide
2. Probability that Traceya's sprinkler was *on* overnight, given that her grass is wet and that Jack's grass is also wet. $p(S{=}1|T{=}1, J{=}1) = 0.1604$

Example: Probability of $p(S{=}1|T{=}1)$

$$
\begin{aligned}
p(S=1|T=1) &= \frac{p(S=1,T=1)}{p(T=1)} \\
&= \frac{\sum_{J,R} p(S=1,J,R,T=1)}{\sum_{J,R,S} p(T=1,J,R,S)} \qquad (4) \\
&= \frac{\sum_{J,R} p(J|R)p(T=1|R,S=1)p(R)p(S=1)}{\sum_{J,R,S} p(J|R)p(T=1|R,S)p(R)p(S)} \\
&= \frac{\sum_{J,R} p(T=1|R,S=1)p(R)p(S=1)}{\sum_{J,R,S} p(T=1|R,S)p(R)p(S)} \qquad (5) \\
&= \frac{0.9\times 0.8\times 0.1 + 1\times 0.2\times 0.1}{.9\times .8\times .1 + 1\times .2\times .1 + 0\times .8\times .9 + 1\times .2\times .9} \\
&= \boxed{0.3382}
\end{aligned}
$$

Uses given belief network in (4) and proof in (5)

Thank You!

Thank you very much for your attention!

Queries ?

(Reference[3])

[3] 1) Book - *AIMA*, ch-14, Russell and Norvig. 2) Book - *Bayeian Reasoning and Machine Learning*, ch-04, David Barber.