Bayesian Decision Theory

Instructor: Dr. Muhammad Fahim

INNOPOLIS UNIVERSITY

Contents

- Bayesian Learning
- Importance of Bayesian Learning
- Bayes Theorem
- Choosing Hypothesis (MAP–Hypothesis)
- Bayes Classifiers
- Bayes Classifiers – Example
- Observations and Application
- Bayesian Network
- Casual Network
- Summary

Bayesian Learning

- Bayesian reasoning provides a probabilistic approach to inference

- **Basic Assumption:**
  - It is based on the assumption that the quantities of interest are governed by probability distributions

  - Optimal decisions can be made by reasoning about these probabilities together with observed data

Why Bayesian Learning?

- **First Reason**:
  - They explicit manipulate the probabilities
  - Most practical approaches to certain types of learning problems
  - For example: Bayes classifier is competitive with decision tree and neural network learning

- **Second Reason:**
  - They provide a useful perspective for understanding many learning algorithms that do not explicitly manipulate probabilities.
  - For example: Algorithms like Find-S and Candidate elimination algorithm

Provides “gold standard” for evaluating other learning algorithms

Bayes Theorem

- In machine learning we are often interested in determining the best hypothesis from space H, given the observed training data D.

- One way to specify what we mean by the best hypothesis is to say that we demand the most probable hypothesis, given the data D plus any initial knowledge about the prior probabilities of the various hypotheses in H

- Bayes theorem provides a direct method for calculating such probabilities.

Bayes Theorem – Prior Probabilities

- Lets consider P(h) to denote the initial probability that hypothesis h holds, before we have observed the training data. P(h) is often called the prior probability of h.
- If we have no such prior knowledge, then we might simply assign the same prior probability to each candidate hypothesis.
- Similarly, we will write P(D) to denote the prior probability that training data D will be observed.
- The probability of D given no knowledge about which hypothesis holds.

Bayes Theorem – Posterior Probabilities

- We will write P(D|h) to denote the probability of observing data D given some world in which hypothesis h holds. It is also known as likelihood.

- In machine learning problems we are interested in the probability P(h|D) that h holds given the observed training data D.

- P(h|D) is called the posterior probability of h, because it reflects our confidence that h holds after we have seen the training data D.

Bayes Theorem

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

Thomas Bayes
1701 – 1761

- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

Choosing Hypothesis (MAP–Hypothesis)

- In many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis h∈H given the observed training data D

- Any maximally probable hypothesis is called
  - *Maximum A Posteriori (***MAP***)* hypotheses

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$
$$= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$
$$= \arg\max_{h \in H} P(D|h)P(h)$$

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

Note: that P(D) can be dropped, because it is a constant independent of h

Choosing Hypothesis (ML–Hypothesis)

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P(h|D) \\
&= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\
&= \arg\max_{h \in H} P(D|h)P(h)
\end{aligned}
$$

If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$
h_{ML} = \arg\max_{h_i \in H} P(D|h_i)
$$

Bayes Classifiers

- **Assumption:** Training set consists of instances of different classes described c_j as conjunctions of attributes values
- **Task:** Classify a new instance *d* based on a tuple of attribute values into one of the classes $c_j \in C$
- **Key idea:** Assign the most probable class c_{MAP} using Bayes Theorem.

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_n)$$

$$= \underset{c_j \in C}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_n)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

Bayes Classifiers—Parameters estimation

- $P(c_j)$: Can be estimated from the frequency of classes in the training examples.
- $P(x_1,x_2,\ldots,x_n/c_j)$: Can be estimated from the available training examples.
- **Independence Assumption:** Attribute values are conditionally independent given the target value: ***naïve*** *Bayes*.

$$P(x_1, x_2, \ldots, x_n \mid c_j) = \prod_i P(x_i \mid c_j)$$

$$\underset{c_j \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

$$c_{NB} = \underset{c_j \in C}{\arg\max}\, P(c_j) \prod_i P(x_i \mid c_j)$$

Naïve Bayes Classifiers – Properties

- Estimating $P(x_i \mid c_j)$ instead of $P(x_1, x_2, \ldots, x_n \mid c_j)$ greatly reduces the number of parameters
- The learning step in Naïve Bayes consists of estimating $P(x_i \mid c_j)$ and $P(c_j)$ based on the frequencies in the training data
- An unseen instance is classified by computing the class that maximizes the posterior
- When conditioned independence is satisfied, Naïve Bayes corresponds to MAP classification.

Naïve Bayes Classifiers – Example

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

For the day <sunny, cool, high, strong>

What's the play prediction?

$$h_{NB} = \arg\max_{h \in [yes, no]} P(h)P(\mathbf{x} \mid h) = \arg\max_{h \in [yes, no]} P(h)\prod_{t} P(a_t \mid h)$$

Naïve Bayes Classifiers – Example

- Based on the examples in the table, classify the data x=(Outlook=Sunny, Temp=Cool, Hum=High, Wind=strong)
- That means: Play tennis or not?

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$h_{NB} = \arg\max_{h \in [yes,no]} P(h)P(\mathbf{x} \mid h) = \arg\max_{h \in [yes,no]} P(h)\prod_t P(a_t \mid h)$$

$$= \arg\max_{h \in [yes,no]} P(h)P(Outlook = sunny \mid h)P(Temp = cool \mid h)P(Humidity = high \mid h)P(Wind = strong \mid h)$$

Naïve Bayes Classifiers – Solution

- Based on the examples in the table, classify the data

  x=(Outlook=Sunny, Temp=Cool, Hum=High, Wind=strong)

- That means: Play tennis or not?

$$h_{NB} = \arg\max_{h \in [yes,no]} P(h)P(\mathbf{x} \mid h) = \arg\max_{h \in [yes,no]} P(h)\prod_t P(a_t \mid h)$$

$$= \arg\max_{h \in [yes,no]} P(h)P(Outlook = sunny \mid h)P(Temp = cool \mid h)P(Humidity = high \mid h)P(Wind = strong \mid h)$$

- Probabilities Calculation

$$P(PlayTennis = yes) = 9/14 = 0.64$$

$$P(PlayTennis = no) = 5/14 = 0.36$$

$$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$$

$$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$$

.....

$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$

$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = \mathbf{0.0206}$$

$$\Rightarrow answer : PlayTennis(x) = no$$

Underflow Prevention

- Multiplying lots of probabilities, which are between 0 and 1 by definition, can result in **floating-point underflow.**

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_i P(x_i \mid c_j)$$

$$c_{NB} = \operatorname{argmax}_{c_j \in C} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

- Since *log(xy) = log(x) + log(y),* it is better to perform all computations by summing logs of probabilities rather than multiplying probabilities.

- Class with highest final unnormalized log probability score is still the most probable.

Observations

- **Advantages**
  - It is easy and fast to predict class of test data set. It also perform well in multi class prediction
  - When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
  - It perform well in case of categorical input variables compared to numerical variable(s).
  - For numerical variable, normal distribution is assumed – bell curve, **which is a strong assumption.**

Observations

- **Disadvantages**
  - If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”.
  - **Solution:** To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
  - Another limitation of Naïve Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Applications of Naive Bayes Algorithms

- Real time Prediction
- Multi class Prediction
- Text classification / Spam Filtering / Sentiment Analysis
- Recommendation System

Naïve Bayes classifier assumes that all the variables are conditionally independent

In real life, it is almost impossible that we get a set of predictors which are completely independent

Any Solution?

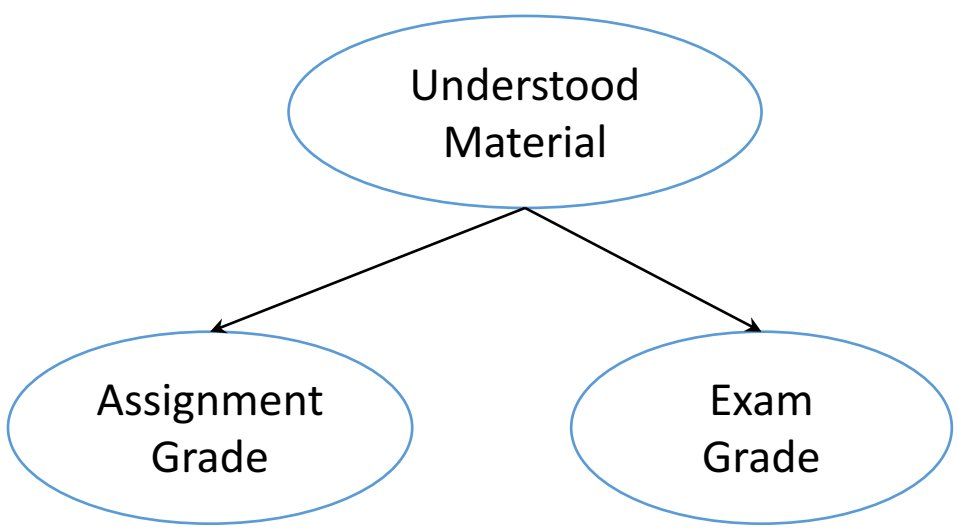
Belief Network (Bayesian Network)

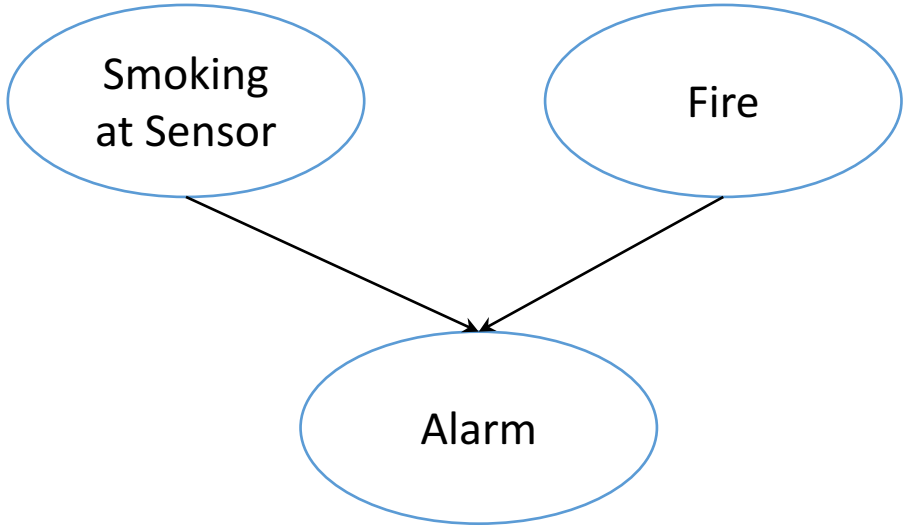
- A Bayesian belief network describes the probability distribution governing a set of variables by
  - specifying a set of conditional independence assumptions
  - along with a set of conditional probabilities.

- It relax the Naïve Bayes classifier assumption – that all the variables are conditionally independent

Belief Network – Intuition

- Some informal Examples





Belief Network

Definition 18 (Belief Network)**.** A Belief Network is a distribution of the form

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i | \text{pa}\,(x_i)) \tag{3.1.2}$$

where $\text{pa}\,(x_i)$ represent the *parental* variables of variable x_i. Written as a Directed Graph, with an arrow pointing from a parent variable to child variable, a Belief Network is a Directed Acyclic Graph (DAG), with the i^{th} vertex in the graph corresponding to the factor $p(x_i|\text{pa}\,(x_i))$.

Source: Chapter 3, Section 3.1, "Bayesian Reasoning and Machine Learning" By: David Barber

Constructing a Simple Bayesian Belief Network

Tracey

Sprinkler

Rain

Example Source: Chapter 3, "Bayesian Reasoning and Machine Learning" By: David Barber

Constructing a Simple Bayesian Belief Network

- One morning Tracey leaves her house and realizes 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.
- This explains away to some extent the possibility that her sprinkler was left on, and she concludes therefore that it has probably been raining.

We can model the above situation by first defining the variables we wish to include in our model.

Making a Model

- In the example, the natural variables are
  - Rain – R
  - Sprinkler – S
  - Jacks – J
  - Tracy – T

Making a Model

- $R \in \{0, 1\}$ R = 1 means that it has been raining, and 0 otherwise
- $S \in \{0, 1\}$ S = 1 means that Tracey has forgotten to turn off the sprinkler, and 0 otherwise
- $J \in \{0, 1\}$ J = 1 means that Jack's grass is wet, and 0 otherwise
- $T \in \{0, 1\}$ T = 1 means that Tracey's Grass is wet, and 0 otherwise

Making a Model

- A model of Tracey's world then corresponds to a probability distribution on the joint set of the variables of interest
  - p(T, J, R, S) (joint probability – the order of the variables is irrelevant).

- Since each of the variables in this example can take one of two states, it would appear that we naively have to specify the values for each of the

$$2^4 = 16 \text{ states}$$

Making a Model

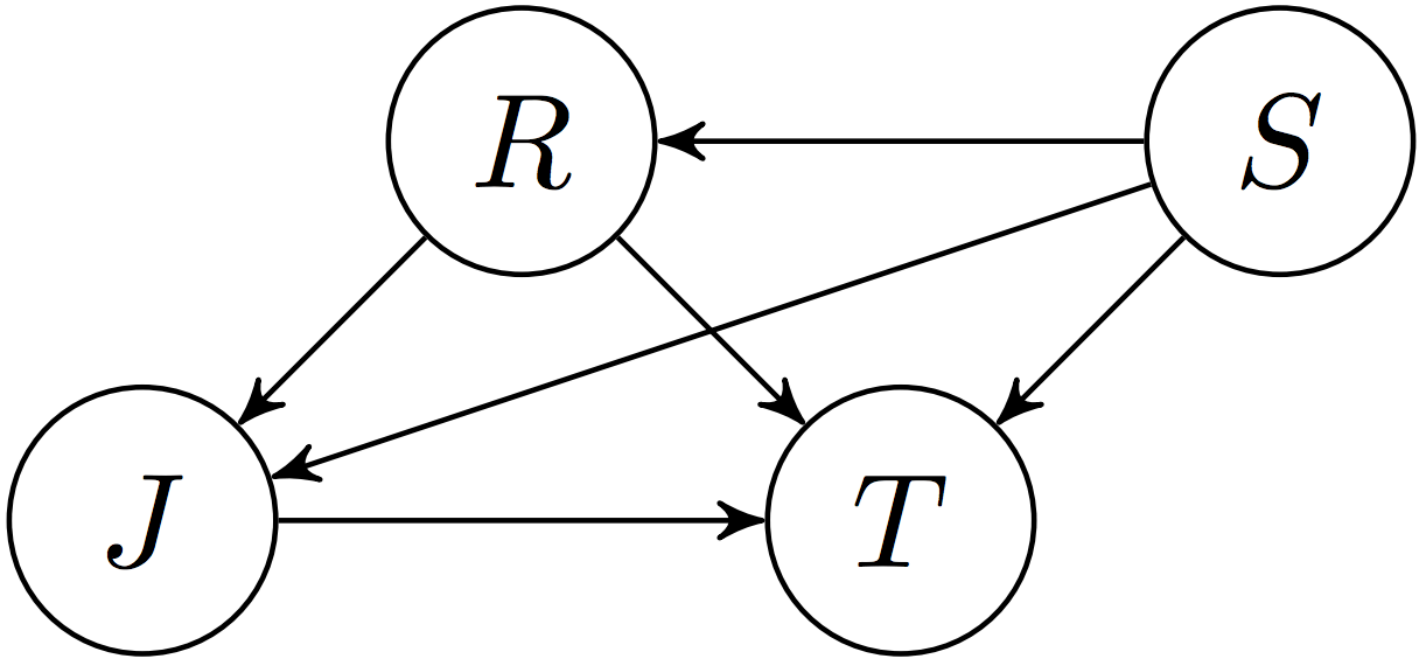
- To see how many states need to be specified, consider the following decomposition.
- Without loss of generality and repeatedly using the definition of conditional probability, we may write:

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

- The first term $p(T|J, R, S)$ requires us to specify $2^3 = 8$ values
- Similarly, we need 4, 2, 1 values for the other factors
- Making a total of $8 + 4 + 2 + 1 = 15$ values in all.
- In general, for a distribution on n binary variables, we need to specify $2^n - 1$ values in the range [0, 1].

Belief Network

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



Modelling Independencies

- The important point here is that the number of values that need to be specified in general scales exponentially increase with the number of variables in the model

This is impractical in general and motivates simplifications

Lets go ☺ …. To make it simple!!

Conditional Independence

- The modeler often knows constraints on the system.

- For instance, in our scenario, we may assume that Tracey's grass is wet depends only directly on whether or not is has been raining and whether or not her sprinkler was on.

- That is, we make a conditional independence assumption

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

- Similarly, we assume that Jack's grass is wet is influenced only directly by whether or not it has been raining, and write $p(J|R, S) = p(J|R)$

Conditional Independence

- Furthermore, we assume the rain is not directly influenced by the sprinkler

$$p(R|S) = p(R)$$

- Which means that our model equation now becomes

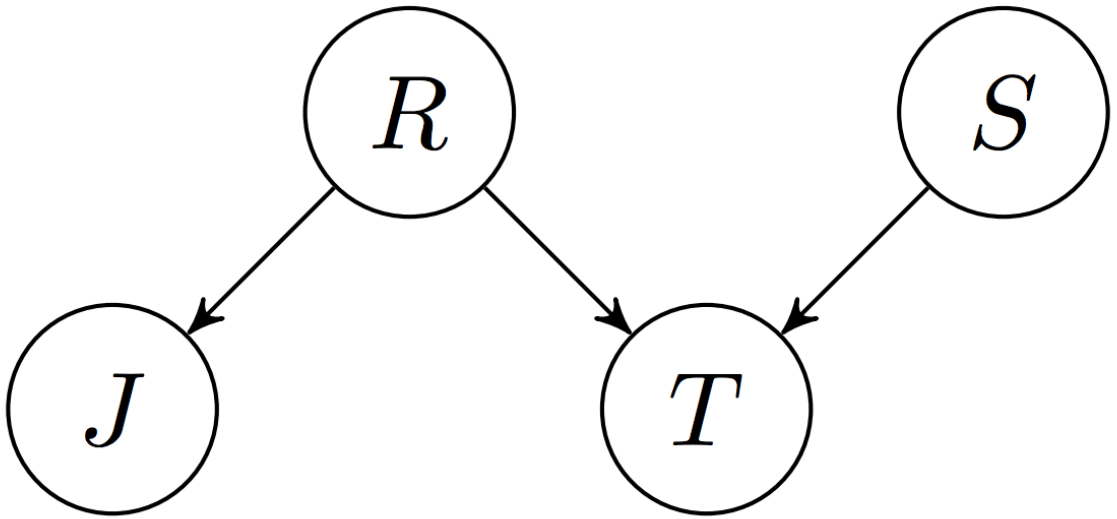
$$p(T, J, R, S) = p(T|R, S)p(J|R)p(R)p(S)$$

- This reduces the number of values that we need to specify to 4 + 2 + 1 + 1 = 8, a saving over the previous 15 values in the case where no conditional independencies had been assumed.

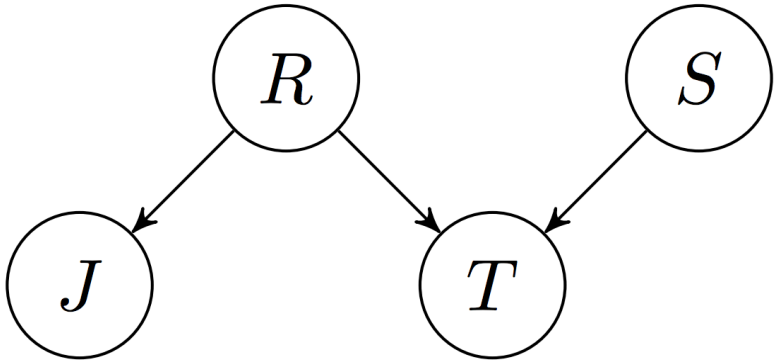
Conditional Independence

- We can represent these conditional independencies graphically

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



Belief Network



- It is a special type of diagram (called a directed graph) together with an associated set of probability tables.
- The graph consists of nodes and arcs.
  - The nodes represent variables, which can be discrete or continuous.
- The arcs represent causal relationships between variables.
- BBNs enable us to model and reason about uncertainty.

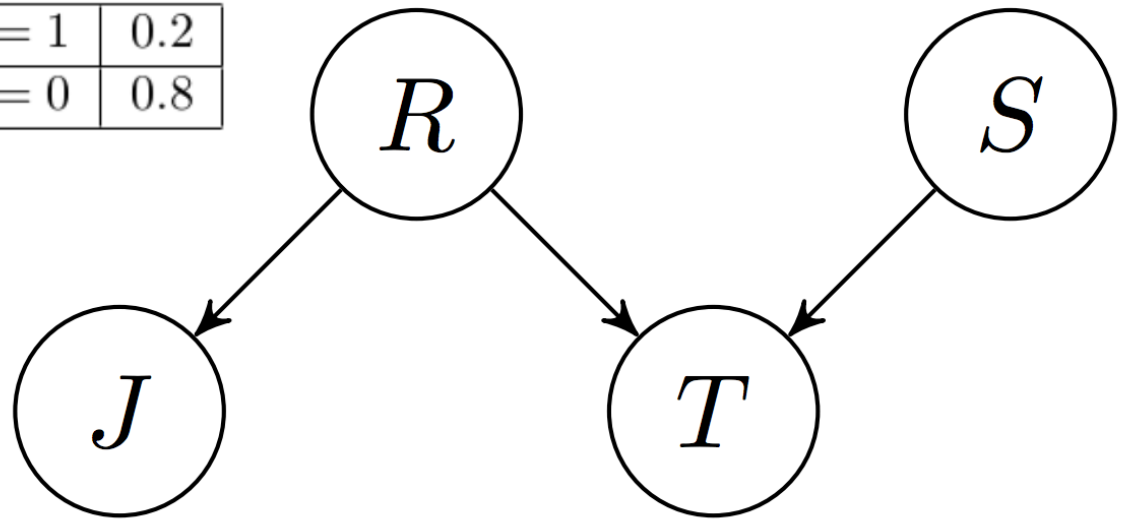
Conditional Independence

- To complete the model, we need to numerically specify the values of each Conditional Probability Table (CPT)
- Let the prior probabilities for R and S be
  - $p(R = 1) = 0.2$ and $p(S = 1) = 0.1$.
- We set the remaining probabilities to
  - $p(J = 1 | R = 1) = 1$,
  - $p(J = 1 | R = 0) = 0.2$ (sometimes Jack's grass is wet due to unknown effects, other than rain),
  - $p(T = 1 | R = 1, S = 0) = 1$,
  - $p(T = 1 | R = 1, S = 1) = 1$,
  - $p(T = 1 | R = 0, S = 1) = 0.9$ (there's a small chance that even though the sprinkler was left on, it didn't wet the grass noticeably),
  - $p(T = 1 | R = 0, S = 0) = 0$.

Conditional Independence

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

- Conditional Probability Table (CPT)



$p(R)$

$R = 1$	0.2
$R = 0$	0.8

$p(S)$

$S = 1$	0.1
$S = 0$	0.9

$p(J \mid R)$

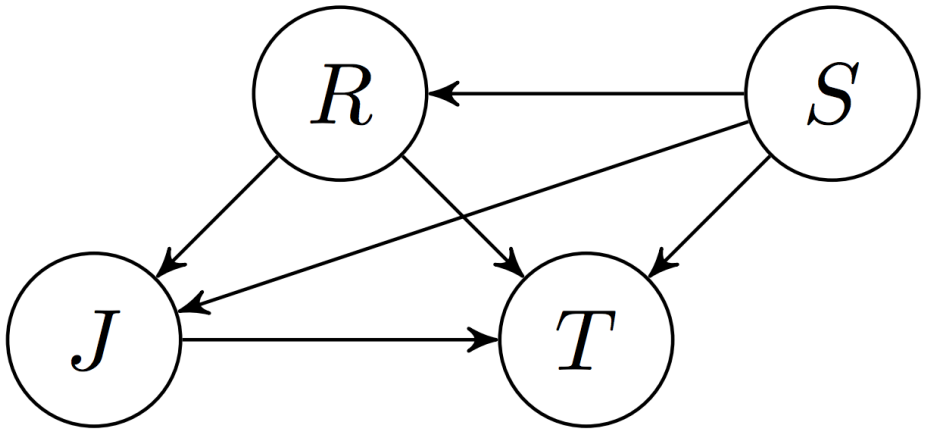
	$R = 1$	$R = 0$
$J = 1$	1	0.2
$J = 0$	0	0.8

$p(T \mid R, S)$

	$R = 1, S = 1$	$R = 1, S = 0$	$R = 0, S = 1$	$R = 0, S = 0$
$T = 1$	1	1	0.9	0
$T = 0$	0	0	0.1	1

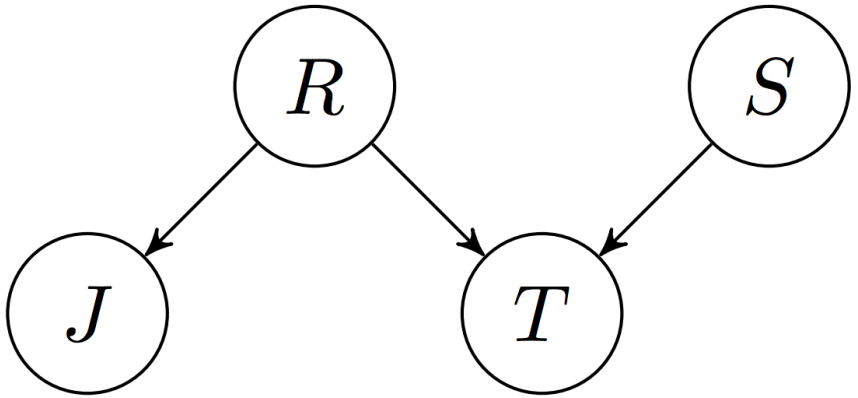
Belief Network

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



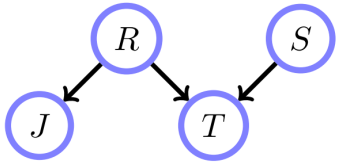
Without constrains

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



With constrains

Inference



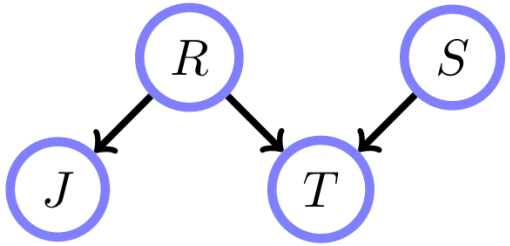
- We have made a model of an environment, we can perform inference

- What is the probability that the sprinkler was on given that we observe that Tracey's grass is wet? – p(S = 1 | T = 1)

$$\mathrm{p}(S = 1 \mid T = 1) = \frac{\mathrm{p}(S = 1, T = 1)}{\mathrm{p}(T = 1)}$$

Inference

- Since T depends on S and R and both S and R are independent of each other



$$\mathsf{p}(T=1) = \sum_{s\in\mathcal{S},r\in\mathcal{R}} \mathsf{p}(T,s,r) = \sum_{s\in\mathcal{S},r\in\mathcal{R}} \mathsf{p}(T=1 \mid s,r)\mathsf{p}(s)\mathsf{p}(r)$$

- Summing over different values of s and r (using CPT) , we get

$$1 \cdot 0.2 \cdot 0.1 + 1 \cdot 0.2 \cdot 0.9 + 0.9 \cdot 0.8 \cdot 0.1 + 0 \cdot 0.8 \cdot 0.9 = 0.272$$

$$\mathsf{p}(T=1) = 0.272$$

Inference

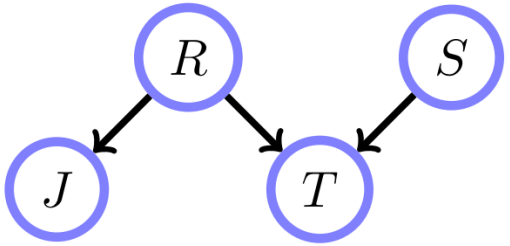
$$\mathrm{p}(T=1) = \sum_{s\in\mathcal{S},r\in\mathcal{R}} \mathrm{p}(T,s,r) = \sum_{s\in\mathcal{S},r\in\mathcal{R}} \mathrm{p}(T=1 \mid s,r)\mathrm{p}(s)\mathrm{p}(r)$$

$$\begin{aligned}
&= \mathrm{p}(\mathrm{T}=1 \mid \mathrm{r}=1, \mathrm{s}=1)\ \mathrm{p}(\mathrm{r}=1)\mathrm{p}(\mathrm{s}=1) + \\
&\quad \mathrm{p}(\mathrm{T}=1 \mid \mathrm{r}=1, \mathrm{s}=0)\ \mathrm{p}(\mathrm{r}=1)\mathrm{p}(\mathrm{s}=0) + \\
&\quad \mathrm{p}(\mathrm{T}=1 \mid \mathrm{r}=0, \mathrm{s}=1)\ \mathrm{p}(\mathrm{r}=0)\mathrm{p}(\mathrm{s}=1) + \\
&\quad \mathrm{p}(\mathrm{T}=1 \mid \mathrm{r}=0, \mathrm{s}=0)\ \mathrm{p}(\mathrm{r}=0)\mathrm{p}(\mathrm{s}=0)
\end{aligned}$$

$$1 \cdot 0.2 \cdot 0.1 + 1 \cdot 0.2 \cdot 0.9 + 0.9 \cdot 0.8 \cdot 0.1 + 0 \cdot 0.8 \cdot 0.9 = 0.272$$

Inference

$$\mathsf{p}(S = 1 \mid T = 1) = \frac{\mathsf{p}(S = 1, T = 1)}{\mathsf{p}(T = 1)}$$



$$\mathsf{p}(S = 1, T = 1) = \mathsf{p}(T = 1 \mid S = 1)\mathsf{p}(S = 1)$$

$$\mathsf{p}(T = 1 \mid S = 1) = \mathsf{p}(T = 1 \mid R = 0, S = 1)\mathsf{p}(R = 0) + \mathsf{p}(T = 1 \mid R = 1, S = 1)\mathsf{p}(R = 1)$$

$$\mathsf{p}(T = 1 \mid S = 1) = 0.9 \times 0.8 + 1 \times 0.2 = 0.92$$

$$\mathsf{p}(S = 1, T = 1) = 0.92 \times 0.1 = 0.092$$

$$\mathsf{p}(S = 1 \mid T = 1) = \frac{0.092}{0.272} = 0.3382$$

Inference

- What is the probability that the sprinkler was on given that we observe that Tracey's and Jack's grass is wet? – p(S = 1 | T = 1, J = 1)

- We use conditional probability again:

$$p(S{=}1|T{=}1, J{=}1) = \frac{p(S{=}1, T{=}1, J{=}1)}{p(T{=}1, J{=}1)} = \frac{\sum_R p(T{=}1, J{=}1, R, S{=}1)}{\sum_{R,S} p(T{=}1, J{=}1, R, S)}$$
$$= \ldots = 0.1604$$

- Jack's wet grass is explaining away the sprinkler as a reason for the wet grass of Tracey

Structure learning

- A Bayesian network is specified by an expert and is then used to perform inference. (In simplest case)
- In other applications the task of defining the network is too complex for humans.
- In this case the network structure and the parameters of the local distributions must be learned from data.

Structure Learning

- Automatically learning the graph structure of a Bayesian network (BN) is a challenge pursued within machine learning.

- For generating the Bayesian Network we can use K2 – Structural learning algorithm

- Another powerful library is: https://github.com/jmschrei/pomegranate

Parameter Learning and Inference

- Often these conditional distributions include parameters that are unknown and must be estimated from data
  - e.g., via the maximum likelihood approach.

- Naturally, we don't wish to carry out such inference calculations by hand all the time!!

- General purpose algorithms exist for this, such as the junction tree algorithm

Issues in Bayesian Networks

- Given data, learning the structure of the Bayesian Network (NP-Complete)
  - Finding the arcs (dependencies) between the nodes
  - Calculating conditional probability tables

- Given the Bayesian Network, finding an efficient algorithm for inference on a given structure (NP-Complete)

Issues in Bayesian Networks

- Your homework task to investigate it.

Applications of Bayesian Networks

- Example applications include:
  - Computer vision
  - Natural language processing
  - Bioinformatics
  - Medical diagnosis
  - Weather forecasting

- Example systems include:
  - PATHFINDER medical diagnosis system at Stanford
  - Microsoft Office assistant and troubleshooters
  - Space shuttle monitoring system at NASA Mission Control Center in Houston

Causal Bayesian Networks

- Build a Bayesian network using casual relationships
  - Choose a set of variables that describes the domain.
  - Draw an arc to a variable from each of its DIRECT causes. (Domain knowledge needed here.)

- Which results a causal Bayesian network, or simply causal networks.
  - Arcs are interpreted as indicating cause-effect relationships.

Summary

- Bayesian Learning
- Importance of Bayesian Learning
- Bayes Theorem
- Choosing Hypothesis (MAP–Hypothesis)
- Bayes Classifiers
- Bayes Classifiers – Example
- Observations and Application
- Bayesian Network
- Casual Network
- Summary

Reference

- Chapter 3 Belief Networks of the book title “Bayesian Reasoning and Machine Learning” By: David Barber (2012)
- Chapter 6 of Tom M. Mitchell

Thank You ☺

Appendix

What is probability?

- Probability theory is the body of knowledge that enables us to reason formally about uncertain events.

- The probability P of an uncertain event A, written P(A), is defined by the frequency of that event based on previous observations.

- This is called *frequency* based probability.

Probability axioms

- P(a) should be a number between 0 and 1.
- If a represents a certain event then P(a)=1.
- If a and b are mutually exclusive events then

$$P(a \text{ or } b)=P(a)+P(b).$$

- Mutually exclusive means that they cannot both be true at the same time.

Probability Distribution

- There is a variable called A.
- The variable A can have many states:

$$A = \{a_1, a_2, ..., a_n\}$$

$$\sum_{i=1}^{n} P(a_i) = 1$$

- We can think of an event as just one state of the variable A.
- The probability distribution of A, written P(A), is simply the set of values $\{p(a_1), p(a_2), ..., p(a_n)\}$

Join Events

- Suppose there are two events A and B that:
- A = {a1, a2, a3} where a1=0, a2=1, a3=">1"
- B = {b1, b2, b3} where b1=0, b2=1, b3=">1"
- The joint event A and B:

$$P(A,B) = P(a_i, b_j) = \begin{Bmatrix} P(a_1,b_1), P(a_1,b_2), P(a_1,b_3), \\ P(a_2,b_1), P(a_2,b_2), P(a_2,b_3), \\ P(a_3,b_1), P(a_3,b_2), P(a_3,b_3) \end{Bmatrix}$$

$$i = 1 \ldots n; \quad j = 1 \ldots m$$

- P(A,B) is called the joint probability distribution of A and B. The general form

$$P(A_1, \ldots, A_n) = \prod_{i=1}^{n} P(A_i \mid parents(A_i))$$

Marginalization

- If we know the joint probability distribution P(A,B) then we can calculate P(A) by a formula called marginalization:

$$P(a) = \sum_i P(a, b_i)$$

- This is because the events (a,b_1) are mutually exclusive.
- When we calculate P(A) in this way from the joint probability distribution we say that the variable B is marginalized out of P(A,B).

Belief Measure

- In general, a person belief in a statement $\boldsymbol{\alpha}$ will depend on some body of knowledge K. We write this as $P(\boldsymbol{\alpha}|K)$.

- The expression $P(\boldsymbol{\alpha}|K)$ thus represents a belief measure.

- Sometimes, for simplicity, when K remains constant we just write $P(\boldsymbol{\alpha})$, but you must be aware that this is a simplification.

Conditional Probability

- The notion of degree of belief P(A|K) is an uncertain event A is ***conditional*** on a body of knowledge K.
- In general, we write P(A|B) to represent a belief in A under the assumption that B is known.
- Strictly speaking, P(A|B) is a shorthand for the expression P(A|B,K) where K represents all other relevant information.
- Only when all other information is irrelevant can we really write P(A|B).
- The traditional approach to defining conditional probability is via joint probabilities:

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

Conditional Probability

- We can rearrange the formula for conditional probability to get the so-called product rule:
  - $$P(A,B) = P(A|\,B)P(B)$$
- We can extend this for three variables:
  - $$P(A,B,C) = P(A|\,B,C)P(B,C) = P(A|\,B,C)P(B|C)P(C)$$
- And to *n* variables:

$$P(A_1, A_2, \ldots, A_n) = P(A_1 \mid A_2, \ldots, A_n)P(A_2 \mid A_3, \ldots, A_n), \ldots, P(A_{n-1} \mid A_n)P(A_n)$$

- In general we refer to this as the chain rule.

Independence and Conditionally Independence

- The conditional probability of A given B is represented by P(A|B). The variables A and B are said to be ***independent*** if P(A)=P(A|B) or alternatively if P(A,B)=P(A)P(B).

- A and B are conditionally independent given C if P(A|C)=P(A|B,C).

Thank You ☺