# MACHINE LEARNING

# Baysian Learning

**Dr G.Kalyani**

**Department of Information Technology**

**Velagapudi Ramakrishna Siddhartha Engineering College**

# Topics

- **Bayes Theorem**
- **Bayes Optimal Classifier**
- **Naïve Byes Classifier**
- **Bayesian Belief Networks**

# Features of Bayesian Learning Methods

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct. This provides a more flexible approach to learning than algorithms that completely eliminate a hypothesis if it is found to be inconsistent with any single example.

- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
  - (1) a prior probability for each candidate hypothesis, and
  - (2) a probability distribution over observed data for each possible hypothesis.

- Bayesian methods can accommodate hypotheses that make probabilistic predictions (e.g., hypotheses such as "this pneumonia patient has a 93% chance of complete recovery").

- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.

# Bayes Theorem

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

- $P(h)$ = 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 the Hypothesis

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

Generally want the most probable hypothesis given the training data

*Maximum a posteriori* hypothesis $h_{MAP}$:

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

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 Theorem-Example

Does patient have cancer or not?

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

$$P(cancer) = \qquad P(\neg cancer) =$$
$$P(+|cancer) = \qquad P(-|cancer) =$$
$$P(+|\neg cancer) = \qquad P(-|\neg cancer) =$$

Calculate the following

$P(Cancer|+)$

P(–Cancer|+)

# Bayes Theorem

- The result of **Bayesian inference depends strongly on the prior probabilities**, which must be available in order to apply the method directly.

# Topics

- **Bayes Theorem**
- **Bayes Optimal Classifier**
- **Naïve Bayes Classifier**
- **Baysian Belief Networks**

# Bayes Optimal Classifier

- So far we have considered the question

  "what is the most probable hypothesis given the training data?"

- In fact, the question that is often of most significance is the closely related question

  "what is the most probable classification of the new instance given the training data?"

**Bayes optimal classification:**

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

# Example for Bayes Optimal Classifier

- consider a hypothesis space containing three hypotheses, hl, h2, and h3.
- Suppose that the posterior probabilities of these hypotheses given the training data are 0.4, 0.3, and 0.3 respectively.
- Thus, hl is the MAP hypothesis.
- Suppose a new instance x is encountered, which is classified positive by hl, but negative by h2 and h3.

$$P(h_1|D) = .4,\ P(-|h_1) = 0,\ P(+|h_1) = 1$$
$$P(h_2|D) = .3,\ P(-|h_2) = 1,\ P(+|h_2) = 0$$
$$P(h_3|D) = .3,\ P(-|h_3) = 1,\ P(+|h_3) = 0$$

# Example for Bayes Optimal Classifier

$$P(h_1|D) = .4,\ P(-|h_1) = 0,\ P(+|h_1) = 1$$
$$P(h_2|D) = .3,\ P(-|h_2) = 1,\ P(+|h_2) = 0$$
$$P(h_3|D) = .3,\ P(-|h_3) = 1,\ P(+|h_3) = 0$$

therefore

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4$$
$$\sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6$$

and

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = -$$

# Topics

- **Bayes Theorem**
- **Bayes Optimal Classifier**
- **Naïve Bayes Classifier**
- **Baysian Belief Networks**

# Naïve Bayesian Classifier

Along with decision trees, neural networks, nearest nbr, one of the most practical learning methods.

When to use

- Moderate or large training set available
- Attributes that describe instances are conditionally independent given classification

Successful applications:

- Diagnosis
- Classifying text documents

# Naïve Bayes Classifier

Assume target function $f : X \rightarrow V$, where each instance $x$ described by attributes $\langle a_1, a_2 \ldots a_n \rangle$. Most probable value of $f(x)$ is:

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

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

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

Naive Bayes assumption:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

which gives

**Naive Bayes classifier:** $v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$

# Algorithm for Naïve Bayesian Classifier

Naive_Bayes_Learn($examples$)

For each target value $v_j$

$\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$

For each attribute value $a_i$ of each attribute $a$

$\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$

Classify_New_Instance($x$)

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

# Example for Naïve Bayesian Classifier

*PlayTennis*: training examples

| Day | Outlook | Temperature | Humidity | Wind | PlayTennis |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |
| D15 | Sunny | Cool | High | Strong | ? |

# Example for Naïve Bayesian Classifier

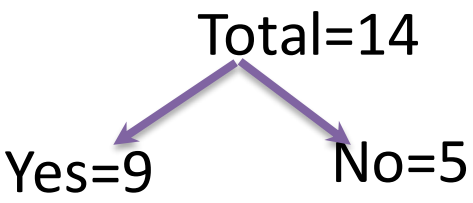
**Naive Bayes classifier:** $v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$

- X={sunny, cool, high, strong}

- P(Yes/X)=P(yes) * P(X/Yes)
  =P(Yes)*P(Sunny/Yes) )*P(Cool/Yes) *P(high/Yes) *P(strong/Yes)

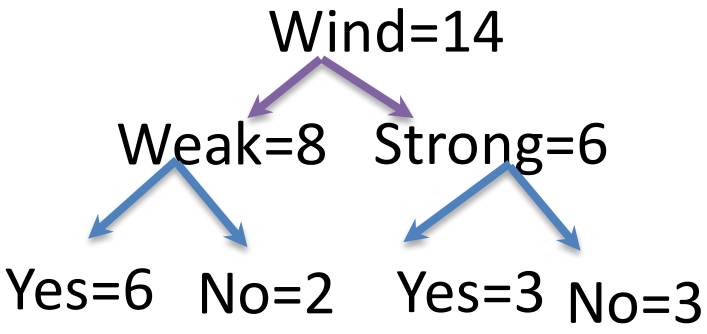
- P(No/X)=P(No) * P(X/No)
  =P(No)*P(Sunny/No)*P(Cool/No) *P(high/No) *P(strong/No)

- Select the class which is having highest probability




P(Yes)=9/14

P(No)=5/14

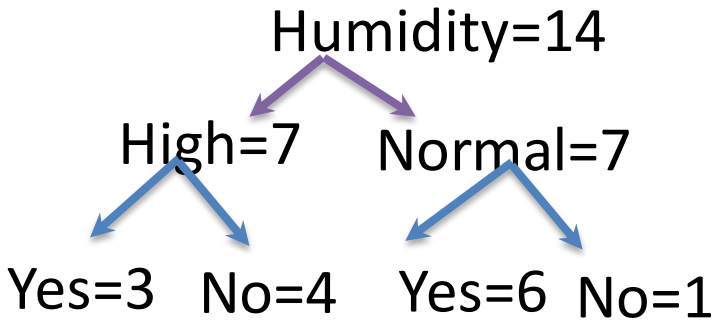



P(Weak/Yes)=6/9

P(Weak/No)=2/5

P(Strong/Yes)=3/9

P(Strong/No)=3/5

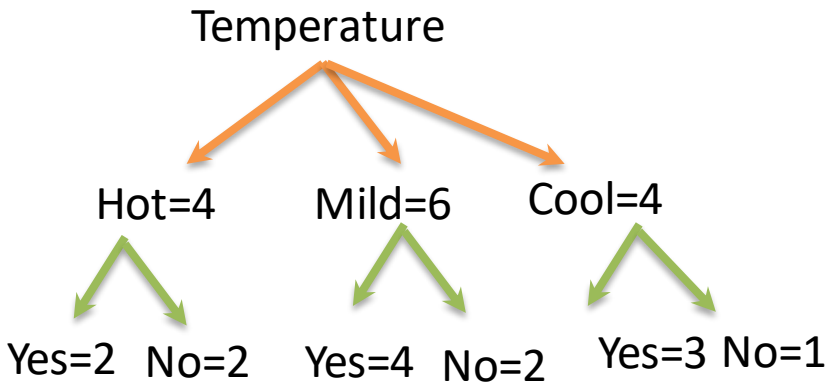



P(High/Yes)=3/9

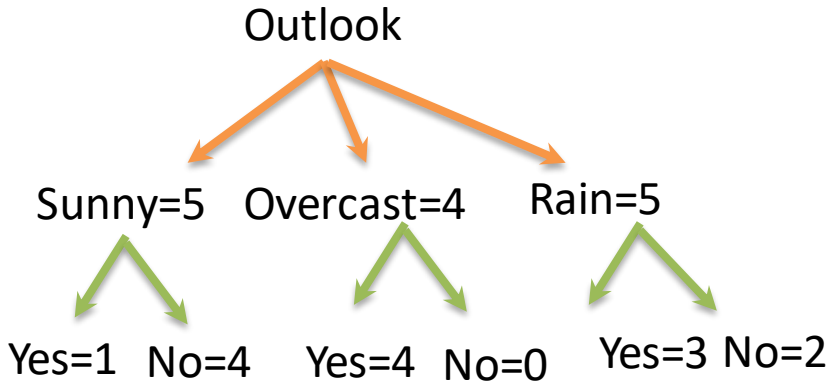
P(High/No)=4/5

P(Normal/Yes)=6/9

P(Normal/No)=1/5




P(Hot/Yes)=2/9
P(Hot/No)=2/5
P(Mild/Yes)=4/5
P(Mild/No)=2/5
P(Cool/Yes)=3/9
P(Cool/No)=1/5




P(Sunny/Yes)=1/9
P(Sunny/No)=4/5
P(Overcast/Yes)=4/9
P(Overcast/No)=0/5
P(Rain/Yes)=3/9
P(Rain/No)=2/5

# Example for Naïve Bayesian Classifier

Naive Bayes assumption:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

- X={sunny, cool, high, strong}

- P(Yes/X)=

  P(Yes)*P(Sunny/Yes) )*P(Cool/Yes) *P(high/Yes) *P(strong/Yes)

  $= \frac{9}{14} * \frac{1}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} = 0.0026$

- P(NO/X)=

  P(No)*P(Sunny/No)*P(Cool/No) *P(high/No) *P(strong/No)

  $= \frac{5}{14} * \frac{4}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} = 0.0274$

- Maximum A Posteriori (MAP)

$$y_{MAP} = \arg\max_Y P(Y|X) = \arg\max_Y \frac{P(X|Y)P(Y)}{P(X)} = \arg\max_Y P(X|Y)P(Y)$$

According to Majority class Rule: Max(0.0026,0.0274)=0.0274

The Answer is NO

# Estimating the Probabilities

what if none of the training instances with target value $v_j$ have attribute value $a_i$? Then

$$\hat{P}(a_i|v_j) = 0, \text{ and...}$$
$$\hat{P}(v_j) \prod_i \hat{P}(a_i|v_j) = 0$$

Typical solution is Bayesian estimate for $\hat{P}(a_i|v_j)$

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n + m}$$

where

- $n$ is number of training examples for which $v = v_j$,
- $n_c$ number of examples for which $v = v_j$ and $a = a_i$

- p is our prior estimate of the probability we wish to determine, and
- m is a constant called the equivalent sample size, which determines how heavily to weight p relative to the observed data.
- A typical method for choosing p in the absence of other information is to assume uniform priors; that is, if an attribute has k possible values we set p = 1/k

# Task on Naïve Bayes Classifier

1. (20 pts) Given the following data set containing three attributes and one class, use Naïve Bayes classifier to determine the class (Yes/No) of Stolen for a Red Domestic SUV.

| Example No. | Color | Type | Origin | Stolen? |
|---|---|---|---|---|
| 1 | Red | Sports | Domestic | Yes |
| 2 | Red | Sports | Domestic | No |
| 3 | Red | Sports | Domestic | Yes |
| 4 | Yellow | Sports | Domestic | No |
| 5 | Yellow | Sports | Imported | Yes |
| 6 | Yellow | SUV | Imported | No |
| 7 | Yellow | SUV | Imported | Yes |
| 8 | Yellow | SUV | Domestic | No |
| 9 | Red | SUV | Domestic | No |
| 10 | Red | Sports | Imported | Yes |

# Topics

- **Bayes Theorem**
- **Bayes Optimal Classifier**
- **Naïve Bayes Classifier**
- **Bayesian Belief Networks**

# Bayesian Belief Networks

- The naive Bayes classifier makes significant use of the **assumption that the values of the attributes $a_1 \ldots a_n$ are conditionally independent** given the target value v.

- However, in many cases this conditional independence assumption is clearly overly restrictive.

- 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**.

# Conditional Independence

- Let X, Y, and Z be three discrete-valued random variables. We say that X is ***conditionally independent*** of Y given Z if the probability distribution governing X is independent of the value of Y given a value for z; that is, if

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

P(X | Y, Z) = P(X | Z).

- This definition of conditional independence can be extended to sets of variables as well. We say that the set of variables X1 . . . Xi is conditionally independent of the set of variables ***Yl*** . . . Ym given the set of variables z1 . . . Zk, if

$$P(X_1 \ldots X_l | Y_1 \ldots Y_m, Z_1 \ldots Z_n) = P(X_1 \ldots X_l | Z_1 \ldots Z_n)$$

- Hence in naïve Bayes,

$$\begin{aligned} P(A_1, A_2 | V) &= P(A_1 | A_2, V) P(A_2 | V) \\ &= P(A_1 | V) P(A_2 | V) \end{aligned}$$

# Bayesian Belief Networks

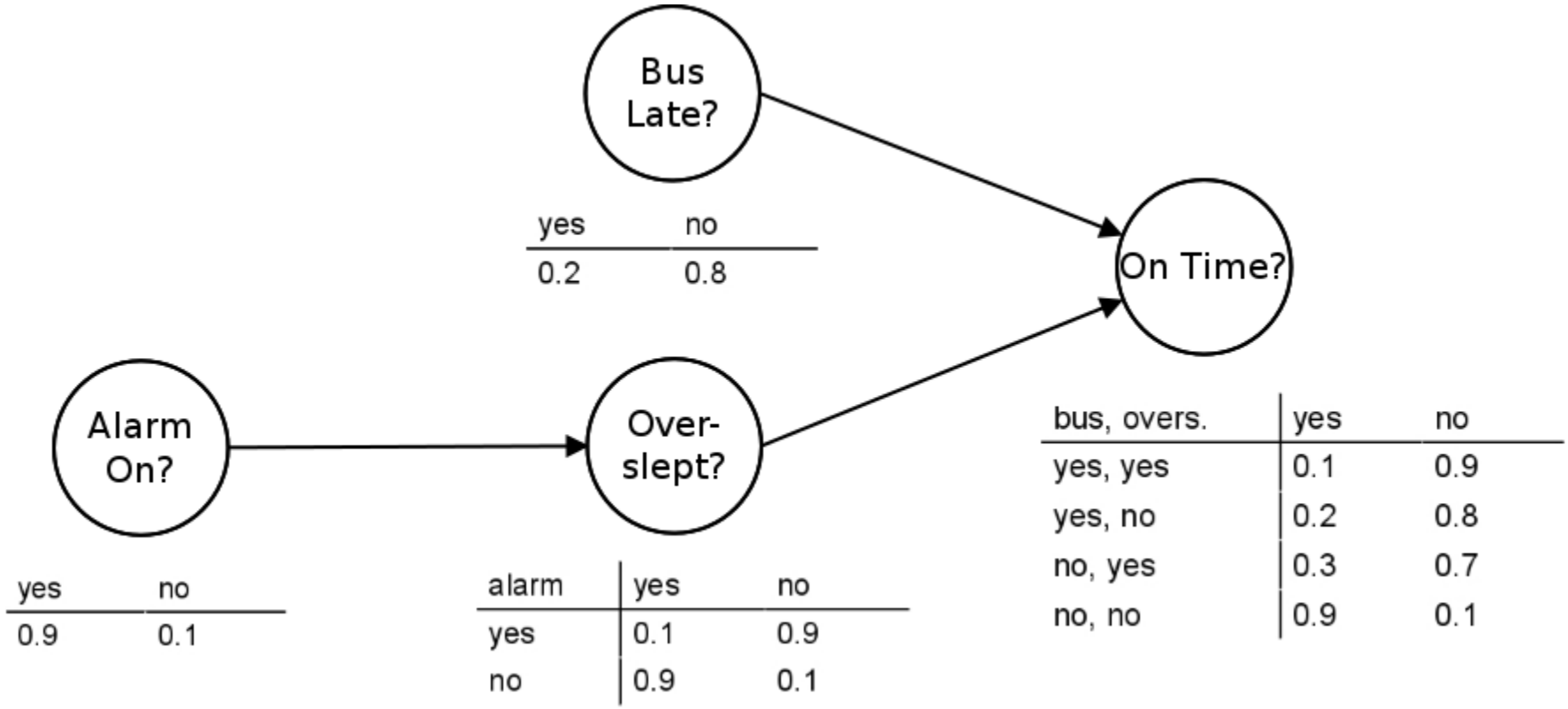
- ***A Bayesian belief network*** (Bayesian network for short) represents the joint probability distribution for a set of variables.
- Each variable in the joint space is represented by a node in the Bayesian network.
- For each variable two types of information are specified.
- First, the network arcs represent the assertion that the variable is conditionally independent of its nondescendants in the network given its immediate predecessors in the network. We say Xj is a ***descendant*** of ***Y*** if there is a directed path from ***Y*** to X.
- Second, a conditional probability table is given for each variable, describing the probability distribution for that variable given the values of its immediate predecessors

# Bayesian Belief Networks

Consider the following example data set with four Attributes and Draw the Bayesian Belief Network

| Alarm status | Over Slept | Bus late | Arrived on time |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

# Bayesian Belief Networks



# Bayesian Belief Networks

- The joint probability for any desired assignment of values ***(yl,*** . . . , ***y,)*** to the tuple of network variables ***(YI*** . . . ***Y,)*** can be computed by the formula

$$P(y_1, \ldots, y_n) = \prod_{i=1}^{n} P(y_i | Parents(Y_i))$$

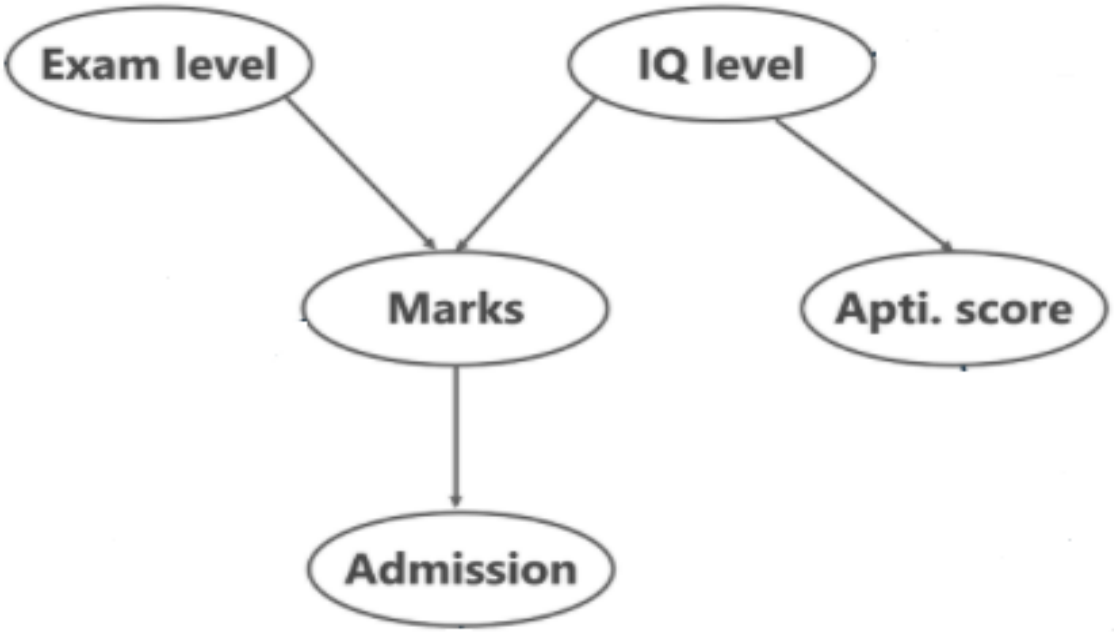
where ***Parents(Yi)*** denotes the set of immediate predecessors of ***Yi*** in the network.

# Example1 for Bayesian Belief Networks

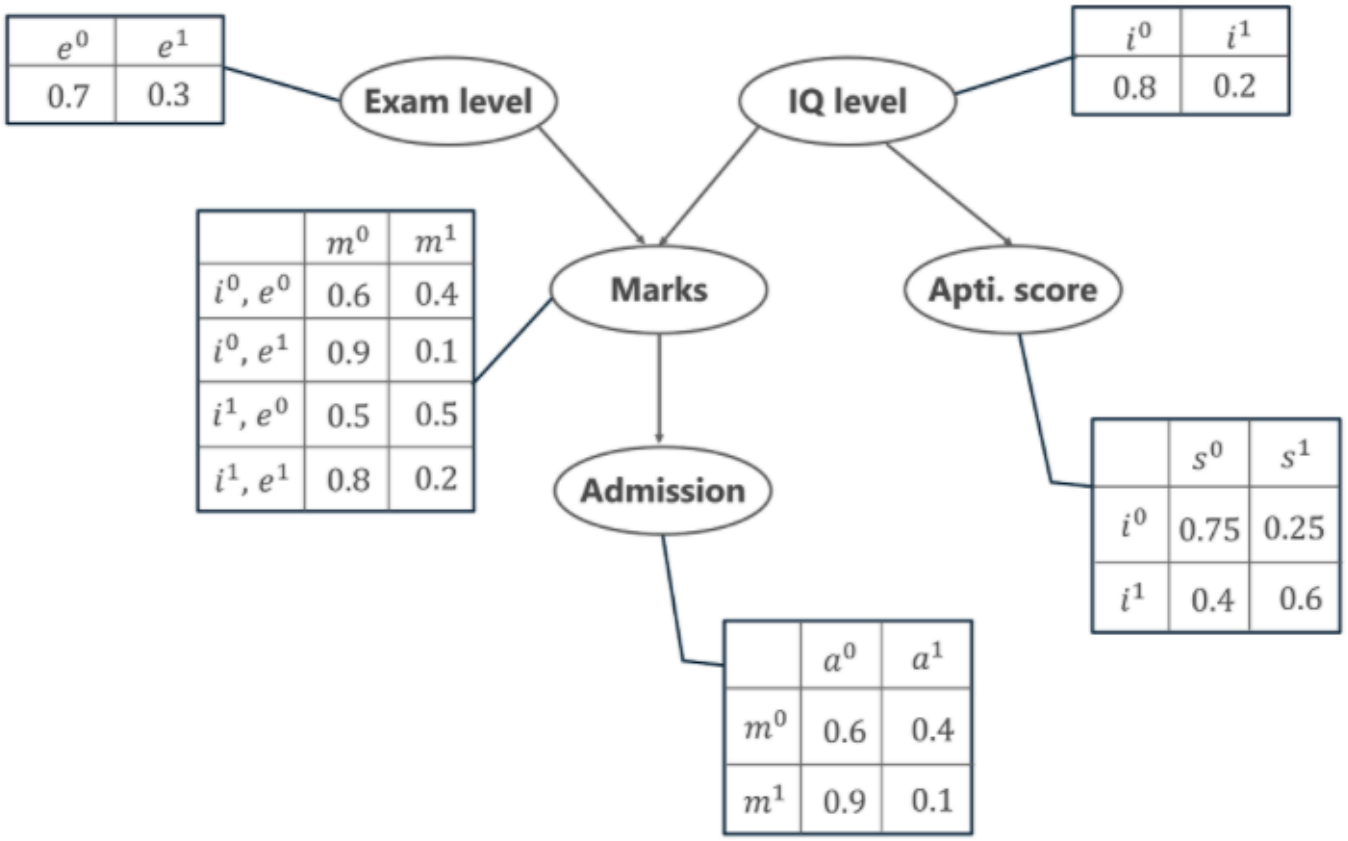
- The dataset of university admissions is given with attributes: Exam level, SIQ level, Apt score, Marks, Admission status.

- Marks depend upon two variables. They are,
  - Exam Level (***e***)– This discrete variable denotes the difficulty of the exam and has two values (0 for easy and 1 for difficult)
  - IQ Level (***i***) – This represents the Intelligence Quotient level of the student and is also discrete in nature having two values (0 for low and 1 for high)
- Additionally, the IQ level of the student also leads us to another variable, which is the Aptitude Score of the student (***s***).
- Now, with marks the student has scored, he can secure admission to a particular university.

# Example1 for Bayesian Belief Network



# Example1 for Bayesian Belief Networks




The Joint Probability Distribution of the 5 variables the formula is given by,

**P[a, m, i, e, s]= P(a | m) . P(m | i, e) . P(i) . P(e) . P(s | i)**

# Example1 for Bayesian Belief Networks

- **Q 1:** Calculate the probability that in spite of the exam level being difficult, the student having a low IQ level and a low Aptitude Score, manages to pass the exam and secure admission to the university.

P[a=1,m=1,i=0,e=1,s=0] =

=P(a=1 | m=1) . P(m=1 | i=0, e=1) . P(i=0) . P(e=1) . P(s=0 | i=0)

=0.1 * 0.1 * 0.8 * 0.3 * 0.75

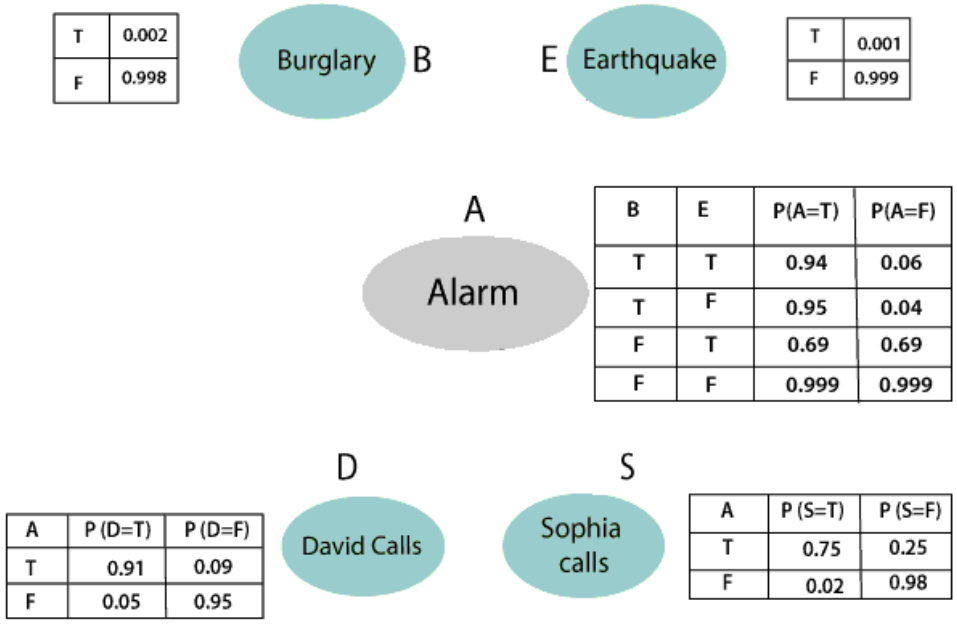
=0.0018

# Example1 for Bayesian Belief Networks

- **Q2:** In another case, calculate the probability that the student has a High IQ level and Aptitude Score, the exam being easy yet fails to pass and does not secure admission to the university.

# Example2 for Bayesian Belief Networks

- Harry installed a new burglar alarm at his home to detect burglary. The alarm reliably responds at detecting a burglary but also responds for minor earthquakes. Harry has two neighbors David and Sophia, who have taken a responsibility to inform Harry at work when they hear the alarm. David always calls Harry when he hears the alarm, but sometimes he got confused with the phone ringing and calls at that time too. On the other hand, Sophia likes to listen to high music, so sometimes she misses to hear the alarm. Here we would like to compute the probability of Burglary Alarm.




| T | 0.002 |
|---|---|
| F | 0.998 |

| T | 0.001 |
|---|---|
| F | 0.999 |

| B | E | P(A=T) | P(A=F) |
|---|---|---|---|
| T | T | 0.94 | 0.06 |
| T | F | 0.95 | 0.04 |
| F | T | 0.69 | 0.69 |
| F | F | 0.999 | 0.999 |

| A | P (D=T) | P (D=F) |
|---|---|---|
| T | 0.91 | 0.09 |
| F | 0.05 | 0.95 |

| A | P (S=T) | P (S=F) |
|---|---|---|
| T | 0.75 | 0.25 |
| F | 0.02 | 0.98 |

# Example2 for Bayesian Belief Networks

- **Q1:** Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and David and Sophia both called the Harry.

- **Q2:** Calculate the probability that alarm has sounded, but there is a burglary, no earthquake occurred, and David called the harry and Sophia not called the Harry.

- **Q3:** Calculate the probability that alarm has not sounded, but there is a burglary, no earthquake occurred, and David & Sophia both not called the Harry.

# Topics

- **Bayes Theorem**
- **Bayes Optimal Classifier**
- **Naïve Bayes Classifier**
- **Bayesian Belief Networks**