Naïve Bayes Classifier

CDS503: Machine Learning

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

Thomas Bayes
1702 - 1761

Topic 5

Classification Parametric - **Bayes**

Assoc. Prof. DR UMI KALSOM YUSOF
SCHOOL OF COMPUTER SCIENCES
UNIVERSITI SAINS MALAYSIA (USM)

Contents

- Conditional Probability
- Bayes Theorem
- Classifying with Probability Theory
- Naïve Bayes Classifier
- Types of Naïve Bayes Algorithm
- Bayesian Belief Network

Conditional Probability

What is Conditional Probability

- In probability theory, conditional probability is a measure of the probability of an event given that another event has already occurred.

- If the event of interest is A and the event B is assumed to have occurred, "the conditional probability of A given B", or "the probability of A under the condition B", is usually written as P(A|B), or sometimes PB(A).

Example of Conditional Probability

Chances of cough

The probability that any given person has a cough on any given day maybe only 5%.

But if we know or assume that the person has a cold, then they are much more likely to be coughing.

The conditional probability of **coughing** given that person have a cold might be a much higher 75%.

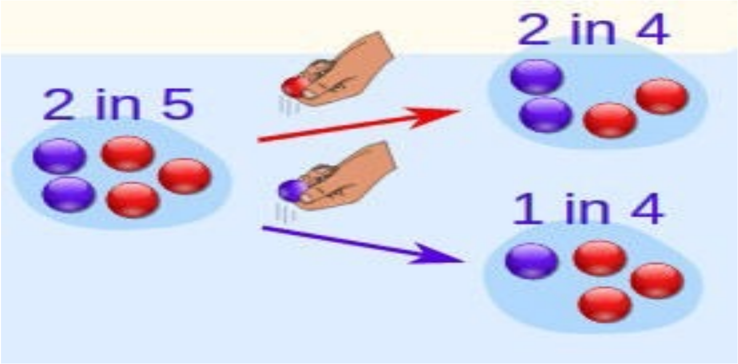
Marbles in a Bag

2 blue and 3 red marbles are in a bag.

What are the chances/probability of getting a blue marble?

Answer: -The chance is 2 in 5

Marbles in a Bag



But after taking one out of these chances,

situation may change!

So the next time:

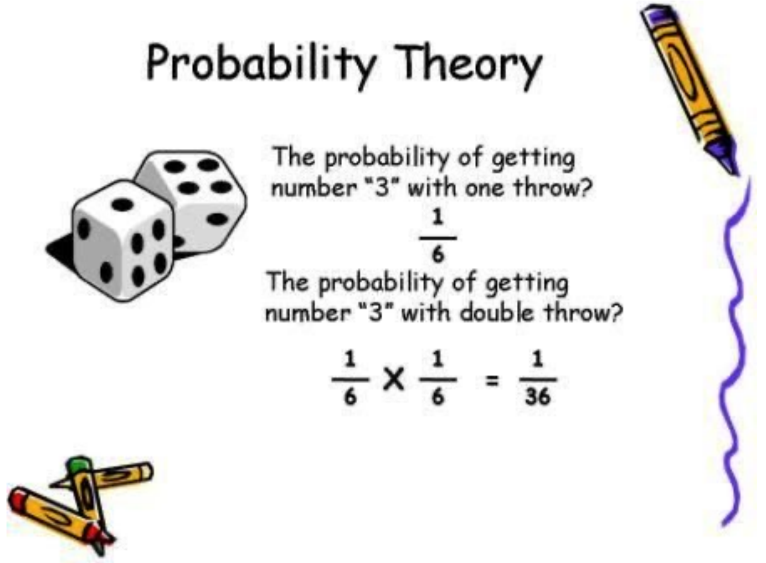
1. if we got a red marble before, then the chance of a blue marble next is 2 in 4
2. if we got a blue marble before, then the chance of a blue marble next is 1 in 4

Bayes Theorem

Bayes Theorem

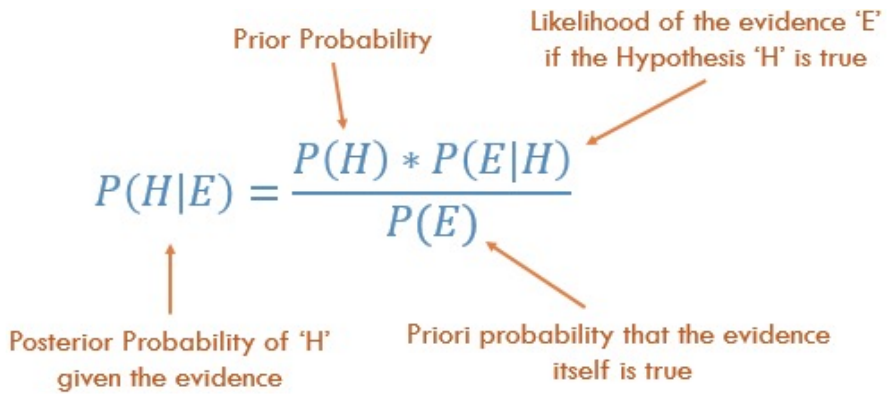
- In probability theory and statistics, Bayes' theorem (alternatively Bayes' law or Bayes' rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

- For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately to assess the probability that they have cancer, compared to the assessment of the probability of cancer made without knowledge of the person's age.



The Formula for Bayes' theorem

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$



where

1. **P(H|E)** is the probability of the hypothesis given that the evidence is there.
2. **P(H)** is the probability of hypothesis H being true. This is known as the prior probability.
3. **P(E)** is the probability of the evidence (regardless of the hypothesis).
4. P(E|H) is the probability of the evidence given that hypothesis is true.

Bayesian Theorem: Basics

- Let **X** be a data sample ("*evidence*"): class label is unknown
- Let H be a *hypothesis* that X belongs to class C
- Classification is to determine P(H|X), the probability
- that the hypothesis holds given the observed data sample X
- **P(H)** (*prior probability*), the initial probability
  - E.g., **X** will buy computer, regardless of age, income, …
- P(**X**): probability that sample data is observed
- P(**X**|H) (*posteriori probability*), the probability of observing the sample **X**, given that the hypothesis holds
  - E.g., Given that **X** will buy computer, the probability that X is age 31..40, medium income, student

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

Bayesian Theorem

- Posteriori probability of a hypothesis H, *given* training data **X**, P(H|**X**), follows the Bayes theorem

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

- Informally, this can be written as

  posteriori = likelihood x prior/evidence

- Predicts **X** belongs to C_i iff the probability $P(C_i|\mathbf{X})$ is the highest among all the $P(C_k|X)$ for all the *k* classes
- Practical **difficulty**: require initial knowledge of many probabilities, significant computational cost

Example of Bayes

Training Dataset

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

Q: Is this person with these information will be buying the computer or not?

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = Fair)

age	income	student	credit_rating	s_com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

The Example for Bayes' theorem

Suppose there are three bowls B1, B2, B3 and bowl B1 has 2 red and 4 blue coins; bowl B2 has 1 red and 2 blue coins; bowl B3 contains 5 red and 4 blue coins.

Bowl	Red	Blue	Total
B1	2	4	6
B2	1	2	3
B3	5	4	9

Suppose the probabilities for selecting the bowls is not the same but are:-

P(B1) = 1/3
P(B2) = 1/6
P(B3) = 1/2

B1 B2 B3

Now, assuming that a red coin was drawn;
What will be the probability that it came from bowl B1?

The Example for Bayes' theorem

We need to find out **P(B1|R) = ???**

And according to Bayes' theorem

$$P(B1|R) = \frac{P(R|B1) * P(B1)}{P(R)}$$

Bayes Theorem

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

For that first, we need to calculate some probabilities which are:-

1. Probability to select a red coin i.e **P(R)**
2. Probability to select the bowl 1 (B1) i.e **P(B1)** which is already given 1/3
3. Probability to select a red coin from B1 i.e **P(R|B1)**

The Example for Bayes' theorem

Bowl	Red	Blue	Total
B1	2	4	6
B2	1	2	3
B3	5	4	9

P(B1) = 1/3
P(B2) = 1/6
P(B3) = 1/2

1. **P(R)** = P(B1 ∩ R) + (B2 ∩ R) + P(B3 ∩ R)

where

- P(B1 ∩ R) is probability to select bowl 1 and red coin
- P(B2 ∩ R) is probability to select bowl 2 and red coin
- P(B3 ∩ R) is probability to select bowl 3 and red coin

= P(selecting B1) * P(Number of Red coins in B1 / total number of coins in B1) + P(selecting B2) * P(Number of Red coins in B2 / total number of coins in B2) + P(selecting B3) * P(Number of Red coins in B3/ total number of coins in B3)

$= (1/3) * 2/6 + (1/6) * 1/3 + (1/2)* 5/9$

$= 4/9$

So P(R) = 4/9

The Example for Bayes' theorem

3. **P(R|B1)**

The probability of selecting a red coin given that it will be drawn from B1

is **2/6 P(B1)** was given i.e 1/3.

By putting all the values in formula:

P(B1|R) = P(R|B1) * P(B1) / P(R)

P(B1|R) = (2/6 *1/3) / 4/9 = 2/8 = 0.25

So we can say that if a red coin was drawn that it will be 25% chances that it was drawn from bowl 1 i.e B1.

Bayesian Classification: Why?

- A statistical classifier: performs *probabilistic prediction*, *i.e.*, predicts class membership probabilities
- Foundation: Based on Bayes' Theorem.
- Performance: A simple Bayesian classifier, *naïve Bayesian classifier*, has comparable performance with decision tree and selected neural network classifiers
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data

Towards Naïve Bayesian Classifier

- Let **D** be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector **X** = (x_1, x_2, …, x_n)
- Suppose there are *m* classes C_1, C_2, …, C_m.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- This can be derived from Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only needs to be maximized

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

Derivation of Naïve Bayes Classifier

- A simplified assumption: attributes are conditionally independent (i.e., no dependence relation between attributes):

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

- This greatly reduces the computation cost: Only counts the class distribution
- If A_k is categorical,
  - $P(x_k|C_i)$ is the # of tuples in C_i having value x_k for A_k divided by $|C_{i,D}|$ (# of tuples of C_i in D)
- If A_k is continous-valued,
  - $P(x_k|C_i)$ is usually computed based on Gaussian distribution with a mean μ and standard deviation σ and $P(x_k|C_i)$ is

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

Naïve Bayes Classifier

What is Naïve Bayes

- Naive Bayes is a kind of classifier which uses the Bayes Theorem.
- It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class.
- The class with the highest probability is considered as the most likely class. This is also known as **Maximum A Posteriori (MAP).**

MAP(H)
= max (P(H|E))
= max ((P(E|H)*P(H))/P(E))
= max (P(E|H)*P(H))

Bayes Theorem

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

P(E) is evidence probability, and it is used to normalize the result.

What is Naïve Bayes

Naive Bayes classifier assumes that all the features are unrelated to each other.

Presence or absence of a feature does not influence the presence or absence of any other feature.

"A fruit may be considered to be an apple if it is red, round, and about 4" in diameter.

Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple."

What is Naïve Bayes

- In real datasets, we test a hypothesis given multiple evidence (feature). So, calculations become complicated.
- To simplify the work, the feature independence approach is used to 'uncouple' multiple evidence and treat each as an independent one.

max ((P(E|H)*P(H))/P(E))

$$P(H|\text{Multiple Evidences}) = \frac{P(E_1|H) * P(E_2|H) \ldots\ldots * P(E_n|H) * P(H)}{P(\text{Multiple Evidences})}$$

Example of Naïve Bayes

For understanding a theoretical concept, the best procedure is to try it on an example.

Let's consider a training dataset with 1500 records and 3 classes.
We presume that there are no missing values in our data.

We have 3 classes associated with Animal Types:

- Parrot
- Dog
- Fish.



Example of Naïve Bayes

The Predictor features set consists of 4 features as

- Swim
- Wings
- Green Color
- Dangerous Teeth

Swim,Wings,Green Color, Dangerous Teeth.
All the features are categorical variables with either of the 2 values:

T(True) or F(False).

Example of Naïve Bayes

Swim	Wings	Green	Teeth	Animal Type
50/500	500/500	400/500	0	Parrot 
450/500	0	0	500/500	Dog 
500/500	0	100/500	50/500	Fish 

The table shows a frequency table of our training data.

Parrots have:

- 50 (10%) parrot can swim,
- 500 out of 500 (100%) parrots have wings,
- 400 out of 500 (80%) parrots are Green and
- 0 (0%) parrots have Dangerous Teeth.

Dogs shows that:

- 450 out of 500 (90%) can swim,
- 0 (0%) dogs have wings,
- 0 (0%) dogs are of Green color and
- 500 out of 500 (100%) dogs have Dangerous Teeth.

Fishes shows that:

- 500 out of 500 (100%) can swim,
- 0 (0%) fishes have wings,
- 100 (20%) fishes are of Green color and
- 50 out of 500 (10%) Fishes have Dangerous Teeth.

Example of Naïve Bayes

Now, it’s time to work on predict classes using the Naive Bayes model.

We have taken 2 records that have values in their feature set, but the target variable needs to predicted.

	Swim	Wings	Green	Teeth
1	True	False	True	False
2	True	False	True	True

Example of Naïve Bayes

- We have to predict animal type using the feature values. We have to predict whether the animal is a Dog, a Parrot or a Fish

P(H|Multiple Evidences) =

$$P(E_1|H) * P(E_2|H) \ldots\ldots * P(E_n|H) * P(H) / P(\text{Multiple Evidences})$$

Let's consider the first record.

The Evidence here is Swim & Green.

The Hypothesis can be an animal type to be Dog, Parrot, Fish.

Example of Naïve Bayes

For Hypothesis testing for the animal to be a **Dog**:

P(Dog | Swim, Green) =
P(Swim|Dog) * P(Green|Dog) * P(Dog) / P(Swim, Green)
= 0.9 * 0 * 0.333 / P(Swim, Green) = 0

For Hypothesis testing for the animal to be a **Parrot**:

P(Parrot| Swim, Green) =
P(Swim|Parrot) * P(Green|Parrot) * P(Parrot) / P(Swim, Green)
= 0.1 * 0.80 * 0.333 / P(Swim, Green)
= 0.0264/ P(Swim, Green)

For Hypothesis testing for the animal to be a **Fish**:
P(Fish| Swim, Green) =
P(Swim|Fish) * P(Green|Fish) * P(Fish) / P(Swim, Green)
= 1 * 0.2 * 0.333 / P(Swim, Green)
= 0.0666/ P(Swim, Green)

Example of Naïve Bayes

The denominator of all the above calculations is same i.e, P(Swim, Green).

The value of P(Fish| Swim, Green) is greater that P(Parrot| Swim, Green).

Using Naive Bayes, we can predict that the class of this record is Fish.

Class Exercise

Example of Naïve Bayes

Let’s consider the second record.

The Evidence here is Swim, Green & Teeth.

The Hypothesis can be an animal type to be Dog, Parrot, Fish.

Class Exercise

Solution

The Evidence here is Swim, Green & Teeth

For Hypothesis testing for the animal to be a **Dog**:
P(Dog | Swim, Green, Teeth) = P(Swim|Dog) * P(Green|Dog) * P(Teeth|Dog) * P(Dog) / P(Swim, Green, Teeth)
= 0.9 * 0 * 1 * 0.333 / P(Swim, Green, Teeth) = 0

For Hypothesis testing for the animal to be a **Parrot**:
P(Parrot| Swim, Green, Teeth) = P(Swim|Parrot) * P(Green|Parrot)* P(Teeth| Parrot) * P(Parrot) / P(Swim, Green, Teeth)
= 0.1 * 0.80 * 0 *0.333 / P(Swim, Green, Teeth)
= 0

For Hypothesis testing for the animal to be a **Fish**:
P(Fish|Swim, Green, Teeth) = P(Swim|Fish) * P(Green|Fish) * P(Teeth|Fish) *P(Fish) / P(Swim, Green, Teeth)
= 1 * 0.2 * 0.1 * 0.333 / P(Swim, Green, Teeth)
= 0.00666 / P(Swim, Green, Teeth)

Solution

The value of P(Fish| Swim, Green, Teeth) is the only positive value greater than 0.

Using Naive Bayes, we can predict that the class of this record is **Fish**.

Another Example of Naïve Bayes

Class:

C1:buys_computer = 'yes'

C2:buys_computer = 'no'

Q: Is this person with these information will be buying the computer or not?

X = (age <=30,

Income = medium,

Student = yes

Credit_rating = Fair)

Training Dataset

age	income	student	credit_rating	s_comp
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayes Classifier

Q: Is this person with these information will be buying the computer or not?

X = (age <=30, Income = medium, Student = yes Credit_rating = Fair)

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

1. $P(C_i)$: P(buys_computer = "yes") = 9/14 = **0.643**
   P(buys_computer = "no") = 5/14 = **0.357**

2. Compute $P(X|C_i)$ for each class
   P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
   P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
   P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
   P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
   P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
   P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
   P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
   P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

3. To calculate **X =** (age <= 30 , income = medium, student = yes, credit_rating = fair)

$P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = **0.019**

$P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.044 * 0.643 = **0.028**
P(X|buys_computer = "no") * P(buys_computer = "no") = 0.019 * 0.357 = **0.007**

Therefore, X belongs to class **("buys_computer = yes")**

Another Class Exercise

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

Let's consider the second person:

Q: Is this person with these information will be buying the computer or not?

X = (age =35,
Income = high,
Student = No
Credit_rating = Excellent)

Types of Naïve Bayes Algorithm

1. **Gaussian:** It is used in classification and it assumes that features follow a normal distribution.
2. **Bernoulli:** The binomial model is useful if your feature vectors are binary (i.e. zeros and ones).
   - One application would be text classification with 'bag of words' model where the 1s & 0s are "word occurs in the document" and "word does not occur in the document" respectively.
3. **MultiNomial** Naive Bayes is preferred to use on data that is multinomially distributed.
   - let's say, we have a text classification problem. Here we can consider bernoulli trials which is one step further and instead of "word occurring in the document", we have "count how often word occurs in the document", you can think of it as "number of times outcome number x_i is observed over the n trials".

Advantages vs Disadvantages

Advantages

- Easy to implement
- Good results obtained in most of the cases

Disadvantages

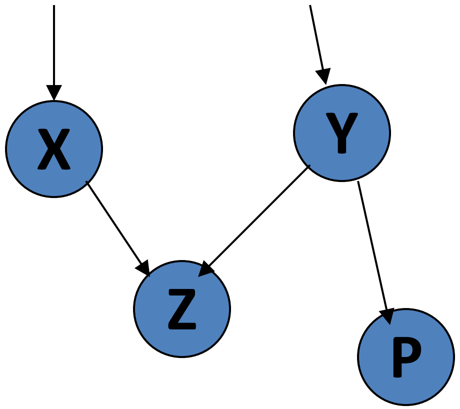
- Assumption: class conditional independence, therefore loss of accuracy
- Practically, dependencies exist among variables
  - E.g., hospitals: patients: Profile: age, family history, etc.
  - Symptoms: fever, cough etc., Disease: lung cancer, diabetes, etc.
  - Dependencies among these cannot be modeled by Naïve Bayesian Classifier

How to deal with these dependencies?

Bayesian Belief Networks

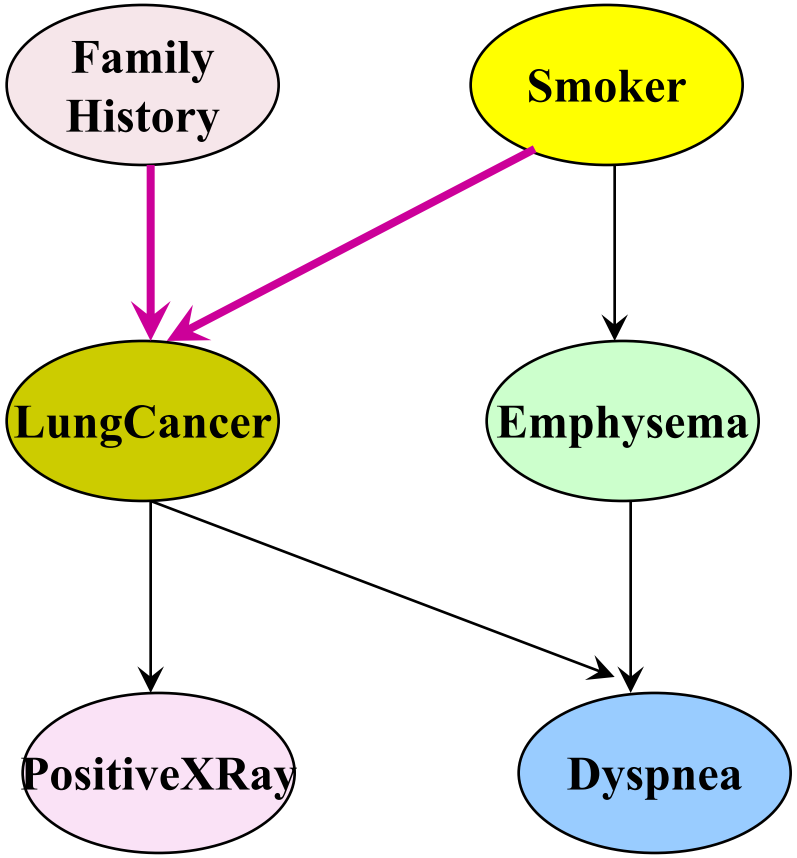
Bayesian Belief Networks

- Bayesian belief network allows a *subset* of the variables conditionally independent
- A graphical model of causal relationships
  - Represents dependency among the variables
  - Gives a specification of joint probability distribution



- Nodes: random variables
- Links: dependency
- X and Y are the parents of Z, and Y is the parent of P
- No dependency between Z and P
- Has no loops or cycles

Bayesian Belief Network: An Example



Bayesian Belief Networks

The **conditional probability** table (CPT) for variable LungCancer:

	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

CPT shows the conditional **probability** for each possible combination of its parents

Derivation of the probability of a particular combination of values of **X**, from CPT:

$$P(x_1,\ldots,x_n) = \prod_{i=1}^{n} P(x_i \mid Parents(Y_i))$$

Simple Example – Wet Grass

You wake up in the morning, see the grass is wet! Wondering what is the cause?

Suppose that there are two events which could cause grass to be wet: either it's raining or the sprinkler is on.

Also, suppose that the rain has a direct effect on the use of the sprinkler (namely that when it rains, the sprinkler is usually not turned on).

Bayesian Belief Network

Simple Example – Wet Grass

Then the situation can be modeled with a **Bayesian network**.

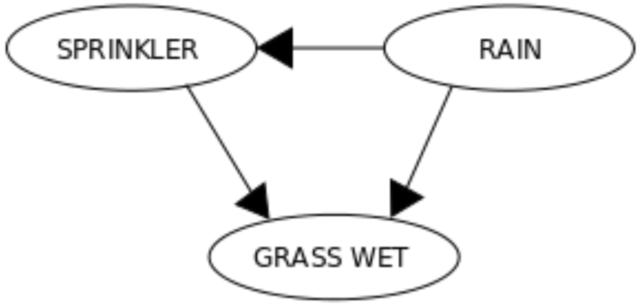
All three variables have two possible values, **T** (for true) and **F** (for false).

The joint probability function is:

$$P(G, S, R) = P(G \mid S, R)P(S \mid R)P(R)$$

where the names of the variables have been abbreviated to *G = Grass wet (yes/no)*, *S = Sprinkler turned on (yes/no)*, and ***R = Raining (yes/no)***

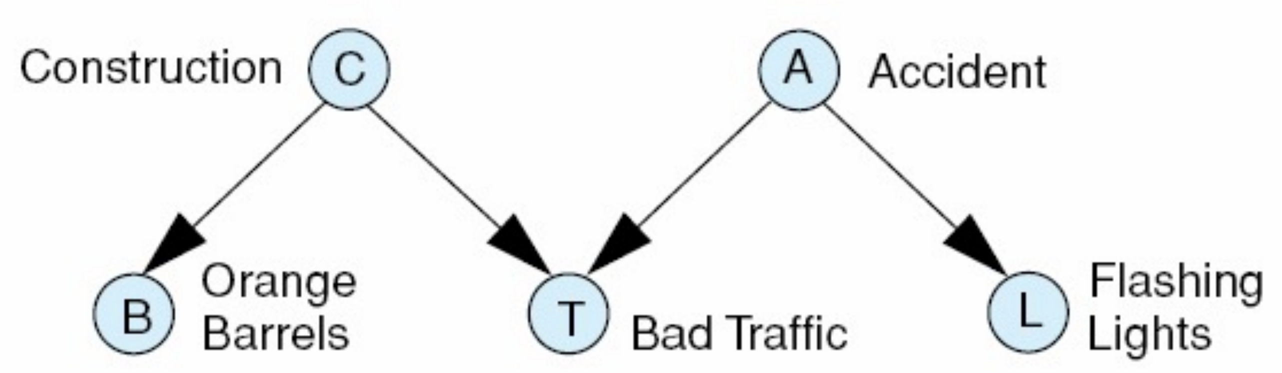
	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Bayesian Belief Network: the Road/Traffic



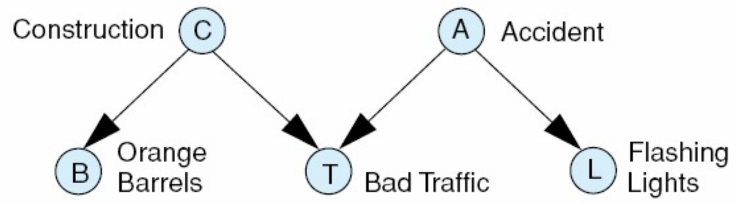
- The computation of the **joint probability** of all the parameters (C, A, B, T, L) using the full Bayesian approach, is as follows:

$$P(C, A, B, T, L) = P(C) \times P(A|C) \times P(B|C,A) \times P(T|C,A,B) \times P(L|C,A,B,T)$$

Bayesian Belief Network: the Road/Traffic

$$P(C, A, B, T, L) = P(C) \times P(A|C) \times P(B|C,A) \times P(T|C,A,B) \times P(L|C,A,B,T)$$

- $2^5 = 32$ parameters in this joint probability (full)
- If we have many **parameters**, the problem becomes complex.



- How to solve this problem?
  - Assumption: parameters of this problem are only dependent on the probabilities of their parents.
    - $P(C, A, B, T, L) = P(C) \times P(A) \times P(B|C) \times P(T|C,A) \times P(L|A)$
  - Also called conditional probability:
    - measures the probability of an event given that (by assumption) another event has occurred

Bayesian Belief Networks

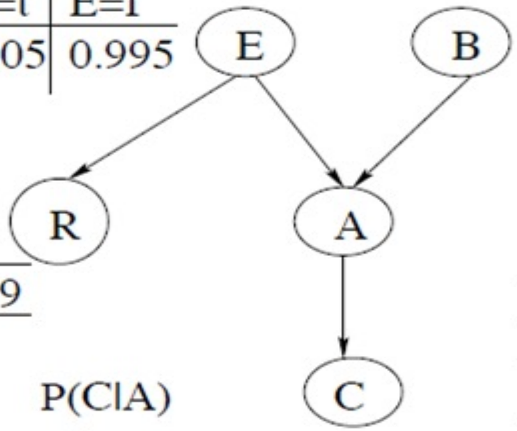
- Another example:

P(E)

E=t	E=f
0.005	0.995

P(B)

B=t	B=f
0.01	0.99



P(R|E)

	R=t	R=f
E=f	0.0001	0.9999
E=t	0.65	0.35

P(A|B,E)

	A=t	A=f
B=f,E=f	0.001	0.999
B=f,E=t	0.3	0.7
B=t,E=f	0.8	0.2
B=t,E=t	0.95	0.05

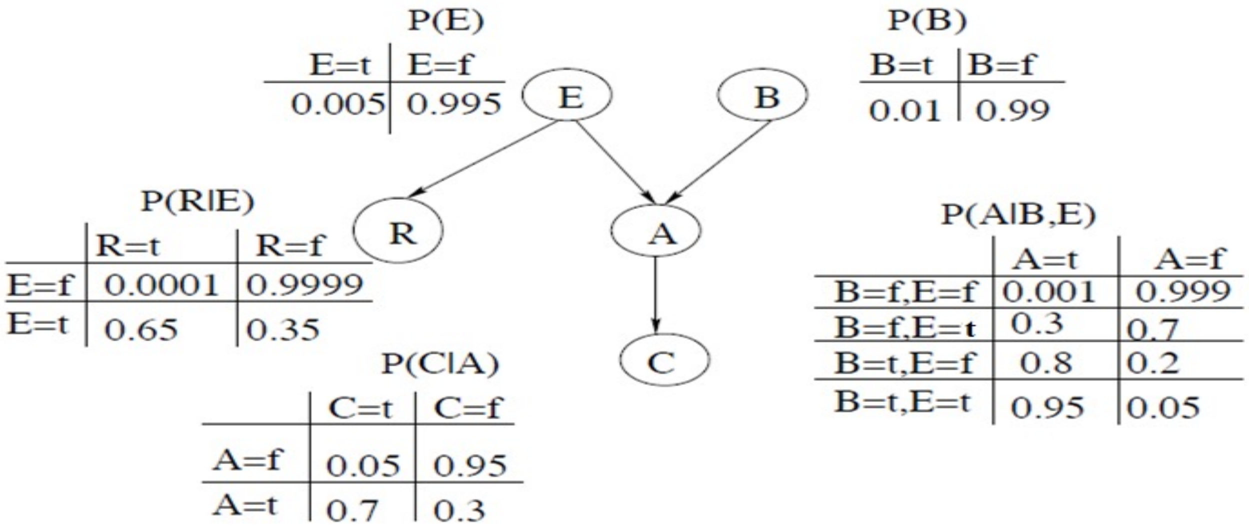
P(C|A)

	C=t	C=f
A=f	0.05	0.95
A=t	0.7	0.3

A Bayesian network is **a directed acyclic graph (DAG),**
G over variables $X_1, \ldots, X_n$, together with a distribution p that factorizes over G.
P is specified as the set of conditional probability distributions associated with G's nodes.

Class Exercise

- Compute P(B = t, E = f, A = t, C = t, R = f) :



- Straight forward.

P(B=t, E=f, A=t, C=t, R=f)

P(B) x P(E) x P(A|B,E) x P(C|A) x P(R|E)

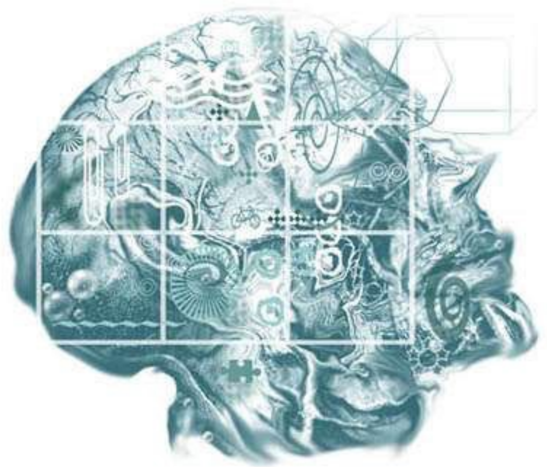
= p(B=t) x p(E=f) x p(A=t|B=t,E=f) x p(C=t|A=t) x p(R=f|E=f)

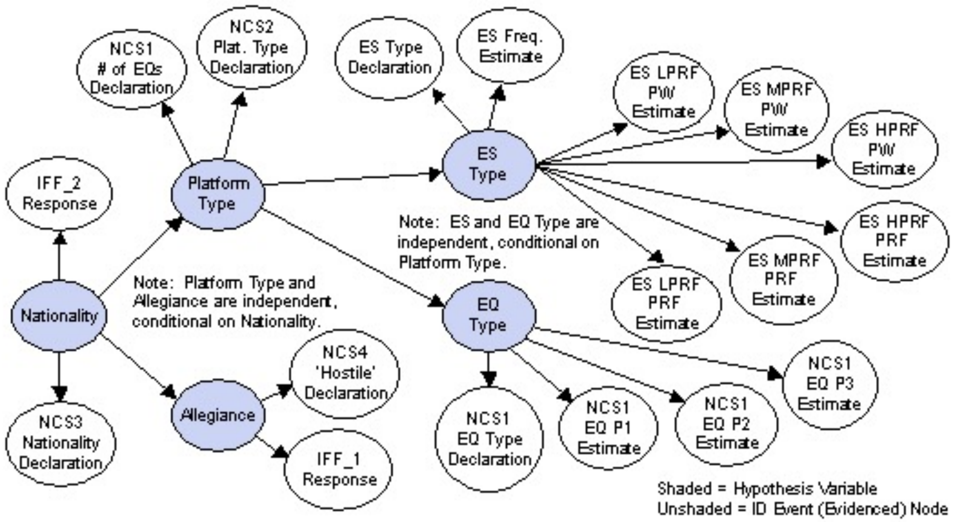
= 0.01 x 0.995 x 0.8 x 0.7 x 0.9999

= 0.0056

Why is Bayes' Theorem so cool?

- Describes what makes something "evidence" and how much evidence it is.
- Science itself is a special case of Bayes' theorem because you are revising a prior probability (hypothesis) in the light of an observation or experience that confirms your hypothesis (experimental evidence) to develop a posterior probability (conclusion)
- Used to judge statistical models and widely applicable in computational biology, medicine, computer science, artificial intelligence, etc.





Thank you