BAYESIAN CLASSIFIERS

Motivation

- Task: predict if a person is at risk of heart disease
- Deciding factors include diet, exercise, excessive smoking, alcohol abuse
- Other factors such as heredity, …

- The class label of a test record cannot be predicted with certainty even though its attribute set is identical to a training record.
- Bayesian classifiers model probabilistic relationships between attributes

Probability Theory

- Probability measures the amount of uncertainty of an event
  - Probability of rain tomorrow
  - Probability of drawing a red ball from a bin containing 6 red and 11 white balls

- Measured as a number between 0 and 1
  - $p(E) = 0$: event E will not occur
  - $p(E) = 1$: event E will occur with certainty

Definitions

- The set of all possible events is called the sample space
  - Forecast space: S = {Rainy, Cloudy, Sunny}
  - Drawing space: S = {Red, White}
- The sum of probabilities of all outcomes of an event is 1:
  - p(Rainy) + p(Cloudy) + p(Sunny) = 1
  - p(Red) + p(White) = 1
- A complimentary event E’ with respect to event E is the event that E does not occur
  - p(E) + p(E’) = 1

Definitions

- Two events are mutually exclusive if they cannot occur together
  - $p(A \cap B) = 0$
- Two events are independent if the chance that each event occurs is independent of the other
- Dependent or independent?
  - Rolling 6 on a die and then rolling 2 on a second roll
  - Picking the first prize winner at a raffle event then picking the second prize winner
- If two events are independent: $p(A \cap B) = p(A)p(B)$

Random Variables

- A variable whose value depends on the outcome of a random experiment
- $P(E)$: the fraction of times E is observed in a potentially unlimited number of experiments
- $P(X=v)$:
  - probability of X having value v
  - probability of all outcomes in which v is observed

Example

- Experiment: toss a coin 4 times
- Let X be the random variable that measures the number of times a head is observed.
- Possible outcomes:

HHHH, HHHT, HHTH, HHTT,
HTHH, HTHT, HTTH, HTTT,
THHH, THHT, THTH, THTT,
TTHH, TTHT, TTTH, TTTT

X	0	1	2	3	4
$P(X)$	1/16	4/16	6/16	4/16	1/16

- *What is* $P(X = 2)$? 6/16
- *What is* $P(X \geq 2)$? 6/16 + 4/16 + 1/16 = 11/16

Continuous Random Variables

- If X can take a continuous range of values:

$$P(a < X < b) = \int_a^b f(x)dx$$

- $f(x)$: probability density function

- $P(X,Y)$: joint probability of two random variables X and Y

If X and Y are independent:

$$P(X,Y) = P(X)P(Y)$$

Conditional Probability

- $P(Y|X)$: conditional probability of Y given X

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)}$$

- If X and Y are independent:

$$P(Y \mid X) = P(Y)$$

Bayes Theorem

- Expresses relationship between conditional probabilities:

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

$$P(X,Y) = P(Y|X)p(X)$$

$$P(X,Y) = P(X|Y)p(Y)$$

Bayes Theorem - Example

- Team 0 wins 65% of the time
- Team 1 wins the remaining matches
- Among games won by Team 0, only 30% come from playing on Team 1's field.
- 75% of victories of Team 1 are obtained at home
- Team 1 will host the next game, who will most likely win?

Bayes Theorem - Example

- Random variables:
  - X: represents the team that will host the game
  - Y: represents the team that will win the game
- Goal: Team 1 will host the next game, who will most likely win?
  - Compute and compare: $P(Y=0 \mid X=1)$ and $P(Y=1 \mid X=1)$

Bayes Theorem - Example

- Team 0 wins 65% of the time:
  - $P(Y=0) = 0.65$
- Team 1 wins the remaining matches:
  - $P(Y=1) = 1 - 0.65 = 0.35$
- Among games won by Team 0, only 30% come from playing on Team 1's field:
  - $P(X=1 \mid Y=0) = 0.3$
- 75% of victories of Team 1 are obtained at home:
  - $P(X=1 \mid Y=1) = 0.75$

Bayes Theorem - Example

Given:

- $P(Y=0) = 0.65$
- $P(Y=1) = 0.35$
- $P(X=1 \mid Y=0) = 0.3$
- $P(X=1 \mid Y=1) = 0.75$

Goal: compute

- $P(Y=0 \mid X=1)$
- $P(Y=1 \mid X=1)$

Solution:

$$
\begin{aligned}
P(Y=1 \mid X=1) &= P(X=1 \mid Y=1)P(Y=1) \,/\, P(X=1) \\
&= P(X=1 \mid Y=1)P(Y=1) \,/\, (\, P(X=1, Y=1) + P(X=1, Y=0) \,) \\
&= P(X=1 \mid Y=1)P(Y=1) \,/\, (\, P(X=1 \mid Y=1)P(Y=1) + P(X=1 \mid Y=0)P(Y=0) \,) \\
&= 0.75\text{x}0.35 \,/\, (0.75\text{x}0.35 + 0.3\text{x}0.65) = 0.5738
\end{aligned}
$$

$$P(Y=0 \mid X=1) = 1 - 0.5738 = 0.4262$$

Classification from Statistical Perspective

- Given a set of attributes X and a class attribute Y
- If Y has nondeterministic relationship with X: treat X, Y as random variables
- In the training phase: learn $P(Y \mid X)$ for every combination of X
- In the test phase: given test record X', find Y' maximizing $P(Y' \mid X')$
- $P(Y \mid X)$: posterior probability of Y
- $P(Y)$: prior probability of Y

Posterior distribution of Y

Prior distribution of Y

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Example

- Task: predict if a borrower will default on his/her payment
- Test record:
  - X = (Home owner = No, Marital Status = Married, Annual Income = $120K)
- Goal: Compute and <u>compare</u> $P(Yes \mid X)$ and $P(No \mid X)$

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Example

- Estimate the posterior probabilities for every X is difficult

- Using Bayes Theorem:
  - $P(Yes \mid X) = P(X \mid Yes)P(Yes) / P(X)$
  - $P(No \mid X) = P(X \mid No)P(No) / P(X)$
  - $P(X)$ is the same in both equations and can be ignored
  - $P(Yes)$ and $P(No)$ can be easily computed from training set

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Example

- Using Bayes Theorem:
  - $P(Yes \mid X) = \boldsymbol{P(X \mid Yes)}P(Yes) \,/\, P(X)$
  - $P(No \mid X) = \boldsymbol{P(X \mid No)}P(No) \,/\, P(X)$

- Remaining sub-problem:
  - Compute $P(X \mid Yes)$ and $P(X \mid No)$: the conditional probability $P(X|Y)$

- 2 Methods:
  - Naïve Bayes Classifier
  - Bayesian Belief Networks

Naïve Bayes Classifier

- Assumes that the attributes X are **conditionally independent** given class label Y

$$X = (X_1, X_2, X_3, \ldots, X_d)$$

$$P(X \mid Y = y) = \prod_{i=1}^{d} P(X_i \mid Y = y) = P(X_1 \mid Y = y) P(X_2 \mid Y = y) \ldots P(X_d \mid Y = y)$$

Naïve Bayes Classifier

- Conditional Independence:

$$P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

$$P(X_1, X_2, \ldots, X_k \mid Y) = P(X_1 \mid Y)P(X_2 \mid Y) \ldots . P(X_k \mid Y)$$

- Idea: instead of computing the class conditional probability for every combination of X, only estimate X_i given Y

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{k} P(X_i \mid Y)}{P(X)}$$

- Find the class (value of Y) that maximizes numerator

Categorical Attributes

- If X_i is categorical attribute, then $P(X_i = x_i \mid Y = y_i)$ = number of instances having both $Y = y_i$ and $X_i = x_i$ divided by number of instances having $Y = y_i$

- P(Home owner = No | No) = 4/7
- P(Marital status = Divorced | Yes) = 1/3

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Continuous Attributes I

- Discretize each continuous attribute then replace each value by its corresponding interval
- $P(X_i \mid Y = y_i)$ = number of instances having both $Y = y_i$ and X_i in the corresponding interval divided by number of instances having $Y = y_i$

Continuous Attributes I

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	100K-130K	No
2	No	Married	75K-100K	No
3	No	Single	60K-75K	No
4	Yes	Married	100K-130K	No
5	No	Divorced	75K-100K	Yes
6	No	Married	60K-75K	No
7	Yes	Divorced	200K-230K	No
8	No	Single	75K-100K	Yes
9	No	Married	60K-75K	No
10	No	Single	75K-100K	Yes

P(Annual Income = 75k-100k | Yes) = 3/3

Continuous Attributes I

- The estimate error depends on the discretization strategy and the number of intervals
- If the number of intervals is too large:
  - Too few records in each interval, so unreliable estimate
- If the number of intervals is too small:
  - May join classes and miss decision boundary

Continuous Attributes II

- Assume a certain probability distribution for the continuous variable
- Estimate parameters of distribution from training sample
- Normal distribution:

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- μ_{ij}: sample mean of attribute X_i of all training records belonging to class y_j
- *σ_{ij}^2: sample variance of same set*
- *Obtain μ_{ij} and σ_{ij}^2 from the training data*

Continuous Attributes II - Example

What is P(Income = 120K|No)?

$\mu_{income,No}$ =(125+100+70+120+60+220+75)/7

= 110

$\sigma^2_{income,No}$ = 2975

$\sigma_{income,No}$ = 54.54

$$P(Income = 120 | No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}}$$

$$= 0.0072$$

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Example of Naïve Bayes Classifier

Test Record: X= (home owner = No, Marital Status = Married, Income = 120K)

More likely to default or not?

Tid	Home Owner (binary)	Marital Status (categorical)	Annual Income (continuous)	Defaulted Borrower (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

P(Home Owner=Yes|No) = 3/7
P(Home Owner=No|No) = 4/7
P(Home Owner=Yes|Yes) = 0
P(Home Owner=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No) = 1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes) = 1/3
P(Marital Status=Married|Yes) = 0

For Annual Income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

Example of Naïve Bayes Classifier

Test Record: X= (home owner = No, Marital Status = Married, Income = 120K)

P(Home Owner=Yes|No) = 3/7
P(Home Owner=No|No) = 4/7
P(Home Owner=Yes|Yes) = 0
P(Home Owner=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No) = 1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes) = 1/3
P(Marital Status=Married|Yes) = 0

For Annual Income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

$$P(Yes \mid X) = \boldsymbol{P(X \mid Yes)P(Yes)} \,/\, P(X)$$
$$P(No \mid X) = \boldsymbol{P(X \mid No)P(No)} \,/\, P(X)$$

P(X | No) = P(Home Owner=No | Class=No)
$\times$ P(Married | Class=No)
$\times$ P(Income=120K | Class=No)
= 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024

P(X | Yes) = P(Home Owner=No | Class=Yes)
$\times$ P(Married | Class=Yes)
$\times$ P(Income=120K | Class=Yes)
= 1 $\times$ 0 $\times$ 1.2 $\times$ 10^{-9} = 0

Since P(X | No)P(No) > P(X | Yes)P(Yes)

Therefore P(No | X) > P(Yes | X)
=> Class = No

Implementation

Functions

- from sklearn.naive_bayes import GaussianNB
- GaussianNB
  - Likelihood of features follows a Gaussian distribution $P(x_i|y) \sim N(\mu_y, \sigma_y)$ for continuous attributes
- BernoulliNB
  - All features are Boolean (True or False, 1 or 0)
- MultinomialNB
  - Multiple classes

Parameters

- Typically not set
- GaussianNB
  - Prior – probability of each class
- BernoulliNB and MultinomialNB
  - fit_prior – learn the class prior probability
  - class_prior – specify probability of each class

Problem with Naïve Bayes Classifier

$$P(Y \mid X) = P\,(X_1|Y)\;P(X_2|Y)\ldots P\,(X_n|Y)P\,(Y)\;/\;P\,(X)$$

Test Record:

X= (home owner = Yes, Marital Status = Divorced, Income = 120K)

$P(X \mid Yes) = 0$ P(home owner = Yes\Y=yes)=0

- If $P(X_i|Y) = 0$ for an attribute X_i then $P(Y|X) = 0$
- If a sample set does not cover all values, the naïve Bayes classifier may not be able to classify some test records

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Solution

- m-estimate approach for estimating conditional probabilities

$$p(x_i \mid y_i) = \frac{n_{xy} + mp}{n_y + m}$$

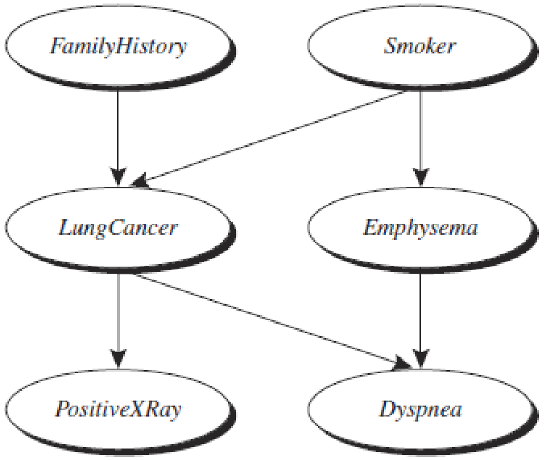
- n_y: total number of instances from class y_i
- n_{xy}: total number of instances from class y_i with value x_i
- p: user specified parameter, prior probability of Y
- m: equivalent sample size parameter

Characteristics

- Robust to noise because noise points averaged in estimations
- Can handle missing values by ignoring records with missing values
- Robust to irrelevant attributes: if X_i is irrelevant, $p(X_i|Y)$ becomes uniformly distributed
- Correlated attributes degrade performance
- Conditional independence may not hold for all attributes
  - Use Bayesian Belief Networks

Bayesian Belief Network

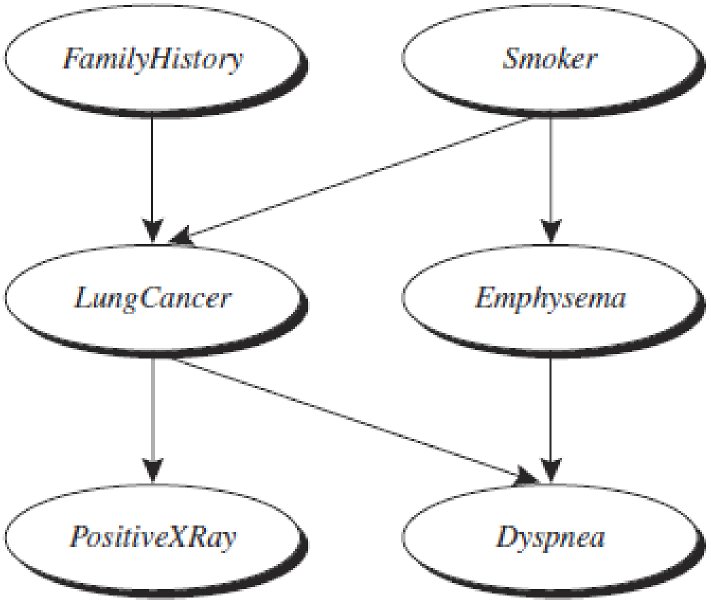
- Specifies the dependencies between attributes
- Two components:
  - A directed acyclic graph: each node represents an attribute
  - A set of conditional probabilities table



	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

Bayesian Belief Network

- *Each variable is conditionally independent of its non descendants given its parents*



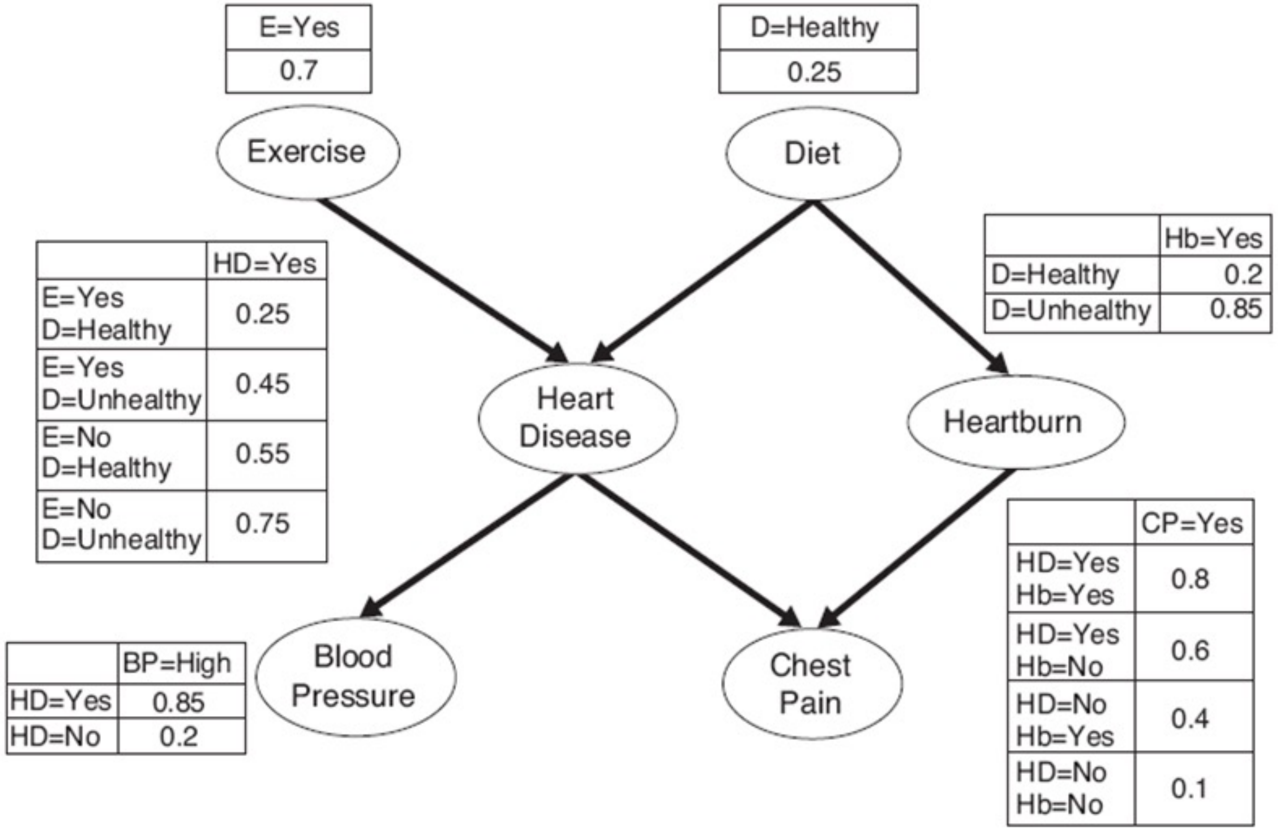
Conditional Probability:
$P(X|\backslash X) = P(X|\pi(X))$

Joint Probability:
P(X) = $\prod_{i=1}^{N} p(X_i|\pi(X_i))$

P(F,S,L,E,P,D) = P(F)P(S)P(L|F,S)P(E|S)P(P|L)P(D|L,E)

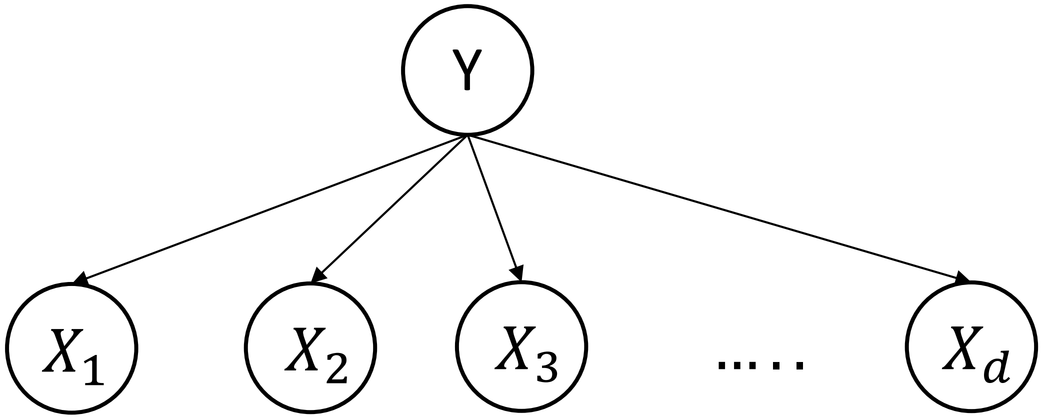
Example

P(E,D,HD,H,B,C) = P(E)P(D)P(HD|E,D)P(H|D)P(B|H)P(C|HD,H)



Naïve Bayes Classifier?

$$P(X \mid Y=y)=\prod_{i=1}^{d} P(X_i \mid Y=y)=P(X_1 \mid Y=y)P(X_2 \mid Y=y)\ldots P(X_d \mid Y=y)$$



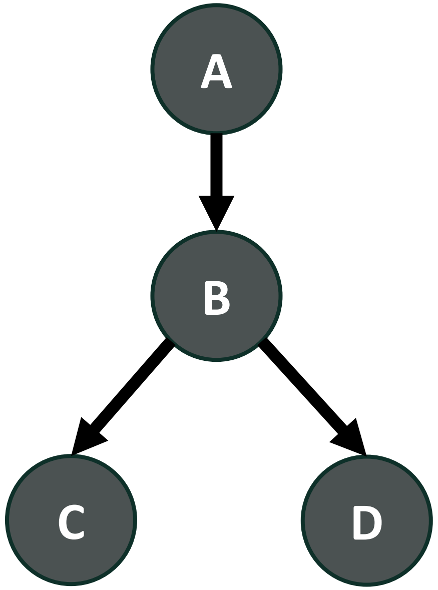
Training Process

- Learn the network topology
  - Constructed by experts
  - Inferred from the data

- If network topology is known:
  - Compute conditional probabilities table

- If network topology is not known:
  - Discrete optimization problem

Prediction



P(A=yes, B=yes, C=yes, D=yes) =
P(A=yes)*P(B=yes | A=yes)*P(C=yes | B=yes)*
P(D=yes | B=yes)

Characteristics

- Captures prior knowledge of a domain using a graphical model
- Network construction may be time consuming
- Well suited for incomplete data
  - Expectation-Maximization (EM) algorithm
- Robust to overfitting
- A popular library in Python is called PyMC3 and provides a range of tools for Bayesian modeling, including graphical models like Bayesian Networks.
- Additionally, BNlearn is a R package with benchmark networks

Bayes Theorem

Posterior distribution of Y

Prior distribution of Y

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

$$\Downarrow$$

$$P(\theta|D) \propto p(D|\theta)p(\theta)$$

θ: parameters; D: dataset

Bayesian Deep Learning