Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability:

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

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

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

The Good Reverend Bayes

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayesian Classifiers

- Consider each attribute and class label as random variables

- Given a record with attributes ($A_1, A_2, \ldots, A_n$)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C| $A_1, A_2, \ldots, A_n$)

- Can we estimate P(C| $A_1, A_2, \ldots, A_n$) directly from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$
  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n \mid C) P(C)$
- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j) P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$
  - Can estimate $P(A_i \mid C_j)$ for all A_i and C_j.
  - New point is classified to C_j if $P(C_j) \Pi P(A_i \mid C_j)$ is maximal.

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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

- Class: $P(C) = N_c/N$
  - e.g., P(No) = 7/10, P(Yes) = 3/10
- For discrete attributes:

  $$P(A_i \mid C_k) = |A_{ik}|/ N_{c_k}$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins
    - one binary attribute per bin
    - violates independence assumption
  - Two-way split: (A < v) or (A > v)
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i|c)$

How to Estimate Probabilities from Data?

Tid	Refund	Marital Status	Taxable Income	Evade
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

- Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  – One for each (A_i,c_i) pair

- For (Income, Class=No):
  – If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income=120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Example of Naïve Bayes Classifier

Given a Test Record:

$$X = (\text{Refund}=\text{No}, \text{Married}, \text{Income}=120\text{K})$$

naive Bayes Classifier:

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

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No) × P(Married| Class=No) × P(Income=120K| Class=No) = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes) × P(Married| Class=Yes) × P(Income=120K| Class=Yes) = 1 × 0 × 1.2 × 10^{-9} = 0

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

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

Example of Naïve Bayes Classifier

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Naïve Bayes (Summary)

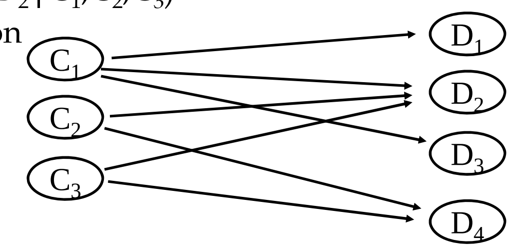
- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)

Sequential Bayesian Inference

- Consider data attributes one by one
  - Prior probabilities $P(C_i)$
  - Observe data D_j
  - Updates priors using Bayes Rule:
  - Repeat for other attributes using the resulting posterior probability as the new prior
- If attributes are conditionally independent, same as doing it all at once
- Allows choice of what attribute to observe (test to perform) next in terms of cost/benefit.

Bipartite Graphs

- Multiple attributes, multiple classifications
- Classifications are probabilistically independent
- Attributes are conditionally independent
- Attribute probabilities depend only the classes exhibiting them
- Attributes with multiple exhibiting classes require joint probabilities $P(D_2 | C_1,C_2,C_3)$
- Information explosion



Noisy OR

- Simplify by assuming only one classification holds at a time
- Probability that all classifications exhibit the attribute is just the probability that at least one does
- Thus an attribute is absent only if no class exhibits it

$$1 - P(D_2 | C_1,C_2,C_3) = (1 - P(D_2 | C_1)) (1 - P(D_2 | C_2)) (1 - P(D_2 | C_3))$$

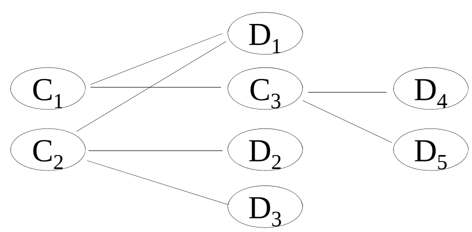
- Use class probabilities for the basic data

$P_c(D_2 | C_1) = P(D_2 | C_1)$, all other C_i absent

- Reduces probability table size: if N classes and K attributes, from $N2^K$ to NK

Polytrees

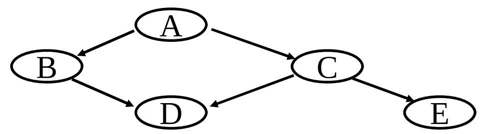
- What if classes are interrelated?
- Polytrees
  - At most one path between any two nodes
- Efficient sequential updating possible



The independence hypothesis…

- … makes computation possible
- … yields optimal classifiers when satisfied
- … but is seldom satisfied in practice, as attributes (variables) are often correlated.
- Attempts to overcome this limitation:
  - Bayesian networks, that combine Bayesian reasoning with causal relationships between attributes
  - Decision trees, that reason on one attribute at the time, considering most important attributes first

Bayesian networks



- Directed acyclic graphs
- Absence of link implies conditional independence

$$P(X_1,\ldots,X_n) = \text{Product } P(X_i \mid \text{parents } (X_i))$$

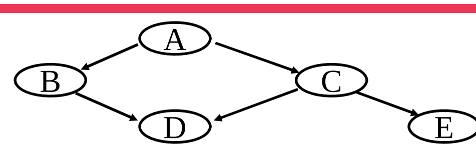
- Specify joint probability tables over parents for each node
- Probability A,B,C,D,E all present:

$$P(A,B,C,D,E) = P(A) * P(B \mid A) * P(C \mid A) * P(D \mid B,C) * P(E \mid C)$$

- Probability A,C,D present and B,E absent:

$$P(A,\neg B,C,D, \neg E) = P(A) * P(\neg B \mid A) * P(C \mid A) * P(D \mid \neg B,C) * P(\neg E \mid C)$$

Computing with partial information



- Probability that A present and E absent:

$$P(A \mid \overline{E}) = \sum_{B,C,D} P(A,B,C,D,\overline{E})$$

$$= \sum_{B,C,D} P(A)P(B \mid A)P(C \mid A)P(D \mid B,C)P(E \mid C)$$

$$= P(A)\sum_{C} P(C \mid A)P(\overline{E} \mid C)\sum_{B} P(B \mid A)\sum_{D} P(D \mid B,C)$$

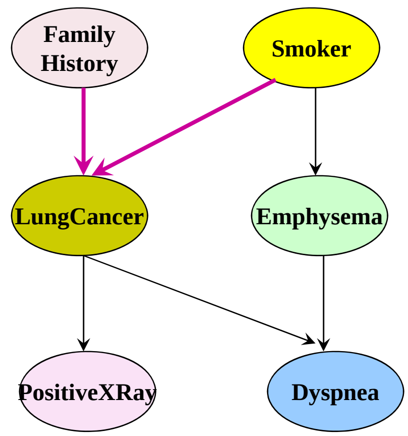
- Graph separators (e.g. C) correspond to factorizations
- General problem of finding separators is NP-hard
  - $P(A \mid \neg E) = P(A, \neg E)/P(\neg E)$

Bayesian networks

- Each node annotated with conditional probability table
  - Probability of node values given values of parent nodes

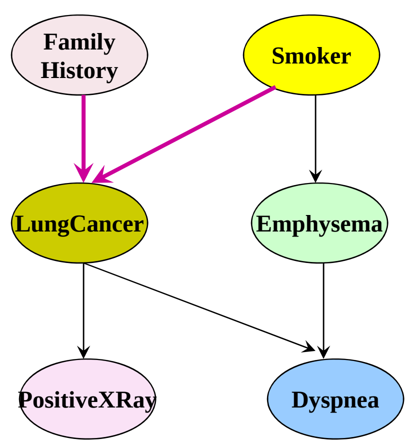
	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

- Conditional probability table for the variable LungCancer



Bayesian networks

- Table of joint probability distribution has $2^6 = 64$ entries
- Bayesian network tables have 8 + 4 + 4 + 8 = 24 entries



Bayesian Belief Networks

- Bayesian belief networks allow subsets of the variables to be conditionally independent
- A graphical model of causal relationships
- Several methods for learning Bayesian belief networks
  - Given both network structure and all the variables: easy
  - Given network structure but only some variables
  - When the network structure is not known in advance