Foundations of Machine Learning

Module 4:
Part A: Probability Basics

Sudeshna Sarkar
IIT Kharagpur

- *Probability* is the study of randomness and uncertainty.
- A *random* experiment is a process whose outcome is uncertain.

  Examples:

  – Tossing a coin once or several times

  – Tossing a die

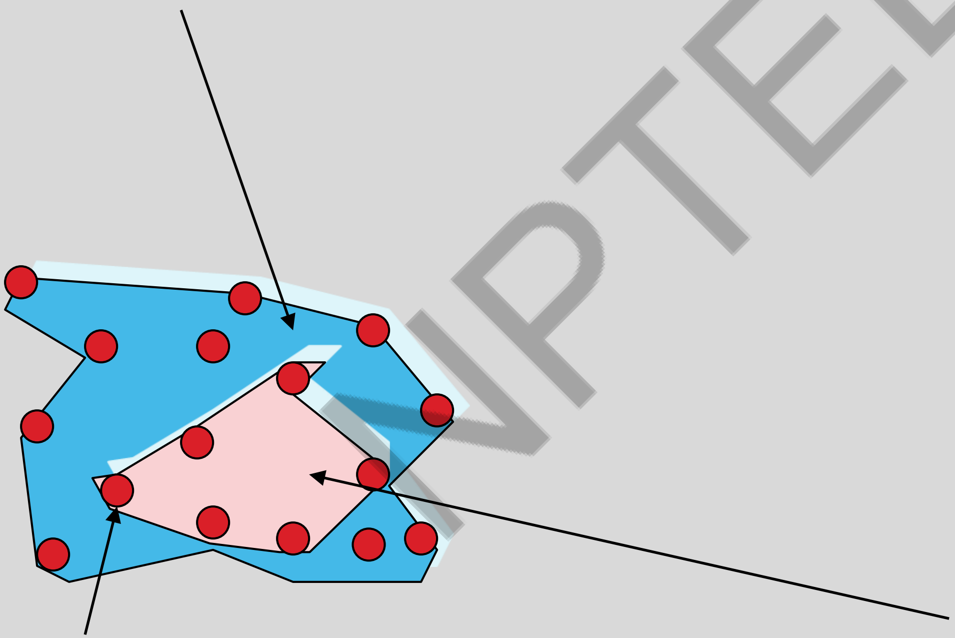
  – Tossing a coin until one gets Heads

  – …

Events and Sample Spaces

Sample Space

The sample space is the set of all possible outcomes.



Simple Events

The individual outcomes are called simple events.

Event

An event is any collection of one or more simple events

Sample Space

- Sample space Ω : the set of all the possible outcomes of the experiment
  - If the experiment is a roll of a six-sided die, then the natural sample space is {1, 2, 3, 4, 5, 6}
  - Suppose the experiment consists of tossing a coin three times.
    $\Omega = \{(hhh, hht, hth, htt, thh, tht, tth, ttt\}$
  - the experiment is the number of customers that arrive at a service desk during a fixed time period, the sample space should be the set of nonnegative integers: $\Omega = Z^{+} = \{0, 1, 2, 3, \ldots\}$

Events

- Events are subsets of the sample space
  - A= {the outcome that the die is even} ={2,4,6}
  - B = {exactly two tosses come out tails}=(htt, tht, tth}
  - C = {at least two heads} = {hhh, hht, hth, thh}

Probability

- A Probability is a number assigned to each event in the sample space.
- Axioms of Probability:
  - For any event A, $0 \leq P(A) \leq 1$.
  - P(Ω) =1 and $P(\phi) = 0$
  - If A_1, A_2, … A_n is a partition of A, then

$$P(A) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

Properties of Probability

- For any event A, $P(A^c) = 1 - P(A)$.
- If $A \subset B$, then $P(A) \leq P(B)$.
- For any two events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

For three events, A, B, and C,

$$P(A \cup B \cup C) =$$

$$P(A) + P(B) + P(C)$$

$$- P(A \cap B) - P(A \cap C) - P(B \cap C)$$

$$+ P(A \cap B \cap C)$$

Intuitive Development (agrees with axioms)

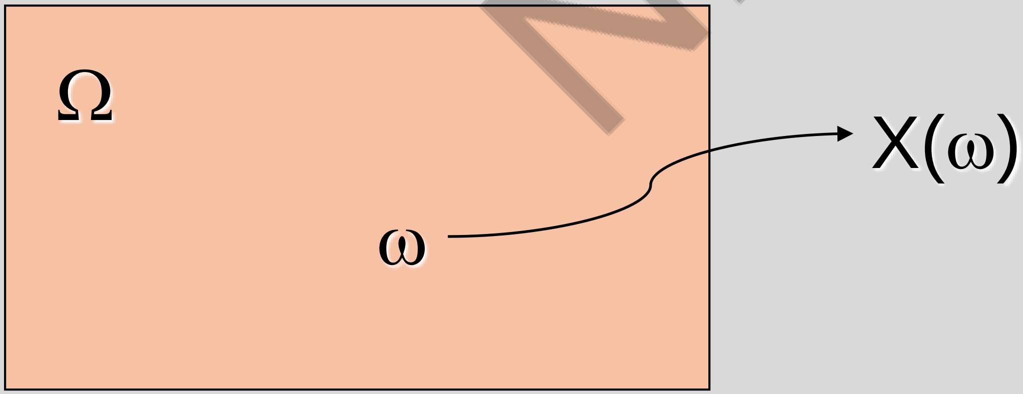
- Intuitively, the probability of an event **a** could be defined as:

$$P(a) = \lim_{n \to \infty} \frac{N(a)}{n}$$

Where N(a) is the number that event a happens in n trials

Random Variable

- A *random variable* is a function defined on the sample space Ω
  - maps the outcome of a random event into real scalar values



Discrete Random Variables

- Random variables (RVs) which may take on only a countable number of distinct values
  - e.g., the sum of the value of two dies

- X is a RV with arity *k* if it can take on exactly one value out of k values,
  - e.g., the possible values that X can take on are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Probability of Discrete RV

- Probability mass function (pmf): $\mathrm{P}\left(\mathrm{X}=x_i\right)$
- Simple facts about pmf
  - $\sum_i \mathrm{P}\left(\mathrm{X}=x_i\right)=1$
  - $\mathrm{P}\left(\mathrm{X}=x_i \cap \mathrm{X}=x_j\right)=0$ if $i \neq j$
  - $\mathrm{P}\left(\mathrm{X}=x_i \cup \mathrm{X}=x_j\right)=\mathrm{P}\left(\mathrm{X}=x_i\right)+\mathrm{P}\left(\mathrm{X}=x_j\right)$if $i \neq j$
  - $\mathrm{P}\left(\mathrm{X}=x_1 \cup \mathrm{X}=x_2 \cup \ldots \cup \mathrm{X}=x_k\right)=1$

Common Distributions

- Uniform $X \sim U[1, \cdots, N]$
  - X takes values 1, 2, …, *N*
  - $\mathrm{P}(\mathrm{X} = i) = 1/N$
  - E.g. picking balls of different colors from a box
- Binomial $X \sim Bin(n, p)$
  - X takes values 0, 1, …, *n*
  - $\mathrm{P}(\mathrm{X} = i) = \binom{n}{i} p^i (1 - p)^{n-i}$
  - E.g. coin flips

Joint Distribution

- Given two discrete RVs X and Y, their **joint distribution** is the distribution of X and Y together
  - e.g.
    you and your friend each toss a coin 10 times
    P(You get 5 heads AND you friend get 7 heads)
- $\sum_{x}\sum_{y} \mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)=1$

$$\sum_{i=0}^{50}\sum_{j=0}^{100} \mathrm{P}\left(\text{You get } i \text{ heads AND your friend get } j \text{ heads}\right)=1$$

Conditional Probability

- $\mathrm{P}(\mathrm{X} = x | \mathrm{Y} = y)$ is the probability of $X = x$, given the occurrence of $Y = y$
  - E.g. you get 0 heads, given that your friend gets 3 heads

- $$\mathrm{P}(\mathrm{X} = x | \mathrm{Y} = y) = \frac{\mathrm{P}(\mathrm{X} = x \cap \mathrm{Y} = y)}{\mathrm{P}(\mathrm{Y} = y)}$$

Law of Total Probability

- Given two discrete RVs X and Y, which take values in $\{x_1,\ldots,x_m\}$ and $\{y_1,\ldots,y_n\}$, We have

$$\mathrm{P}\left(\mathrm{X}=x_i\right)=\sum_j \mathrm{P}\left(\mathrm{X}=x_i \cap \mathrm{Y}=y_j\right)$$

$$=\sum_j \mathrm{P}\left(\mathrm{X}=x_i \middle| \mathrm{Y}=y_j\right)\mathrm{P}\left(\mathrm{Y}=y_j\right)$$

Marginalization

Marginal Probability

Joint Probability

$$
\begin{aligned}
\mathrm{P}\left(\mathrm{X} = x_i\right) &= \sum_j \mathrm{P}\left(\mathrm{X} = x_i \cap \mathrm{Y} = y_j\right) \\
&= \sum_j \mathrm{P}\left(\mathrm{X} = x_i \middle| \mathrm{Y} = y_j\right) \mathrm{P}\left(\mathrm{Y} = y_j\right)
\end{aligned}
$$

Conditional Probability

Marginal Probability

Bayes Rule

- X and Y are discrete RVs…

$$\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Y}=y\right)=\frac{\mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)}{\mathrm{P}\left(\mathrm{Y}=y\right)}$$

↓

$$\mathrm{P}\left(\mathrm{X}=x_i \middle| \mathrm{Y}=y_j\right)=\frac{\mathrm{P}\left(\mathrm{Y}=y_j \middle| \mathrm{X}=x_i\right)\mathrm{P}\left(\mathrm{X}=x_i\right)}{\sum_k \mathrm{P}\left(\mathrm{Y}=y_j \middle| \mathrm{X}=x_k\right)\mathrm{P}\left(\mathrm{X}=x_k\right)}$$

Independent RVs

- X and Y are independent means that $X = x$ does not affect the probability of $Y = y$

- Definition: X and Y are independent iff
  - P(XY) = P(X)P(Y)
  - $\mathrm{P}(\mathrm{X} = x \cap \mathrm{Y} = y) = \mathrm{P}(\mathrm{X} = x)\mathrm{P}(\mathrm{Y} = y)$

More on Independence

$$\begin{tikzcd}[column sep=small]
 & \mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y\right)=\mathrm{P}\left(\mathrm{X}=x\right)\mathrm{P}\left(\mathrm{Y}=y\right) \arrow[dl] \arrow[dr] & \\
\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Y}=y\right)=\mathrm{P}\left(\mathrm{X}=x\right) & & \mathrm{P}\left(\mathrm{Y}=y \middle| \mathrm{X}=x\right)=\mathrm{P}\left(\mathrm{Y}=y\right)
\end{tikzcd}$$

- E.g. no matter how many heads you get, your friend will not be affected, and vice versa

Conditionally Independent RVs

- Intuition: X and Y are conditionally independent given Z means that once Z is **known**, the value of X does not add any **additional** information about Y
- Definition: X and Y are conditionally independent given Z iff

$$\mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y \middle| \mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Z}=z\right)\mathrm{P}\left(\mathrm{Y}=y \middle| \mathrm{Z}=z\right)$$

More on Conditional Independence

$$\begin{tikzcd}
 & \mathrm{P}\left(\mathrm{X}=x\cap \mathrm{Y}=y\middle|\mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{X}=x\middle|\mathrm{Z}=z\right)\mathrm{P}\left(\mathrm{Y}=y\middle|\mathrm{Z}=z\right) \arrow[dl] \arrow[ddr] & \\
\mathrm{P}\left(\mathrm{X}=x\middle|\mathrm{Y}=y,\mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{X}=x\middle|\mathrm{Z}=z\right) & & \\
 & & \mathrm{P}\left(\mathrm{Y}=y\middle|\mathrm{X}=x,\mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{Y}=y\middle|\mathrm{Z}=z\right)
\end{tikzcd}$$

Continuous Random Variables

- What if X is continuous?
- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable *x*.

PDF

- Properties of pdf
  - $f\left(x\right) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f\left(x\right) = 1$
  - $f\left(x\right) \leq 1$???
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

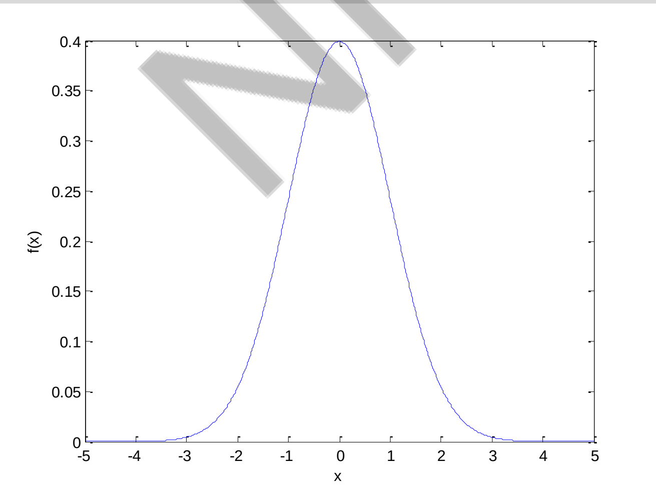
$$\mathrm{P}\left(0 \leq X \leq 1\right) = \int_{0}^{1} f\left(x\right) dx$$

Cumulative Distribution Function

- $F_{\mathrm{X}}(v) = \mathrm{P}(\mathrm{X} \le v)$
- Discrete RVs
  - $F_{\mathrm{X}}(v) = \sum_{v_i} \mathrm{P}(\mathrm{X} = v_i)$
- Continuous RVs
  - $F_{\mathrm{X}}(v) = \int_{-\infty}^{v} f(x)\,dx$
  - $\frac{d}{dx} F_{\mathrm{X}}(x) = f(x)$

Common Distributions

- Normal $X \sim N(\mu, \sigma^2)$
  - $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\dfrac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \mathbb{R}$
  - E.g. the height of the entire population



Multivariate Normal

- Generalization to higher dimensions of the one-dimensional normal

Covariance Matrix

- $$f_{\vec{X}}(x_1,\ldots,x_d)=\frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}(\vec{x}-\mu)^T\,\Sigma^{-1}(\vec{x}-\mu)\right\}$$

Mean

Mean and Variance

- Mean (Expectation): $\mu = E(\mathrm{X})$
  - Discrete RVs: $E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $E(\mathrm{X}) = \int_{-\infty}^{+\infty} xf(x)\,dx$
- Variance: $V(\mathrm{X}) = E(\mathrm{X} - \mu)^2$
  - Discrete RVs: $V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $V(\mathrm{X}) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\,dx$

Mean Estimation from Samples

- Given a set of N samples from a distribution, we can estimate the mean of the distribution by:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance Estimation from Samples

- Given a set of N samples from a distribution, we can estimate the variance of the distribution by:

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2$$

Thank You

Foundations of Machine Learning

Module 4:

Part B: Bayesian Learning

Sudeshna Sarkar

IIT Kharagpur

Probability for Learning

- Probability for classification and modeling concepts.
- Bayesian probability
  - – Notion of probability interpreted as partial belief
- Bayesian Estimation
  - – Calculate the validity of a proposition
    - Based on prior estimate of its probability
    - and New relevant evidence

Bayes Theorem

- **Goal:** To determine the most probable hypothesis, given the data *D* plus any initial knowledge about the prior probabilities of the various hypotheses in *H*.

Bayes Theorem

Bayes Rule: $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- P(h) = prior probability of hypothesis h
- P(D) = prior probability of training data D
- P(h|D) = probability of h given D (posterior density)
- P(D|h) = probability of D given h (likelihood of D given h)

An 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.

$$P(cancer) = .008, P(\neg cancer) = .992$$

$$P(+ | cancer) = .98, P(- | cancer) = .02$$

$$P(+ | \neg cancer) = .03, P(- | \neg cancer) = .97$$

$$P(cancer | +) = \frac{P(+ | cancer)P(cancer)}{P(+)}$$

$$P(\neg cancer | +) = \frac{P(+ | \neg cancer)P(\neg cancer)}{P(+)}$$

Maximum A Posteriori (MAP) Hypothesis

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

The Goal of Bayesian Learning: the most probable hypothesis given the training data (Maximum A Posteriori hypothesis)

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

$$= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D \mid h)P(h)$$

Maximum Likelihood (ML) Hypothesis

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

$$= \arg\max_{h \in H} \frac{P(D \mid h) P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D \mid h) P(h)$$

- If every hypothesis in *H* is equally probable a priori, we only need to consider the likelihood of the data *D* given *h*, ***P(D|h).*** Then, h_{MAP} becomes the ***Maximum Likelihood***,

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

MAP Learner

For each hypothesis h in H, calculate the posterior probability

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

Output the hypothesis h_{MAP} with the highest posterior probability

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

Comments:

Computational intensive

Providing a standard for judging the performance of learning algorithms

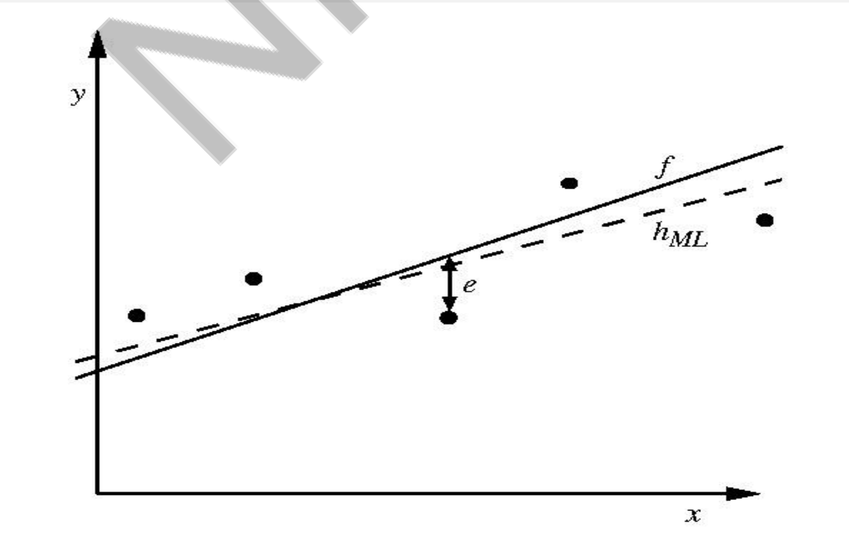
Choosing P(h) and P(D|h) reflects our prior knowledge about the learning task

Maximum likelihood and least-squared error

- Learn a Real-Valued Function:
  - Consider any real-valued target function f.
  - Training examples (x_i,d_i) are assumed to have Normally distributed noise e_i with zero mean and variance σ^2, added to the true target value f(x_i),

d_i satisfies $N(f(x_i), \sigma^2)$

Assume that e_i is drawn independently for each x_i .



Compute ML Hypo

$$h_{ML} = \arg\max_{h \in H} p(D|h)$$

$$= \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{d_i - h(x_i)}{\sigma})^2}$$

$$= \arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}(\frac{d_i - h(x_i)}{\sigma})^2$$

$$= \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

Bayes Optimal Classifier

Question: Given new instance x, what is its most probable classification?

- $h_{MAP}(x)$ is not the most probable classification!

Example: Let P(h1|D) = .4, P(h2|D) = .3, P(h3 |D) =.3

Given new data x, we have h1(x)=+, h2(x) = -, h3(x) = -

What is the most probable classification of x ?

Bayes optimal classification:

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

where *V* is the set of all the values a classification can take and v_j is one possible such classification.

Example:

P(h1| D) =.4,	P(-|h1)=0,	P(+|h1)=1
P(h2|D) =.3,	P(-|h2)=1,	P(+|h2)=0
P(h3|D)=.3,	P(-|h3)=1,	P(+|h3)=0

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

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

Why "Optimal"?

- Optimal in the sense that no other classifier using the same *H* and prior knowledge can outperform it on average

Gibbs Algorithm

- Bayes optimal classifier is quite computationally expensive, if *H* contains a large number of hypotheses.
- An alternative, less optimal classifier Gibbs algorithm, defined as follows:
  1. Choose a hypothesis randomly according to $P(h|D)$, where *D* is the posterior probability distribution over *H*.
  2. Use it to classify new instance

Error for Gibbs Algorithm

- Surprising fact: Assume the expected value is taken over target concepts drawn at random, according to the prior probability distribution assumed by the learner, then (Haussler *et al.* 1994)

$$E_f[error_{X,f}GibbsClassifier] \le 2E_f[error_{X,f}BayesOPtimal],$$

where f denotes a target function, X denotes the instance space.

Thank You

Foundations of Machine Learning

Module 4:

Part C: Naïve Bayes

Sudeshna Sarkar

IIT Kharagpur

Bayes Theorem

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

Naïve Bayes

- Bayes classification

$$P(Y | \mathbf{X}) \propto P(\mathbf{X} | Y)P(Y) = P(X_1,\cdots, X_n | Y)P(Y)$$

Difficulty: learning the joint probability $P(X_1,\cdots,X_n | C)$

- Naïve Bayes classification

Assume all input features are conditionally independent!

$$\begin{aligned} P(X_1, X_2,\cdots, X_n | Y) &= P(\underline{X_1 | X_2,\cdots, X_n, Y})P(X_2,\cdots, X_n | Y) \\ &= P(\underline{X_1 | Y})P(\underline{X_2,\cdots, X_n | Y}) \\ &= P(X_1 | Y)\underline{P(X_2 | Y)\cdots P(X_n | Y)} \end{aligned}$$

Naïve Bayes

Bayes rule:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \dots X_n | Y = y_j)}$$

Assuming conditional independence among X_i's:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, classification rule for $X^{new} = < X_1, \dots, X_n >$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k

      estimate $\pi_k \equiv P(Y = y_k)$

      for each* value x_{ij} of each attribute X_i

          estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates (MLE's):

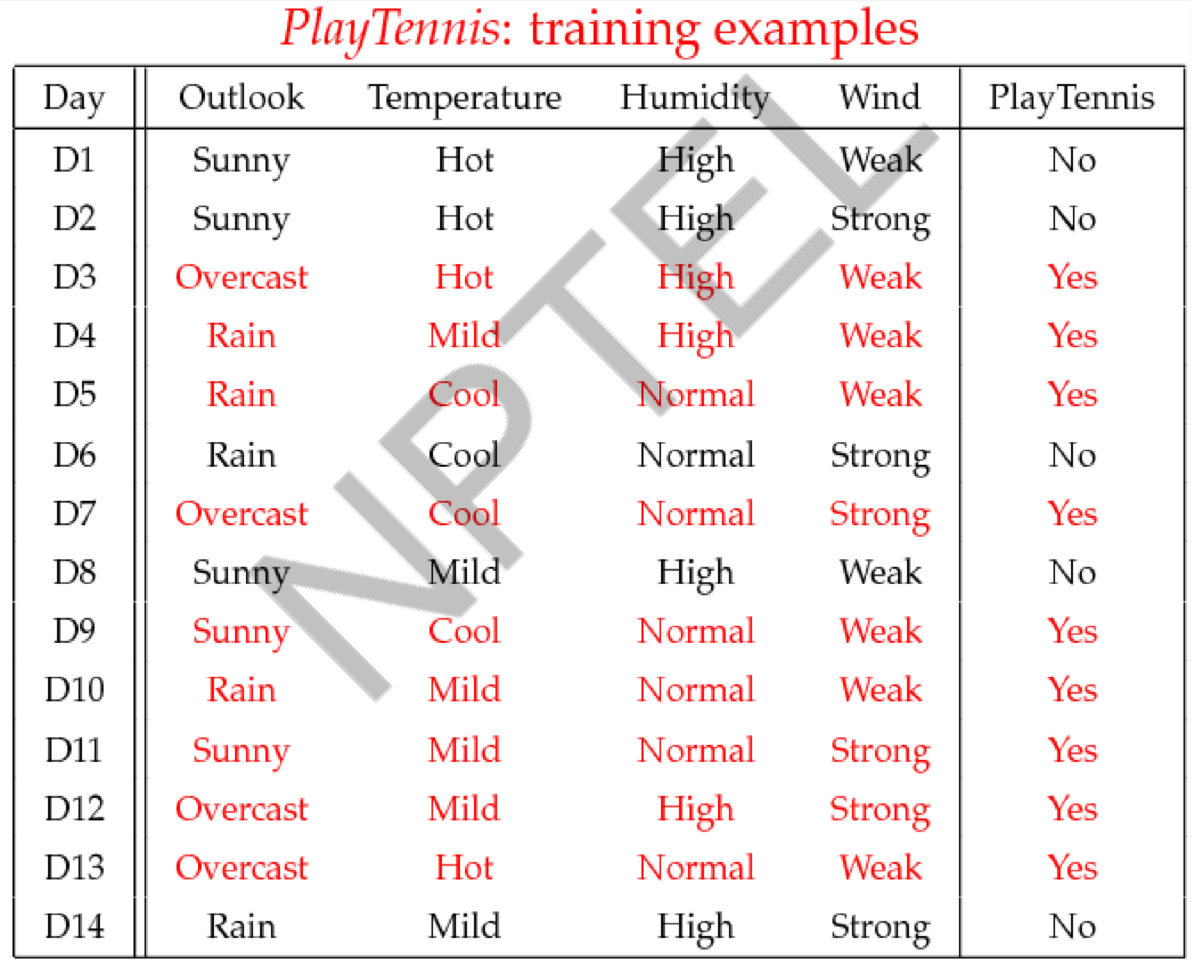
$$\widehat{\pi}_k = \widehat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\widehat{\theta}_{ijk} = \widehat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in set D for which Y=y_k

Example

- Example: Play Tennis



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

Example

Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

P(Play=*Yes*) = 9/14 *P*(Play=*No*) = 5/14

Example

Test Phase

– Given a new instance, predict its label

$\mathbf{x}'$=(Outlook=*Sunny,* Temperature=*Cool,* Humidity=*High,* Wind=*Strong*)

– Look up tables achieved in the learning phrase

P(Outlook=*Sunny*|Play=*Yes*) = 2/9

P(Temperature=*Cool*|Play=*Yes*) = 3/9

P(Huminity=*High*|Play=*Yes*) = 3/9

P(Wind=*Strong*|Play=*Yes*) = 3/9

P(Play=*Yes*) = 9/14

P(Outlook=*Sunny*|Play=*No*) = 3/5

P(Temperature=*Cool*|Play==*No*) = 1/5

P(Huminity=*High*|Play=*No*) = 4/5

P(Wind=*Strong*|Play=*No*) = 3/5

P(Play=*No*) = 5/14

– Decision making with the MAP rule

$P(Yes|\mathbf{x}') \approx [P(Sunny|\text{Yes})P(Cool|\text{Yes})P(High|\text{Yes})P(Strong|\text{Yes})]P(\text{Play}=Yes) = 0.0053$

$P(No|\mathbf{x}') \approx [P(Sunny|\text{No})\, P(Cool|\text{No})P(High|No)P(Strong|No)]P(\text{Play}=No) = 0.0206$

Given the fact $P(Yes|\mathbf{x}') < P(No|\mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

Estimating Parameters: *Y*, X_i discrete-valued

If unlucky, our MLE estimate for $P(X_i \mid Y)$ may be zero.

$$\widehat{\pi}_k = \widehat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\widehat{\theta}_{ijk} = \widehat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates:

$$\widehat{\pi}_k = \widehat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + l}{|D| + lR}$$

Only difference:
"imaginary" examples

$$\widehat{\theta}_{ijk} = \widehat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + l}{\#D\{Y = y_k\} + lM}$$

Naïve Bayes: Assumptions of Conditional Independence

Often the X_i are not really conditionally independent

- We can use Naïve Bayes in many cases anyway
  - often the right classification, even when not the right probability

Gaussian Naïve Bayes (continuous X)

- Algorithm: Continuous-valued Features
  - Conditional probability often modeled with the normal distribution

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \; e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

Sometimes assume variance

- is independent of Y (i.e., σ_i),
- or independent of X_i (i.e., σ_k)
- or both (i.e., σ)

Gaussian Naïve Bayes Algorithm – continuous X_i
(but still discrete Y)

- Train Naïve Bayes (examples)

  for each value y_k

  estimate* $\pi_k \equiv P(Y = y_k)$

  for each attribute X_i estimate

  class conditional mean μ_{ik}, variance σ_{ik}

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

jth training example

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

ith feature

kth class

δ(z)=1 if z true, else 0

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

Naïve Bayes

- Example: Continuous-valued Features
  - Temperature is naturally of continuous value.

Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

No: 27.3, 30.1, 17.4, 29.5, 15.1

- Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$
$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

- **Learning Phase**: output two Gaussian models for P(temp|C)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2\times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2\times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

The independence hypothesis…

- makes computation possible
- yields optimal classifiers when satisfied
- Rarely satisfied in practice, as attributes (variables) are often correlated.
- To overcome this limitation:
  - Bayesian networks combine Bayesian reasoning with causal relationships between attributes

Thank You

Foundations of Machine Learning

Module 4:

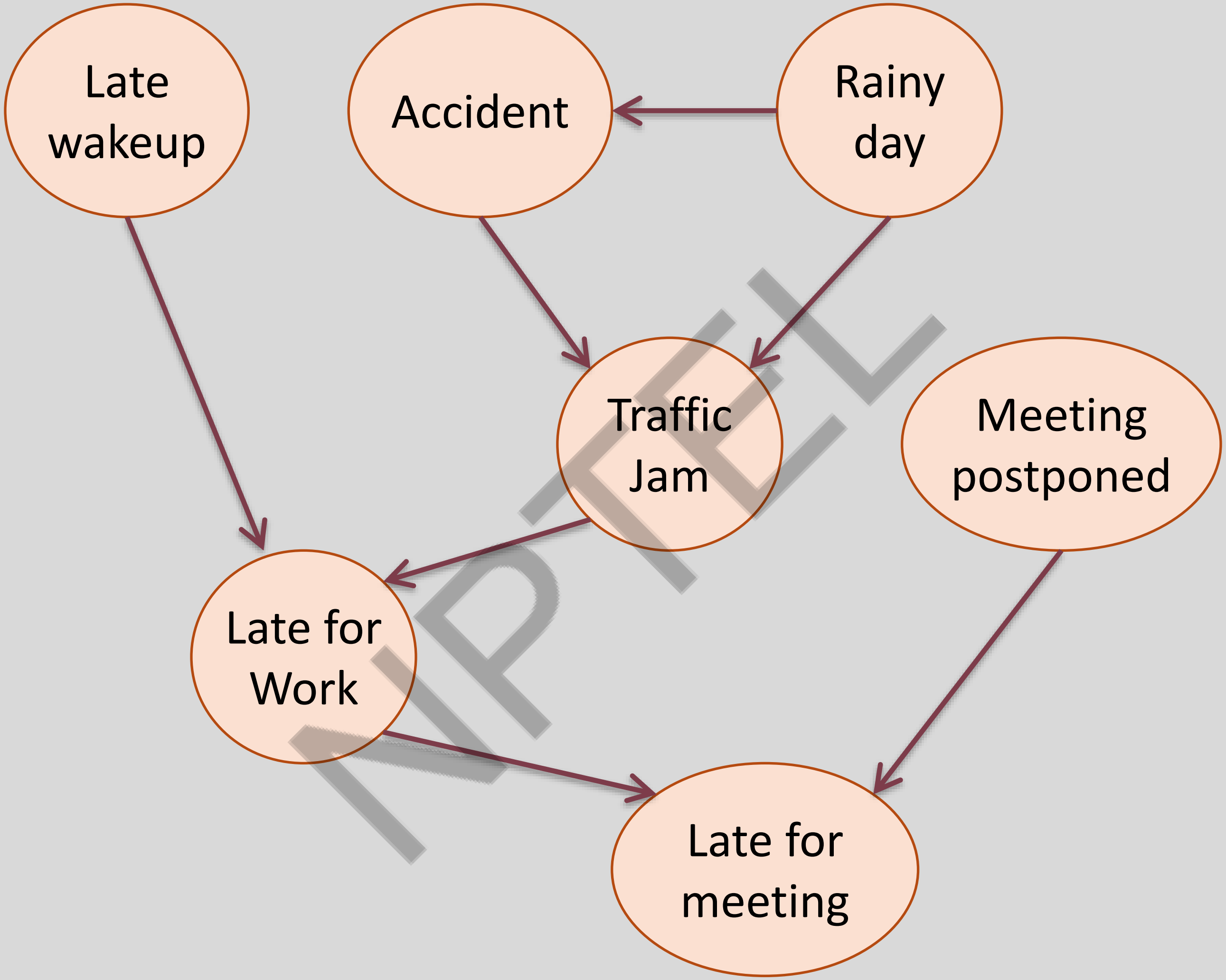
Part D: Bayesian Networks

Sudeshna Sarkar

IIT Kharagpur

Why Bayes Network

- Bayes optimal classifier is too costly to apply
- Naïve Bayes makes overly restrictive assumptions.
  - But all variables are rarely completely independent.
- Bayes network represents conditional independence relations among the features.
- Representation of causal relations makes the representation and inference efficient.



Late wakeup

Accident

Rainy day

Traffic Jam

Meeting postponed

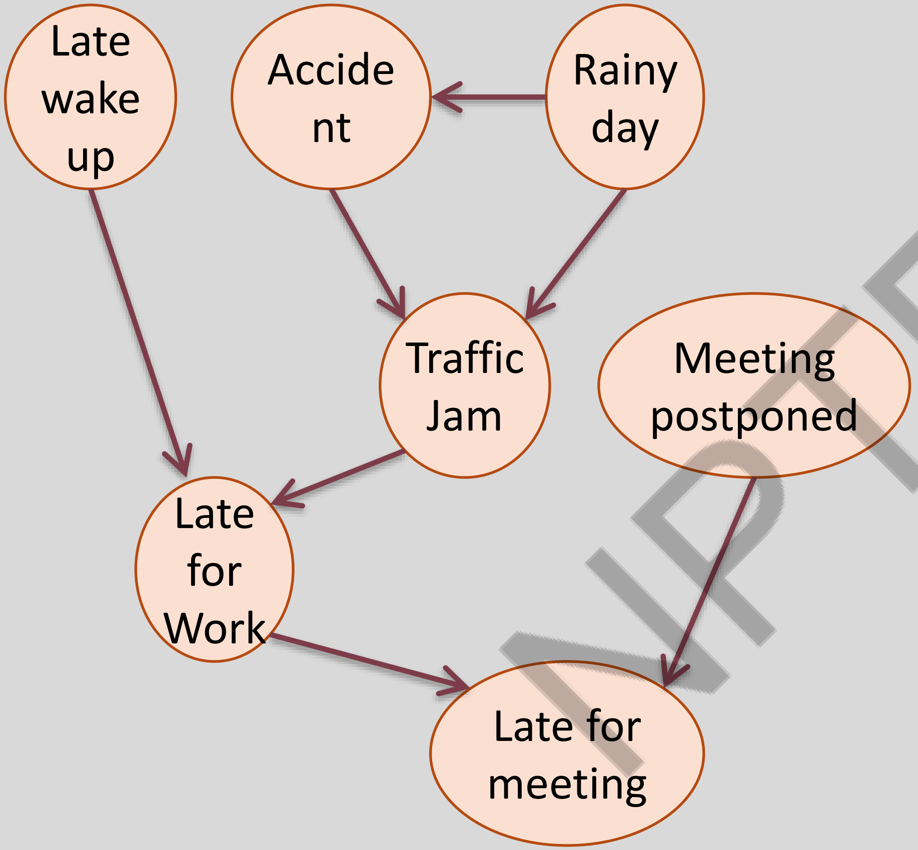
Late for Work

Late for meeting

Bayesian Network

- A graphical model that efficiently encodes the joint probability distribution for a large set of variables
- A Bayesian Network for a set of variables (nodes) X = { X1,…….Xn}
- Arcs represent probabilistic dependence among variables
- Lack of an arc denotes a conditional independence
- The network structure S is a directed acyclic graph
- A set P of local probability distributions at each node (Conditional Probability Table)

Representation in Bayesian Belief Networks



Conditional probability table associated with each node specifies the conditional distribution for the variable given its immediate parents in
the graph

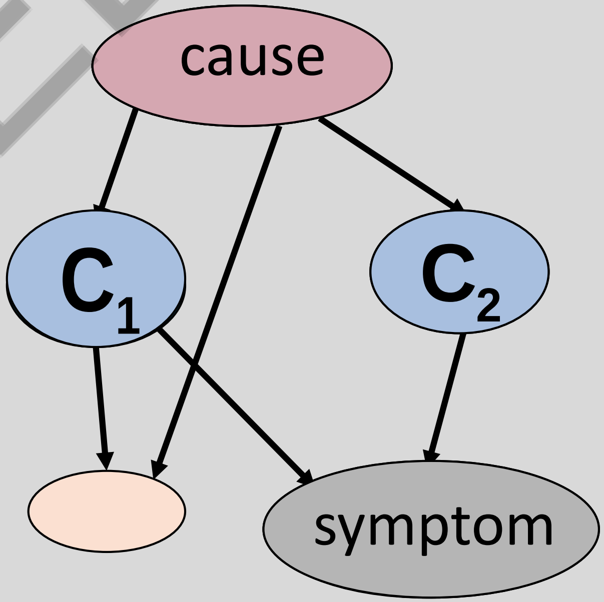
Each node is asserted to be conditionally independent of its non-descendants, given its immediate parents

Inference in Bayesian Networks

- Computes posterior probabilities given evidence about some nodes
- Exploits probabilistic independence for efficient computation.
- Unfortunately, exact inference of probabilities in general for an arbitrary Bayesian Network is known to be NP-hard.
- In theory, approximate techniques (such as Monte Carlo Methods) can also be NP-hard, though in practice, many such methods were shown to be useful.
- Efficient algorithms leverage the structure of the graph

Applications of Bayesian Networks

- Diagnosis: P(cause|symptom)=?
- Prediction: P(symptom|cause)=?
- Classification: P(class|data)
- Decision-making (given a cost function)



Bayesian Networks

- Structure of the graph ⇔ Conditional independence relations

In general,

$$p(X_1, X_2, \ldots X_N) = \Pi\, p(X_i \mid \text{parents}(X_i)\,)$$

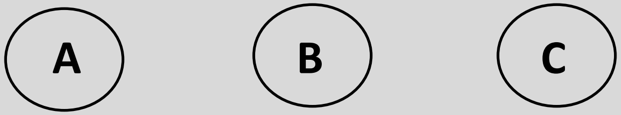
The full joint distribution

The graph-structured approximation

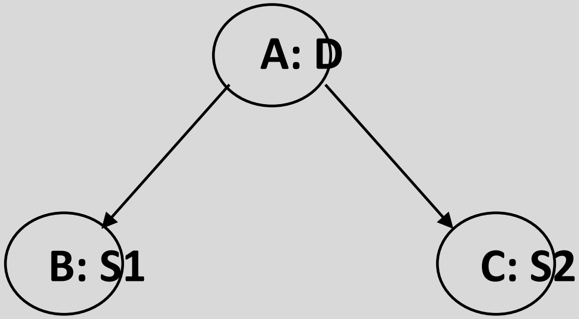
- Requires that graph is acyclic (no directed cycles)

- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)

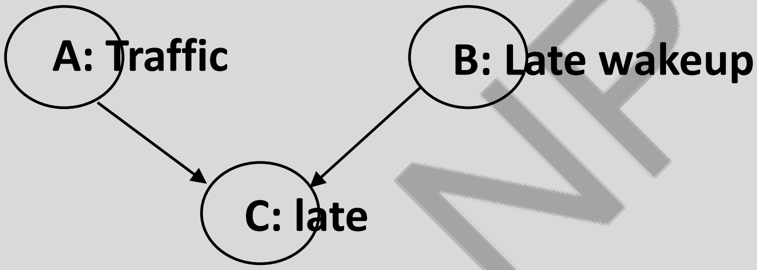
Examples



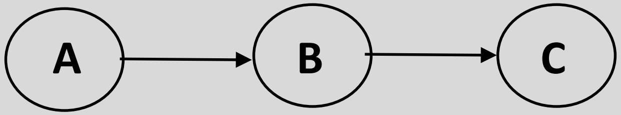
Marginal Independence:
p(A,B,C) = p(A) p(B) p(C)



Conditionally independent effects:
p(A,B,C) = p(B|A)p(C|A)p(A)
B and C are conditionally independent
Given A

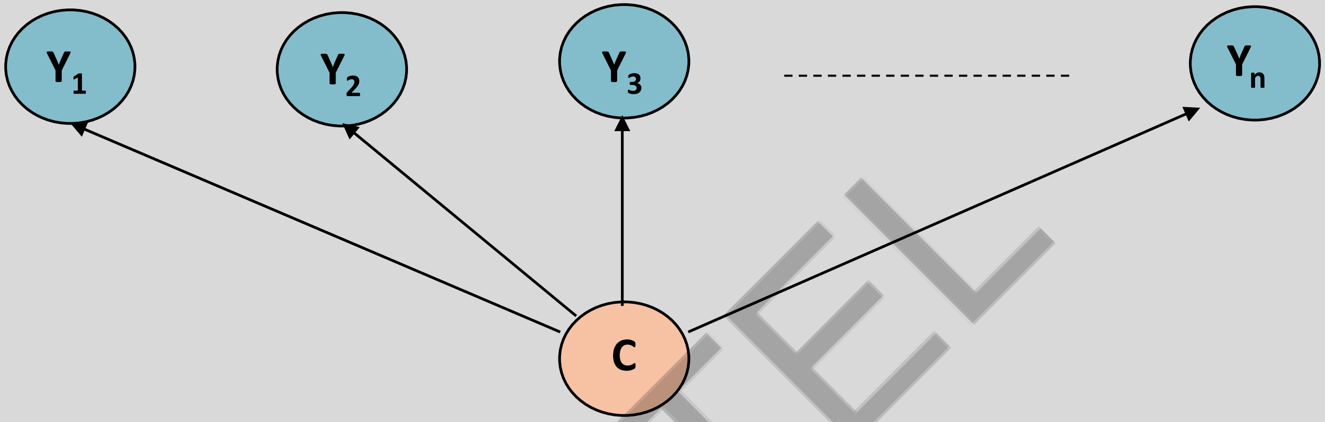


Independent Causes:
p(A,B,C) = p(C|A,B)p(A)p(B)
“Explaining away”

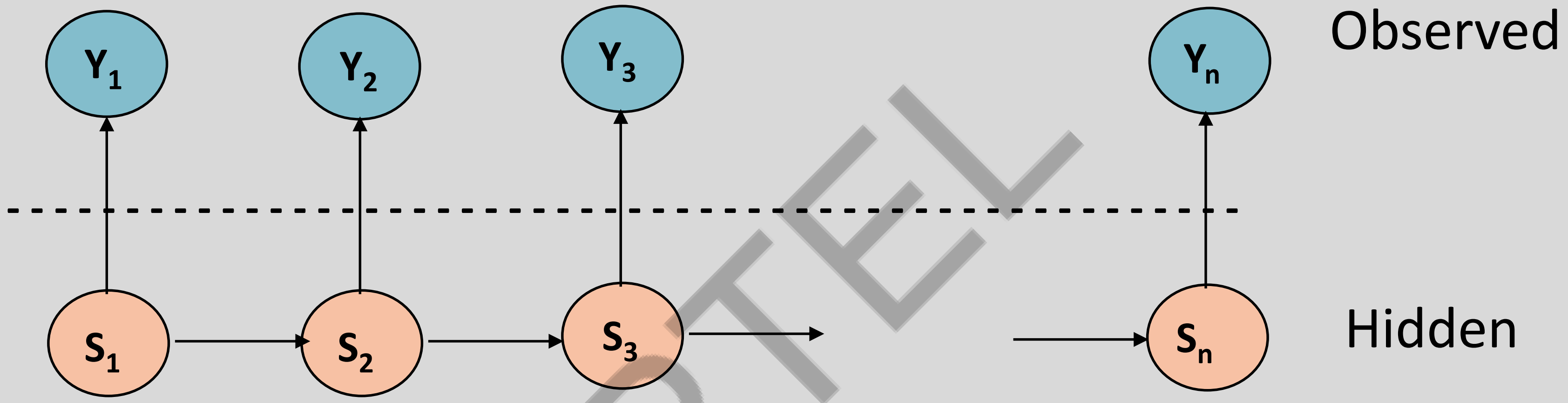


Markov dependence:
p(A,B,C) = p(C|B) p(B|A)p(A)

Naïve Bayes Model



Hidden Markov Model (HMM)



Assumptions:

1. hidden state sequence is Markov
2. observation Y_t is conditionally independent of all other variables given S_t

Widely used in sequence learning eg, speech recognition, POS tagging

Inference is linear in n

Learning Bayesian Belief Networks

1. The network structure is given in advance and all the variables are fully observable in the training examples.

– estimate the conditional probabilities.

2. The network structure is given in advance but only some of the variables are observable in the training data.

– Similar to learning the weights for the hidden units of a Neural Net: Gradient Ascent Procedure

3. The network structure is not known in advance.

– Use a heuristic search or constraint-based technique to search through potential structures.

Thank You