WPI

Artificial Intelligence
CS 534

Week 10

Probabilistic Reasoning

Important dates (tentative schedule)

- Sep. 3: Deadline to report a “NO group” membership
- Sep. 7: Deadline to form groups and select the title/topic of project. Report to me
- Sep. 17: Assignment 1 is due
- Sep. 17: Submit project proposal. Report to me
- Sep. 24: Assignment 2 is due
- Sep. 24: Proposals are returned to the teams
- Oct. 8: Midterm 1
- Oct. 16: Assignment 3 is due
- Nov. 15: Assignment 4 is due
- Nov. 26: Midterm 2
- Dec. 3: Assignment 5 is due
- Dec 3,10: Project presentations
- Dec 4: Beta version of course project paper is due. Report to me
- Dec 13: Final version of course project paper is due. Report to me

Project presentation notes

- Two presentation week slots: Dec 3 and Dec 10
- If a specific slot is need, email Suhas the requested date
- The slots are on the first come first serve basis
- The rest of the teams will be assigned randomly
- We will roughly divide them into two groups: 8 teams first day and 7 teams second day
- Each team is given 20 mins including 2 mins for Q&A (2-3 questions per team)
- The timing will be enforced!
- Each person asking a non-trivial question will be rewarded with 5 points

Project presentation notes

- Presentation should be at most ~12-15 slides and should cover:
  - Problem introduction and motivation: ~4-5 slides
  - Methodology: ~3-4 slides
  - Results: ~3-4
  - Conclusions/Dscussion: ~2 slides

Preliminaries

- Binary events: coin toss (H/T); happens or doesn't happen, *etc*

Preliminaries

- Binary events: coin toss (H/T); happens or doesn't happen, etc.

- Probability: $P(A) = p, \quad P(\neg A) = 1 - p$
  - Example: $\mathrm{P}(x_1{=}\mathrm{H}) = 0.5$ (unbiased coin)

Preliminaries

- Binary events: coin toss (H/T); happens or doesn't happen, etc.

- Probability: $P(A) = p, \quad P(\neg A) = 1 - p$
  - Example: $\mathrm{P}(x_1{=}\mathrm{H}) = 0.5$ (unbiased coin)

- Independence: $X \perp Y : P(X)P(Y) = P(X,Y)$
  - $\mathrm{P}(X,Y)$ is called *joint probability*, $\mathrm{P}(X)$ and $\mathrm{P}(Y)$ are called *marginals*

- Dependence: $P(X \mid Y) = \dfrac{P(X,Y)}{P(Y)}$

Preliminaries

- Binary events: coin toss (H/T); happens or doesn't happen, etc.

- Probability: $P(A) = p, \quad P(\neg A) = 1 - p$
  - Example: $\mathrm{P}(x_1\text{=H}) = 0.5$ (unbiased coin)

- Independence: $X \perp Y : P(X)P(Y) = P(X,Y)$
  - $\mathrm{P}(X,Y)$ is called *joint probability*, $\mathrm{P}(X)$ and $\mathrm{P}(Y)$ are called *marginals*

- Dependence: $P(X \mid Y) = \dfrac{P(X,Y)}{P(Y)}$
  - Example: Assume that while $\mathrm{P}(x_1\text{=H}) = 0.5$ (and $\mathrm{P}(x_1\text{=T}) = 0.5$), we have $\mathrm{P}(x_2\text{=H} \mid x_1\text{=H}) = 0.9$ and $\mathrm{P}(x_2\text{=T} \mid x_1\text{=T}) = 0.8$

    Then: $\mathrm{P}(x_2\text{=H}) = \mathrm{P}(x_2\text{=H} \mid x_1\text{=H})\,\mathrm{P}(x_1\text{=H}) + \mathrm{P}(x_2\text{=H} \mid x_1\text{=T})\,\mathrm{P}(x_1\text{=T}) =$
    $= 0.5*0.9 + (1 - \mathrm{P}(x_2\text{=T} \mid x_1\text{=T}))\,\mathrm{P}(x_1\text{=T}) = 0.45 + 0.2*0.5 = 0.55$

Preliminaries

- Total probability:

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

- Negation:

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

but NOT:

$$P(X \mid \neg Y) \neq 1 - P(X \mid Y)$$

Preliminaries

- Total probability:

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

- Negation:

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

but NOT:

$$P(X \mid \neg Y) \neq 1 - P(X \mid Y)$$

Example: Day can be sunny (S) and rainy (R) :

$$P(D_1{=}S) = 0.9 \Rightarrow P(D_1{=}R) = 0.1$$

now assume that

$$P(D_2{=}S \mid D_1{=}S) = 0.8 \Rightarrow P(D_2{=}R \mid D_1{=}S) = 0.2$$
$$P(D_2{=}S \mid D_1{=}R) = 0.6 \Rightarrow P(D_2{=}R \mid D_1{=}R) = 0.4$$

Then: $P(D_2{=}S) = 0.8*0.9 + 0.6*0.1 = 0.78$

Intuition about conditional probability

- A cancer example: P(Cancer|Test) = P(A|B)
  - B is the test (we don't care) or it's evidence
  - A is what we care about!
  - Diagnostic reasoning in P(A|B): from evidence to its causes
  - Causal reasoning in P(B|A): hypothetically, if we knew the cause, what is the probability of observing it in the test?
  - We typically know probabilities P(B|A) (by statistics on the known cases and controls)
  - How to get P(A|B) from it???

Intuition about conditional probability

- A cancer example: P(Cancer|Test) = P(A|B)
  - B is the test (we don't care) or it's evidence
  - A is what we care about!
  - Diagnostic reasoning in P(A|B): from evidence to its causes
  - Causal reasoning in P(B|A): hypothetically, if we knew the cause, what is the probability of observing it in the test?
  - We typically know probabilities P(B|A) (by statistics on the known cases and controls)
  - How to get P(A|B) from it???

- Bayes Rule:

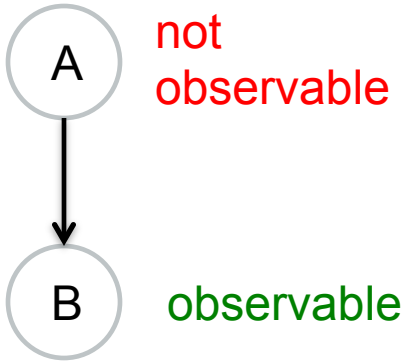
likelihood *prior*

$$\underset{\text{posterior}}{P(A \mid B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

marginal likelihood

Bayes rule graphically

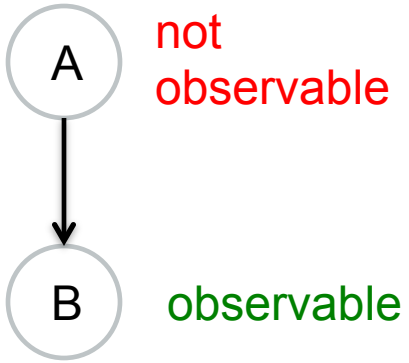
- We have a primitive Bayes Net which is composed of 2 variables



- We know prior $P(A)$
- We know: $P(B \mid A)$ and $P(B \mid \neg A)$
- We don't know: $P(A \mid B)$ and $P(A \mid \neg B)$
-

Bayes rule graphically

- We have a primitive Bayes Net which is composed of 2 variables



- We know prior $P(A)$
- We know: $P(B \mid A)$ and $P(B \mid \neg A)$
- We don't know: $P(A \mid B)$ and $P(A \mid \neg B)$
- Note, that in total we need 3 numerical parameters:

$$P(A), P(B \mid A), \text{ and } P(B \mid \neg A)$$

Example: Cancer diagnostics

- $P(C) = 0.01 \Rightarrow P(\neg C) = 0.99$
- Let's assume, we have a test (could be positive or negative):

$$P(+ \mid C) = 0.9 \Rightarrow P(- \mid C) = 0.1$$
$$P(+ \mid \neg C) = 0.2 \Rightarrow P(- \mid \neg C) = 0.8$$

Example: Cancer diagnostics

- $P(C) = 0.01 \Rightarrow P(\neg C) = 0.99$
- Let's assume, we have a test (could be positive or negative):

$$P(+ \mid C) = 0.9 \Rightarrow P(- \mid C) = 0.1$$
$$P(+ \mid \neg C) = 0.2 \Rightarrow P(- \mid \neg C) = 0.8$$

- First, calculate joint probabilities:

$$P(+, C) = P(+ \mid C)P(C) = 0.9 * 0.01 = 0.009$$
$$P(-, C) = P(- \mid C)P(C) = 0.1 * 0.01 = 0.001$$
$$P(+, \neg C) = P(+ \mid \neg C)P(\neg C) = 0.2 * 0.99 = 0.198$$
$$P(-, \neg C) = P(- \mid \neg C)P(\neg C) = 0.8 * 0.99 = 0.792$$

-

-

Example: Cancer diagnostics

- $P(C) = 0.01 \Rightarrow P(\neg C) = 0.99$
- Let's assume, we have a test (could be positive or negative):

$$P(+ \mid C) = 0.9 \Rightarrow P(- \mid C) = 0.1$$
$$P(+ \mid \neg C) = 0.2 \Rightarrow P(- \mid \neg C) = 0.8$$

- First, calculate joint probabilities:

$$P(+, C) = P(+ \mid C)P(C) = 0.9 * 0.01 = 0.009$$
$$P(-, C) = P(- \mid C)P(C) = 0.1 * 0.01 = 0.001$$
$$P(+, \neg C) = P(+ \mid \neg C)P(\neg C) = 0.2 * 0.99 = 0.198$$
$$P(-, \neg C) = P(- \mid \neg C)P(\neg C) = 0.8 * 0.99 = 0.792$$

- Second, calculate P(+):

$$P(+) = P(+ \mid C)P(C) + P(+ \mid \neg C)P(\neg C) = 0.009 + 0.198 = 0.207$$

-

Example: Cancer diagnostics

- $P(C) = 0.01 \Rightarrow P(\neg C) = 0.99$
- Let's assume, we have a test (could be positive or negative):

$$P(+ \mid C) = 0.9 \Rightarrow P(- \mid C) = 0.1$$
$$P(+ \mid \neg C) = 0.2 \Rightarrow P(- \mid \neg C) = 0.8$$

- First, calculate joint probabilities:

$$P(+, C) = P(+ \mid C)P(C) = 0.9 * 0.01 = 0.009$$
$$P(-, C) = P(- \mid C)P(C) = 0.1 * 0.01 = 0.001$$
$$P(+, \neg C) = P(+ \mid \neg C)P(\neg C) = 0.2 * 0.99 = 0.198$$
$$P(-, \neg C) = P(- \mid \neg C)P(\neg C) = 0.8 * 0.99 = 0.792$$

- Second, calculate P(+):

$$P(+) = P(+ \mid C)P(C) + P(+ \mid \neg C)P(\neg C) = 0.009 + 0.198 = 0.207$$

- Last, let's calculate the diagnostic reasoning:

$$P(C \mid +) = P(+, C) / P(+) = 0.009 / 0.207 = 0.4348$$

Bayesian trick: Idea

- Recall Bayes rule:

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

- P(B|A)P(A) is easy to calculate

- P(B) is hard to calculate (but it does not depend on P(A)!)

- Write down Bayes rule for the $P(\neg A)$:

$$P(\neg A \mid B) = \frac{P(B \mid \neg A)P(\neg A)}{P(B)}$$

- Note that it has the same denominator

- We also know that: $P(A \mid B) + P(\neg A \mid B) = 1$

Bayesian trick

- Idea: Let's ignore the normalizer term P(B)

$$P'(A \mid B) = P(B \mid A)P(A) \quad P(A \mid B) = \mu P'(A \mid B)$$
$$P'(\neg A \mid B) = P(B \mid \neg A)P(\neg A) \quad P(\neg A \mid B) = \mu P'(\neg A \mid B)$$

- μ is our normalization term

- Note that $\mu = \dfrac{1}{P'(A \mid B) + P'(\neg A \mid B)}$

A useful expansion of Bayes Rule

- Given:

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

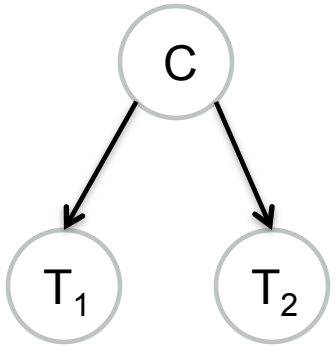
- Prove:

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

Example: two-tests cancer

- Given: $P(C) = 0.01 \Rightarrow P(\neg C) = 0.99$

  $P(+ \mid C) = 0.9, \quad P(- \mid \neg C) = 0.8$



- Find the probability of cancer, given that both (independent) tests came positive: $P(C \mid T_1{=}+, T_2{=}+)$

Example: two-tests cancer (contd)

	$prior$	$P(+\mid\bullet)$	$P(+\mid\bullet)$	$p' = prior * P(+\mid\bullet) * P(+\mid\bullet)$	$p(\bullet\mid{+}{+})$
C	0.01	0.9	0.9	0.0081	$0.0081 / 0.0477 = 0.1698$
$\neg C$	0.99	0.2	0.2	0.0396	$0.0356 / 0.0477 = 0.8302$

Hmm… What about $C(+,-)$?

	$prior$	$P(+\mid\bullet)$	$P(-\mid\bullet)$	$p' = prior * P(+\mid\bullet) * P(+\mid\bullet)$	$p(\bullet\mid{+}{+})$
C	0.01	0.9	0.1	0.0009	0.0056
$\neg C$	0.99	0.2	0.8	0.1584	0.9943

And what about $C(-,-)$?

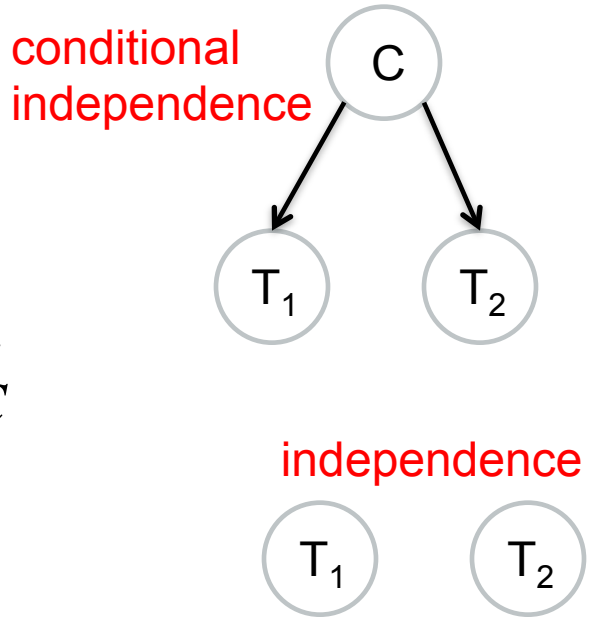
	$prior$	$P(-\mid\bullet)$	$P(-\mid\bullet)$	$p' = prior * P(-\mid\bullet) * P(-\mid\bullet)$	$p(\bullet\mid{-}{-})$
C	0.01	0.1	0.1	0.0001	0.0002
$\neg C$	0.99	0.8	0.8	0.6336	0.9998

Important note

- Here we assumed that tests T_1 and T_2 are conditionally independent ($T_1 \perp T_2 \mid C$):

$$P(T_2 \mid C, T_1) = P(T_2 \mid C)$$

- Conditional independence is a key property in Bayesian Networks

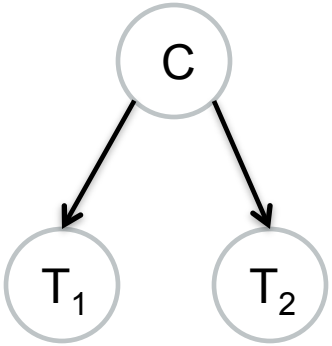


- Note:
  1. From $T_1 \perp T_2 \mid C$ does not follow $T_1 \perp T_2$
  2. From $T_1 \perp T_2$ does not follow $T_1 \perp T_2 \mid C$

Example

- Given: $P(C) = 0.01 \Rightarrow P(\neg C) = 0.99$

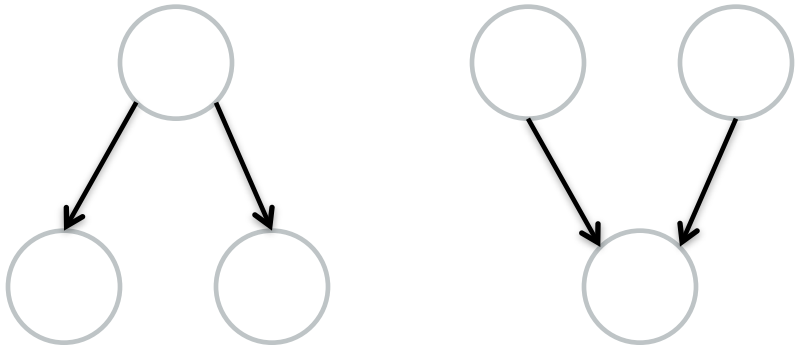
  $P(+ \mid C) = 0.9, \quad P(- \mid \neg C) = 0.8$



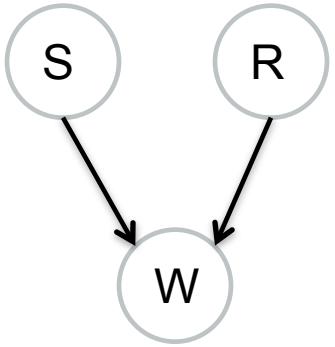
- Find the conditional probability of the second test to be positive given the first came positive

$$P(+_2 \mid +_1) = P(+_2 \mid +_1, C)P(C \mid +_1) + P(+_2 \mid +_1, \neg C)P(\neg C \mid +_1) =$$
$$= P(+_2 \mid C)P(C \mid +_1) + P(+_2 \mid \neg C)P(\neg C \mid +_1)$$

There are different types of dependences



- Example: Wet bounce house (W) dependent on sprinkler (S) or rainy weather (R)



Naïve Bayes: An example

- Let's consider an email SPAM filter (SPAM vs HAM)

- Approach: Bag of Words
  - Uses count of each word from the dictionary
  - Example: sentence "Win $10,000,000, subscribe to win!"

win	$10,000,000	subscribe	to
2	1	1	1

- Example: Spam vs. Ham

SPAM	HAM
OFFER IS SECRET	PLAY SPORTS TODAY
CLICK SECRET LINK	SECRET SPORTS EVENT
SECRET SPORTS LINK	SPORTS
	SECRET IS SECRET
	GO MIZZOU

P(HAM)=5/8

P(SPAM)=3/8

Naïve Bayes: An example (contd.)

- Assume: each of the items is drown independently with the same distribution

- Let's have the following 8 messages observed: SSSHHHHH

- What we are interested in is what's our *prior* probability of SPAM that maximizes the likelihood of the data:
  P(S)=p_0 => P(y_i)= p_0 if y_i=S and P(y_i)= (1−p_0) otherwise

- If we think of our events as 11100000, then we can define:

$$P(y_i) = p_0^{y_i}(1-p_0)^{1-y_i}$$

Naïve Bayes: An example (contd.)

- And for the entire data set:

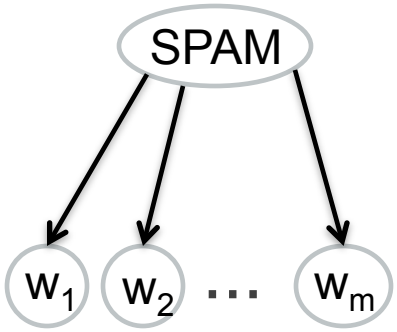
$$P(data) = \prod_{i=1}^{n} {p_0}^{y_i} = {p_0}^3 (1 - p_0)^5$$

- Maximum of ${p_0}^3(1\text{-}p_0)^5$ correspond to the maximum of log; obtain it by taking a derivative and making it zero:

$$(\log(P(data)))' = 3/p_0 + 5/(1\text{-}p_0) = 0$$

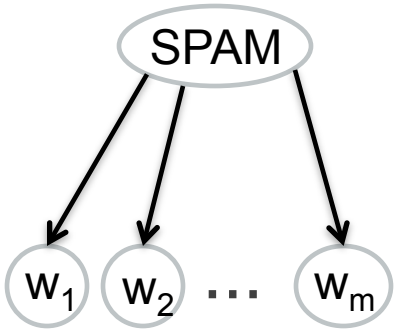
- Thus $p_0 = 3/8$, which is nothing but empirical counting!

Naïve Bayes: Basic ideas



Words in dictionary

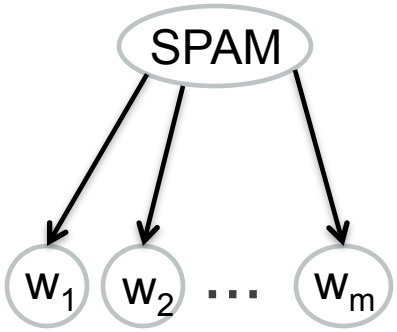
Naïve Bayes: Basic ideas



Words in dictionary

- Here, $P(w_i|\ SPAM)$ and $P(w_i|\ HAM)$ are found from data
- Let's say we have a dictionary of 12 words, then how many parameters we need?
- Answer: 23. Indeed $p(SPAM)$ is one parameter $P(w_i|\ SPAM)$ takes 12-1=11 and $P(w_i|\ HAM)$ takes another 11 parameters

Naïve Bayes: Basic ideas



Words in dictionary

- Here, $P(w_i|\,SPAM)$ and $P(w_i|\,HAM)$ are found from data
- Let's say we have a dictionary of 12 words, then how many parameters we need?
- Answer: 23. Indeed $p(SPAM)$ is one parameter $P(w_i|\,SPAM)$ takes 12-1=11 and $P(w_i|\,HAM)$ takes another 11 parameters
- Given a message *M*, we calculate the probability of it being a spam:

$$P(SPAM \mid M) = \frac{P(M \mid SPAM)P(SPAM)}{P(M)}$$

Naïve Bayes: A general approach

- Let $\boldsymbol{x} \in \mathbb{R}^{\ell}$ and the goal is to estimate $P(\boldsymbol{x} \mid \omega_i)$ $i = 1, 2, \ldots, M$. For a "good" estimate of the pdf one would need, say, N^{ℓ} points.

- Assume $x_1, x_2, \ldots, x_{\ell}$ mutually independent. Then:

$$p(\boldsymbol{x} \mid \omega_i) = \prod_{j=1}^{\ell} p(x_j \mid \omega_i)$$

- In this case, one would require, roughly, N points for each pdf. Thus, a number of points of the order $N \cdot \ell$ would suffice

- It turns out that the Naïve – Bayes classifier works reasonably well even in cases that violate the independence assumption

Naïve Bayes

- Probability is then estimated as

$$P(\omega_i \mid x_j) \propto P(\boldsymbol{x} \mid \omega_i) P(\omega_i) = P(\omega_i) \prod_{j=1}^{\ell} P(x_i \mid \omega_j)$$

- Maximize it by the same log likelihood maximization

- Priors are estimated based on frequencies of occurrences

- Conditional probabilities are estimated from data

Laplasian smoothing

- What happened when we have a word not from the dictionary?

$$P\left(\omega_i \mid x_j\right) \propto P\left(\boldsymbol{x} \mid \omega_i\right) P\left(\omega_i\right) = P\left(\omega_i\right) \prod_{j=1}^{\ell} P\left(x_i \mid \omega_j\right)$$

- The probability is zero according to maximum likelihood, since one word was never observed in the class

- Are we overfitting?

- Yes, since a single word can alter probability so drastically

Approach

- Classical maximum likelihood:

$$p(x) = \frac{count(x)}{N}$$

Approach

- Classical maximum likelihood:

$$p(x) = \frac{count(x)}{N}$$

- Laplas smoothing

$$p(x) = \frac{count(x) + k}{N + kN_w}$$

  - *count*(x) – number of values
  - *k* – a tunable parameter
  - N_W – number of classes
  - *N* – total number of ocurrences

Approach

- Classical maximum likelihood:

$$p(x) = \frac{count(x)}{N}$$

- Laplas smoothing

$$p(x) = \frac{count(x) + k}{N + kN_w}$$

  - *count*(x) – number of values
  - *k* – a tunable parameter
  - N_W – number of classes
  - *N* – total number of ocurrences

- Example: If $k=1$, then for 10 messages and 6 spam messages we have:

$$p(x) = \frac{6+1}{10+1*2} = 0.583 \text{ (compare with 0.6 in ML)}$$

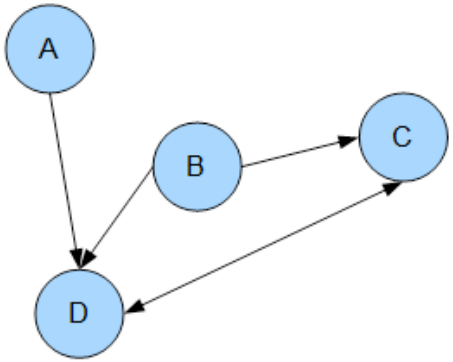
Naïve Bayes algorithm

TRAINMULTINOMIALNB($\mathbb{C}, \mathbb{D}$)

1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c / N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

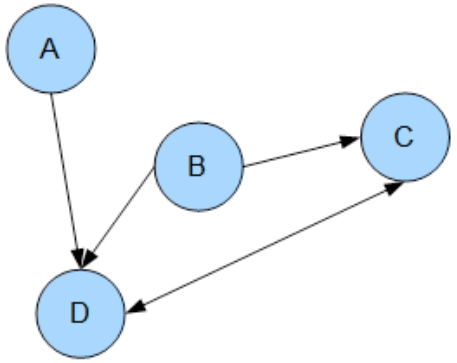
Types of graphical models

1. Bayesian Network
   - Represented as a directed acyclic graph (DAG)
   - Nodes: random variables, edges: conditional dependencies
   - Each node is associated with a probability function that takes as input a set of values for the node's parent variables



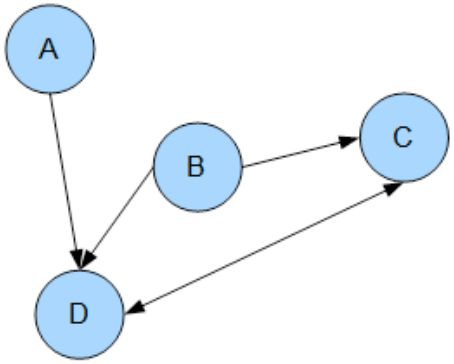
Types of graphical models

1. Bayesian Network
   - Represented as a directed acyclic graph (DAG)
   - Nodes: random variables, edges: conditional dependencies
   - Each node is associated with a probability function that takes as input a set of values for the node's parent variables
   - Local Markov property: each variable is conditionally independent of its non-descendants given its parent variables



Types of graphical models

1. Bayesian Network
   - Represented as a directed acyclic graph (DAG)
   - Nodes: random variables, edges: conditional dependencies
   - Each node is associated with a probability function that takes as input a set of values for the node's parent variables
   - Local Markov property: each variable is conditionally independent of its non-descendants given its parent variables

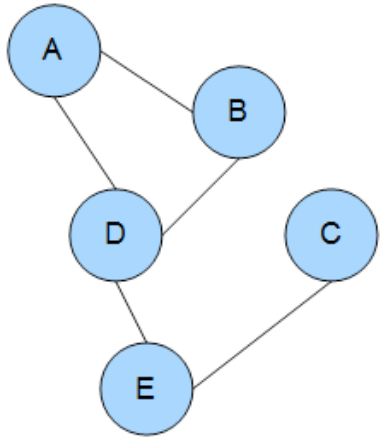


   - Important steps (supervised learning):
     - Inferring unobserved variables: BN can be used to find out updated knowledge of the state of a subset of variables (posterior distribution) when other variables (evidence) are observed a.k.a probabilistic inference.
     - Parameter learning: specify for each node x the probability distribution for x conditional upon its parents
     - Structure learning

Types of graphical models

Markov random fields

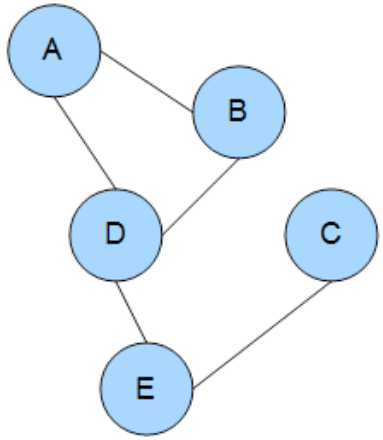
- Represented as an undirected graph (may be cyclic)
- Nodes: random variables, edges: dependencies for the Markov properties



Types of graphical models

Markov random fields

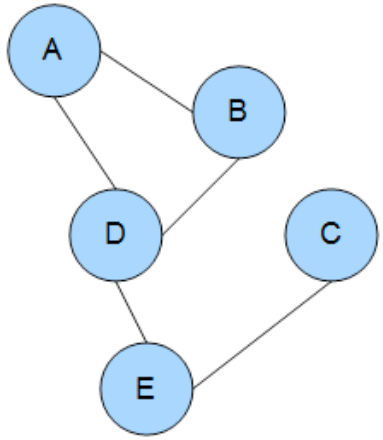
- Represented as an undirected graph (may be cyclic)
- Nodes: random variables, edges: dependencies for the Markov properties
- Markov properties:
  - Pairwise Markov property: Any two non-adjacent variables are conditionally independent given all other variables
  - Local Markov property: A variable is conditionally independent of all other variables given its neighbors
  - Global Markov property: Any two subsets of variables are conditionally independent given a separating subset (where every path from a node in one subset to a node in the second subset passes through separating subset)



Types of graphical models

Markov random fields

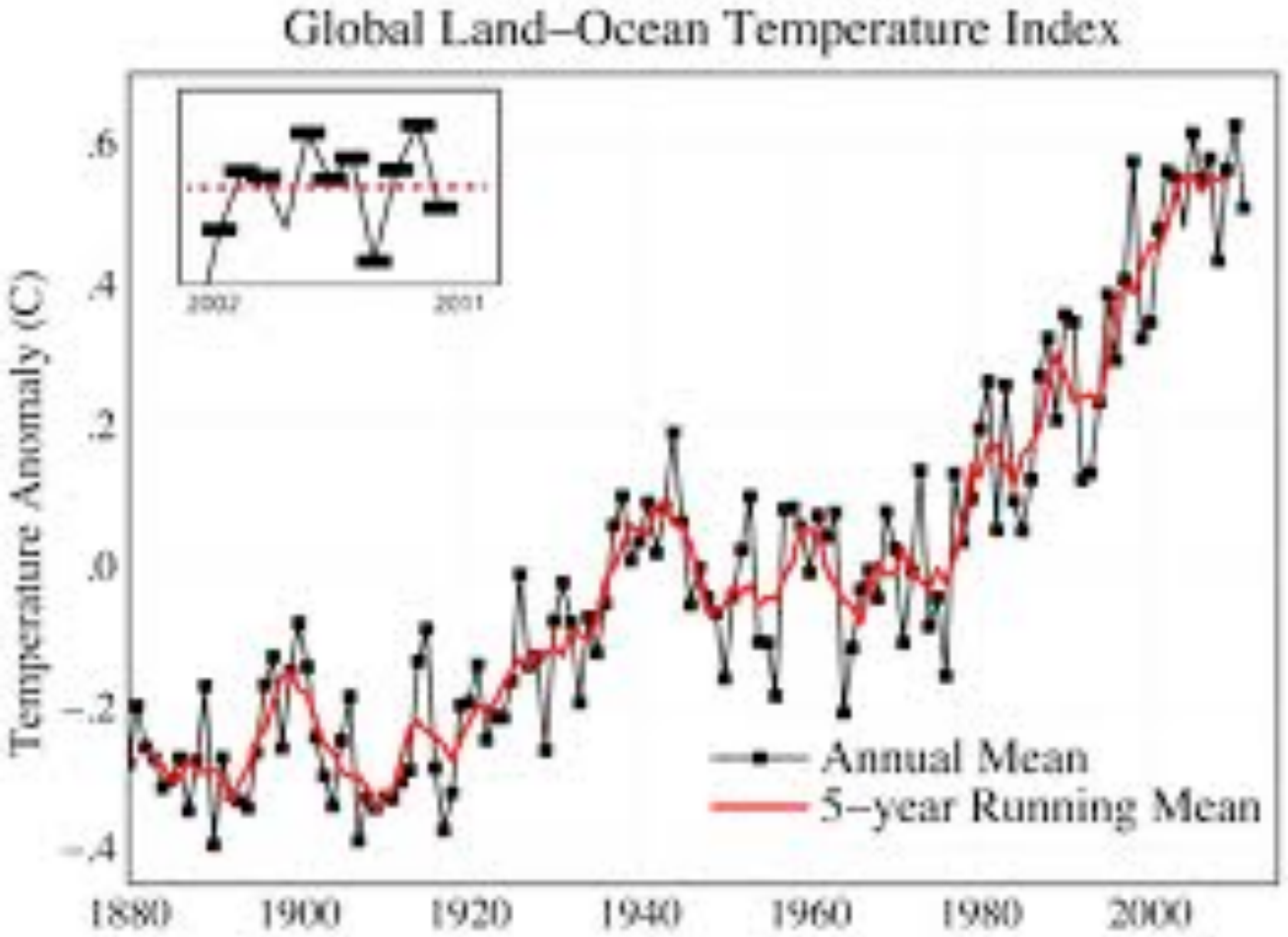
- Represented as an undirected graph (may be cyclic)
- Nodes: random variables, edges: dependencies for the Markov properties
- Markov properties:
  - Pairwise Markov property: Any two non-adjacent variables are conditionally independent given all other variables
  - Local Markov property: A variable is conditionally independent of all other variables given its neighbors
  - Global Markov property: Any two subsets of variables are conditionally independent given a separating subset (where every path from a node in one subset to a node in the second subset passes through separating subset)



- Pros: MRF can represent certain dependencies that a BN cannot (cyclic dependencies). Cons: it can't represent certain dependencies that a BN can (such as induced dependencies).

Hidden Markov Models (HMMs)

Intuitive example



- Global warming

Intuitive example: Global warming

- Challenge: derive temperature outside the years measured by measurement device
- Cannot observe the temperature directly
-
-

Intuitive example: Global warming

- Challenge: derive temperature outside the years measured by measurement device
- Cannot observe the temperature directly
- Can we do it indirectly? Yes!!!

- Suppose that modern evidence indicates that the probability of a hot year followed by another hot year is 0.7 and the probability that a cold year is followed by another cold year is 0.6

Intuitive example: Global warming (contd)

- Suppose that current research indicates a correlation between the size of tree growth rings and temperature
- For simplicity, we only consider three different tree ring sizes, small, medium and large: S, M and L,

Intuitive example: Global warming (contd)

- Suppose that current research indicates a correlation between the size of tree growth rings and temperature
- For simplicity, we only consider three different tree ring sizes, small, medium and large: S, M and L,

- Finally, suppose that based on available evidence, the probabilistic relationship between annual temperature and tree ring sizes is given by

$$\begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{ccc} S & M & L \end{array} \\ \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ H \\ C \end{array} \begin{array}{c} \begin{array}{cc} H & C \end{array} \\ \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \end{array}$$

Basic concepts in HMMs

- State is the average annual temperature (H or C)

- Transition from one state to the next is a Markov process of order one 1: dependent only on the previous year

Basic concepts in HMMs

- State is the average annual temperature (H or C)

- Transition from one state to the next is a Markov process of order one 1: dependent only on the previous year

- However, the actual states are hidden since we can't directly observe the temperature in the past

- Observations: we can observe the size of tree rings. Tree rings provide us with probabilistic information regarding the temperature

Basic concepts in HMMs

- State is the average annual temperature (H or C)

- Transition from one state to the next is a Markov process of order one 1: dependent only on the previous year

- However, the actual states are hidden since we can't directly observe the temperature in the past

- Observations: we can observe the size of tree rings. Tree rings provide us with probabilistic information regarding the temperature

- Since the states are hidden, this type of system is known as a Hidden Markov Model (HMM)

- Our goal is to make effective and efficient use of the observable information to gain insight into aspects of the Markov process.

Notation

T = length of the observation sequence

N = number of states in the model

M = number of observation symbols

Q = $\{q_0, q_1, \ldots, q_{N-1}\}$ = distinct states of the Markov process

V = $\{0, 1, \ldots, M-1\}$ = set of possible observations

A = state transition probabilities

B = observation probability matrix

π = initial state distribution

$\mathcal{O}$ = $(\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_{T-1})$ = observation sequence.

X = $(X_0, X_1, \ldots, X_{T-1})$

Notation

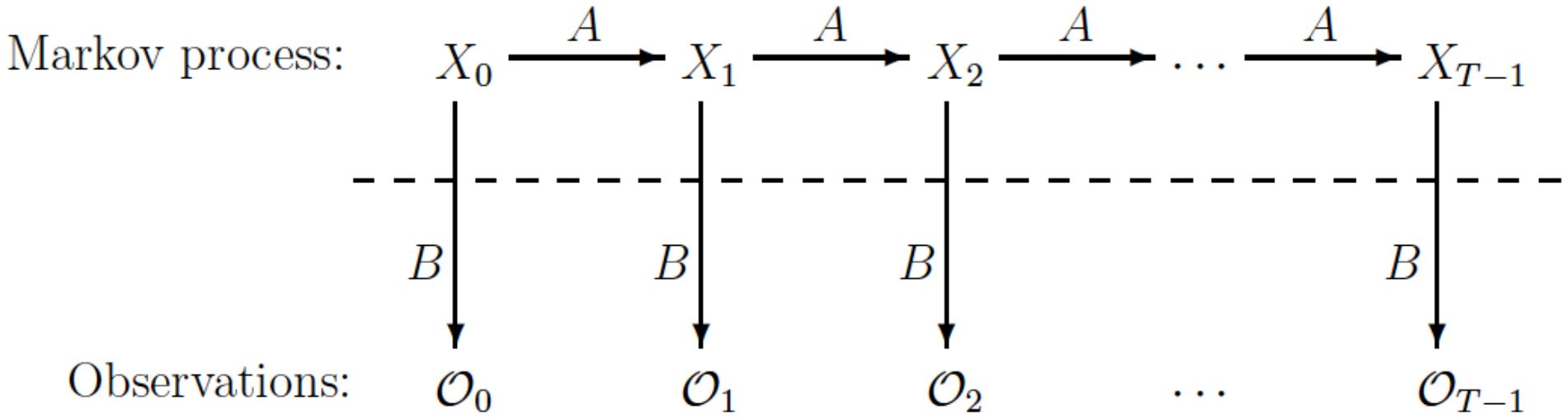
T = length of the observation sequence
N = number of states in the model
M = number of observation symbols
Q = $\{q_0, q_1, \ldots, q_{N-1}\}$ = distinct states of the Markov process
V = $\{0, 1, \ldots, M-1\}$ = set of possible observations
A = state transition probabilities
B = observation probability matrix
π = initial state distribution
$\mathcal{O}$ = $(\mathcal{O}_0, \mathcal{O}_1, \ldots, \mathcal{O}_{T-1})$ = observation sequence.
$\boldsymbol{X}$ = $(X_0, X_1, \ldots, X_{T-1})$

In our case:

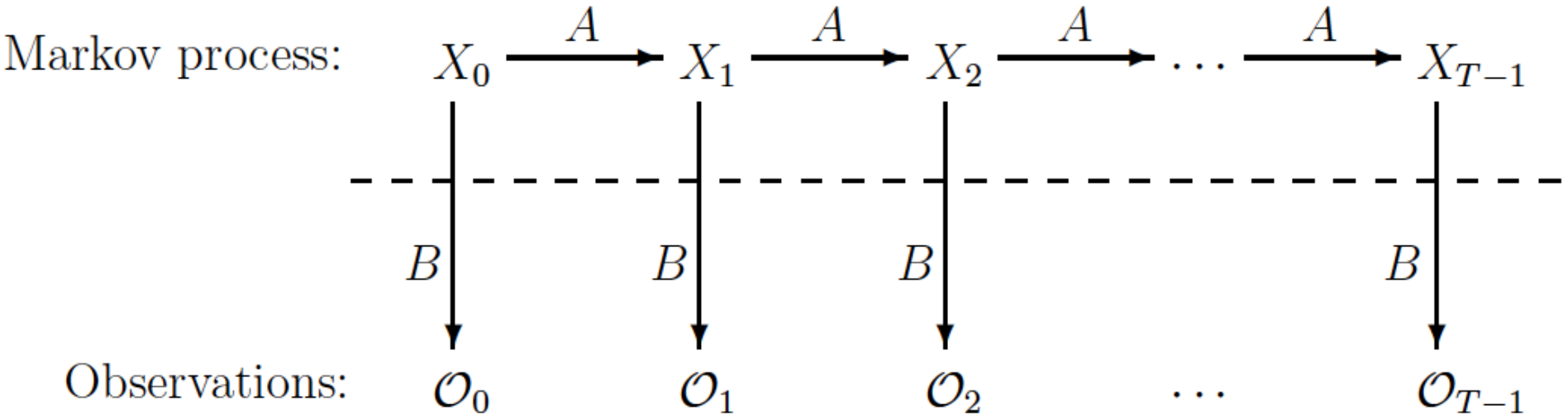
$$A = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \quad B = \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix} \quad \pi = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$$

er Polytechnic Institute

General structure of a HMM



General structure of a HMM



$A = \{a_{ij}\}$ is $N \times N$ with

$$a_{ij} = P(\text{state } q_j \text{ at } t+1 \,|\, \text{state } q_i \text{ at } t)$$

$B = \{b_j(k)\}$ is an $N \times M$ with

$$b_j(k) = P(\text{observation } k \text{ at } t \,|\, \text{state } q_j \text{ at } t).$$

Back to our example

- Consider a sequence of four observations:

$$T = 4,\ N = 2,\ M = 3,\ Q = \{H, C\},\ V = \{0, 1, 2\}$$

- That is, 0 represent S, 1 represent M and 2 represent L
- Assume that our observation sequence is $\mathcal{O} = (0, 1, 0, 2)$

Back to our example

- Consider a sequence of four observations:

$$T = 4,\ N = 2,\ M = 3,\ Q = \{H, C\},\ V = \{0, 1, 2\}$$

- That is, 0 represent S, 1 represent M and 2 represent L
- Assume that our observation sequence is

$$\mathcal{O} = (0, 1, 0, 2)$$

- Goal: determine the most likely state sequence of the Markov process given the observations. That is, determine the most likely average annual temperatures over the four-year period of interest

- Note: This is not quite as straight-forward as it seems, since there are different possible interpretations of “most likely."

Back to our example (contd)

- π_{x0} is the probability of starting in state x_0
- $\mathrm{b}_{x0}(O_0)$ is the probability of initially observing O_0
- $a_{x0,x1}$ is the probability of transiting from state x_0 to x_1

Back to our example (contd)

- π_{x0} is the probability of starting in state x_0
- $\mathrm{b}_{x0}(O_0)$ is the probability of initially observing O_0
- $a_{x0,x1}$ is the probability of transiting from state x_0 to x_1

- The probability of the state sequence X is given by

$$P(X) = \pi_{x_0} b_{x_0}(\mathcal{O}_0) a_{x_0,x_1} b_{x_1}(\mathcal{O}_1) a_{x_1,x_2} b_{x_2}(\mathcal{O}_2) a_{x_2,x_3} b_{x_3}(\mathcal{O}_3)$$

Back to our example (contd)

- π_{x0} is the probability of starting in state x_0
- $\mathrm{b}_{x0}(O_0)$ is the probability of initially observing O_0
- $a_{x0,x1}$ is the probability of transiting from state x_0 to x_1

- The probability of the state sequence X is given by

$$P(X) = \pi_{x_0} b_{x_0}(\mathcal{O}_0) a_{x_0,x_1} b_{x_1}(\mathcal{O}_1) a_{x_1,x_2} b_{x_2}(\mathcal{O}_2) a_{x_2,x_3} b_{x_3}(\mathcal{O}_3)$$

- That is, in our case the probability of a specific sequence of years:

$$P(HHCC) = 0.6(0.1)(0.7)(0.4)(0.3)(0.7)(0.6)(0.1) = 0.000212$$

Back to our example (contd)

- π_{x0} is the probability of starting in state x_0
- $\mathrm{b}_{x0}(O_0)$ is the probability of initially observing O_0
- $a_{x0,x1}$ is the probability of transiting from state x_0 to x_1

- The probability of the state sequence X is given by

$$P(X) = \pi_{x_0} b_{x_0}(\mathcal{O}_0) a_{x_0,x_1} b_{x_1}(\mathcal{O}_1) a_{x_1,x_2} b_{x_2}(\mathcal{O}_2) a_{x_2,x_3} b_{x_3}(\mathcal{O}_3)$$

- That is, in our case the probability of a specific sequence of years:

$$P(HHCC) = 0.6(0.1)(0.7)(0.4)(0.3)(0.7)(0.6)(0.1) = 0.000212$$

- Dynamic Programming sense: compute all possible combinations and select the highest

Probabilities for all combinations of 4 years

state	probability	normalized probability
$HHHH$	.000412	.042787
$HHHC$	.000035	.003635
$HHCH$	.000706	.073320
$HHCC$	.000212	.022017
$HCHH$	.000050	.005193
$HCHC$	.000004	.000415
$HCCH$	.000302	.031364
$HCCC$	.000091	.009451
$CHHH$	.001098	.114031
$CHHC$	.000094	.009762
$CHCH$	.001882	.195451
$CHCC$	.000564	.058573
$CCHH$	.000470	.048811
$CCHC$	.000040	.004154
$CCCH$	.002822	.293073
$CCCC$	.000847	.087963

Solving the problem in HMM sense

- To find the optimal sequence in the HMM sense, we choose the most probable symbol at each position
- We first sum the probabilities in Table 1 that have an H in the first position: we find the (normalized) probability of H in the first position is 0.18817
- Hence the probability of C in the first position is 0:81183, and the HMM chooses the first element of the
- optimal sequence to be C
- We then repeat this for each element of the sequence, obtaining the following probabilities:

Solving the problem in HMM sense

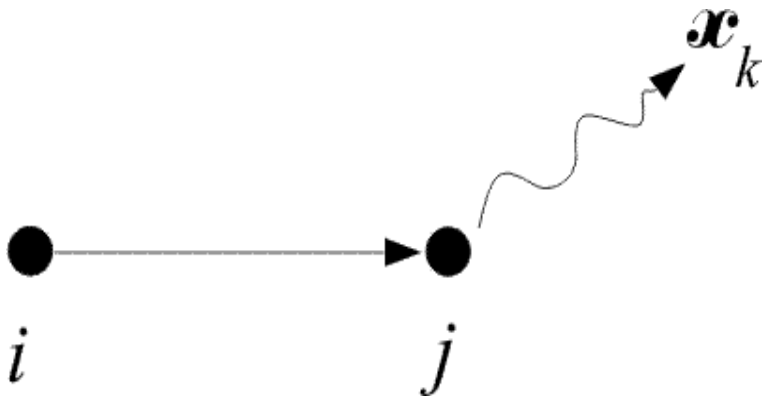
	element			
	0	1	2	3
$P(H)$	0.188182	0.519576	0.228788	0.804029
$P(C)$	0.811818	0.480424	0.771212	0.195971

- From this table, we find that the optimal sequence in the HMM sense is CHCH

- Note: in this example, the optimal DP sequence differs from the optimal HMM sequence and all state transitions are valid

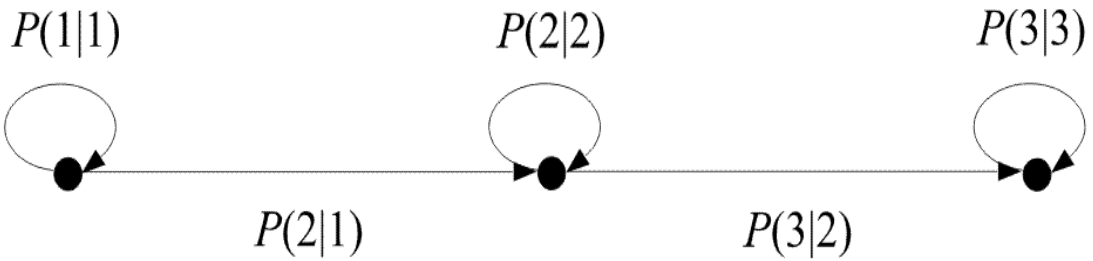
HMM as a stochastic automaton

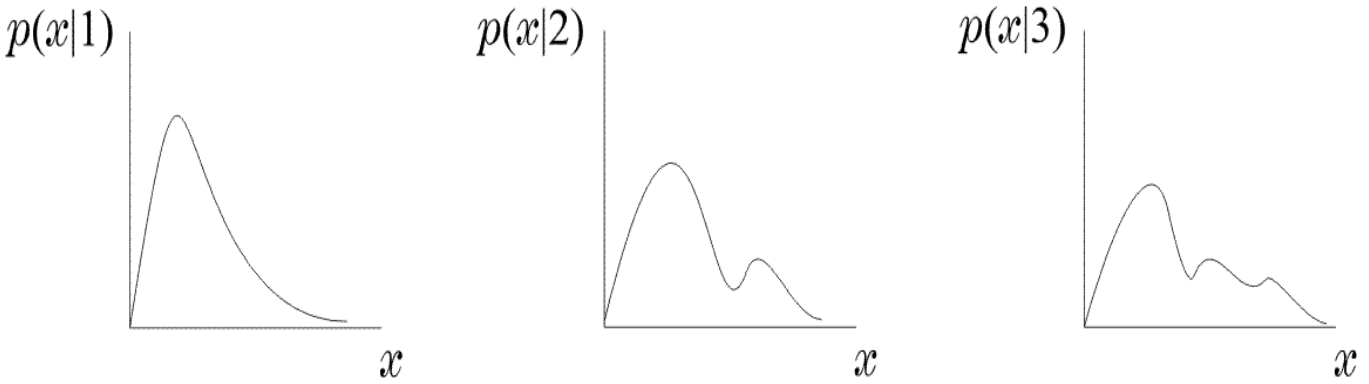
- An HMM is a stochastic finite state automaton, that generates the observation sequence, $x_1, x_2, \ldots, x_N$
- We assume that: The observation sequence is produced as a result of **successive** transitions between states, upon arrival at a state:



HMM as a stochastic automaton

This type of modeling is used for non-stationary stochastic processes that undergo distinct transitions among a set of different stationary processes.





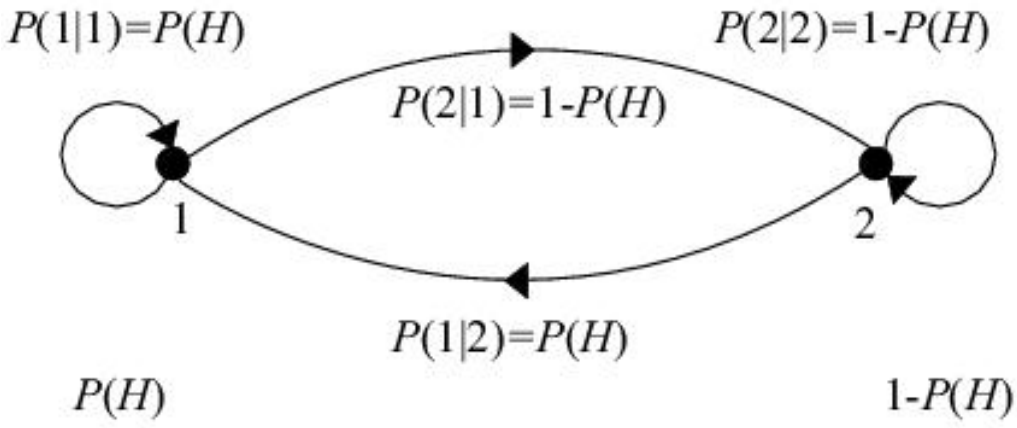
More examples: Single coin case

The single coin case: Assume a coin that is tossed behind a curtain. All it is available to us is the outcome, i.e., *H* or *T*. Assume the two states to be:

$$S = 1 \rightarrow H$$
$$S = 2 \rightarrow T$$

This is also an example of a random experiment with observable states. The model is characterized by a single parameter, e.g., $P(H)$. Note that

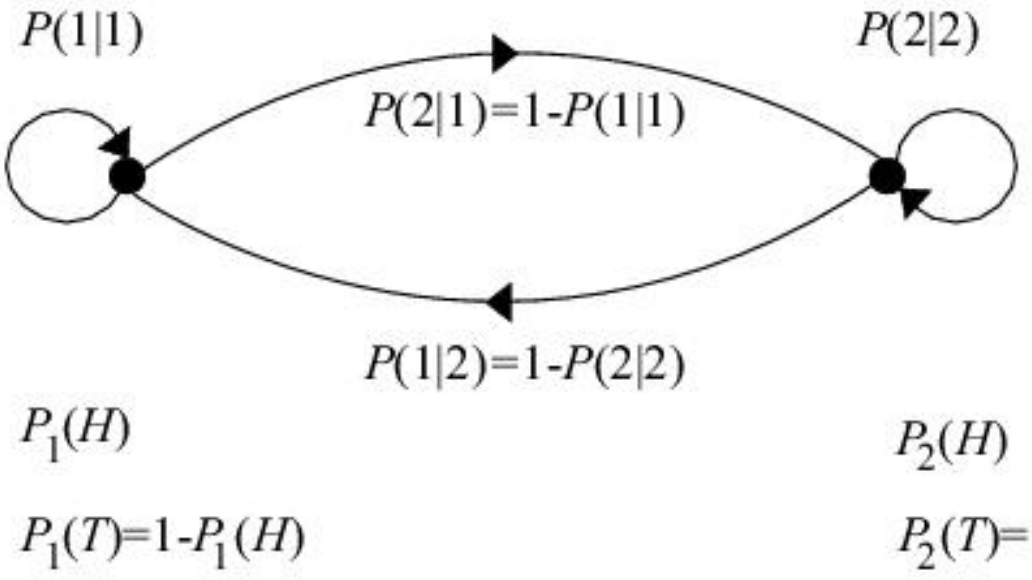
$$P(1|1) = P(H)$$
$$P(2|1) = P(\mathrm{T}) = 1 - P(H)$$



Single coin case (contd)

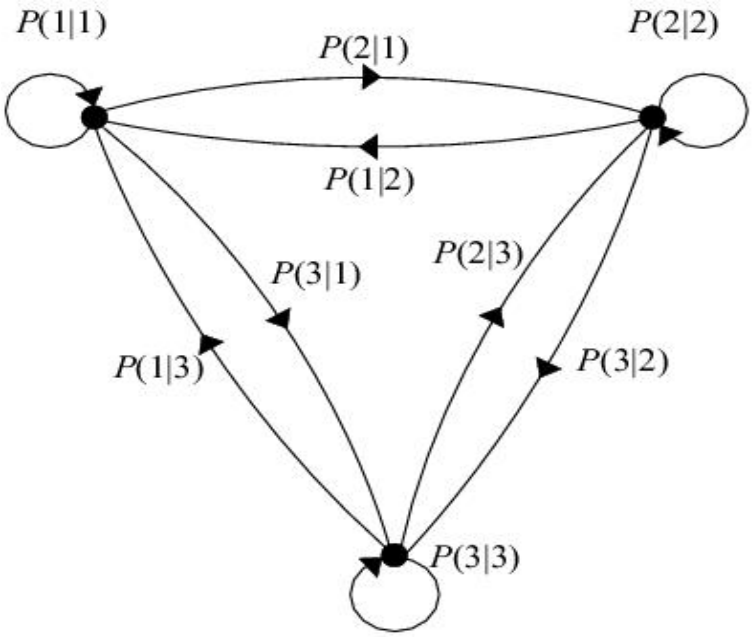
The two-coins case: For this case, we observe a sequence of *H* or *T*. However, we have no access to know which coin was tossed. Identify one state for each coin. This is an example where states are not observable. *H* or *T* can be emitted from either state. The model depends on four parameters.

$$P_1(H), P_2(H),$$
$$P(1|1), P(2|2)$$



Even more examples: Three coin case

- The three-coins case :



- Note that in all previous examples, specifying the model is equivalent to knowing:
  - The probability of each observation (H,T) to be emitted from each state
  - The transition probabilities among states: P(i|j).

Parameters for HMM model

A general HMM model is characterized by the following set of parameters

- N, number of states
- $A(i|j),\ i,j=1,2,...,N$
- $B(\boldsymbol{x}|i),\ i=1,2,...,N$
- $\pi(i),\ i=1,2,...,K$, initial state probabilities, $P(.)$

Patter recognition problem

That is:

$$S = \{A(i|j),\ \mathrm{B}(\boldsymbol{x}|i), \pi(i),\ N\}$$

What is the problem in Pattern Recognition

- Given *M* reference patterns, each described by an HMM, find the parameters, *S*, for each of them (training).
- Given an unknown pattern, find to which one of the *M*, known patterns, matches best (recognition).

Acknowledgements

- “A Revealing Introduction to Hidden Markov Models” by Mark Stamp, San Jose State University