Probability

Paul Rad, Ph.D.

Associate Professor
Cyber Analytics and AI
Information Systems and Cyber Security
College of Business School
210.872.7259

Outline

Unsupervised Learning

Sequences

Probability

Markov Models

Unsupervised or Supervised Learning?

- Unsupervised learning is for when we want to discover a patter or model a distribution but don't have any labels
- Supervised learning is for when we have labels, and want to be able to accurately predict the labels
- Hidden Markov Models are for modeling sequences: x(1), x(2), …x(n)
  - There is no label here
- But can be used for classification as well, e.g. recognizing if a sound sound signal is a male voice or female voice

Prediction based on previous values

Suppose we can predict next day's weather based on previous days

Ex. Monday is sunny, Tuesday is sunny, Wednesday is sunny, then the probability it will be sunny on Thursday is 90%

Many applications

Ex. NLP

"The Macbook was created by" → "Apple"? "Microsoft" ? " Whole Foods" ?

P("Apple" | "The Macbook was created by") = High Probability

Stock Market Prediction

Probability

Probability *is the formal study of the laws of chance. Probability allows us to* ***manage uncertainty****.*

The ***sample space*** *is the set of all* ***outcomes****. For example, for a die we have 6 outcomes:*

{ 1,2,3,4,5,6}

Experiments and Events

Experiment: Record an age

- A: person is 30 years old
- B: person is older than 65

Experiment: Toss a die

- A: observe an odd number
- B: observe a number greater than 2

The events are subsets of the sample space

Even = {2,4,6} , Odd = {1,3,5} , GreaterThan4 = {5,6}

We assign probabilities to these events:

P(Even) = ½ P(Odd) = 3/6 = ½ P(GreaterThan4) = 2/6 = 1/3

Axiom (I)

Probability of event = p

- 0 <= p <= 1
- 0 = certain non-occurrence
- 1 = certain occurrence

Axiom (II)

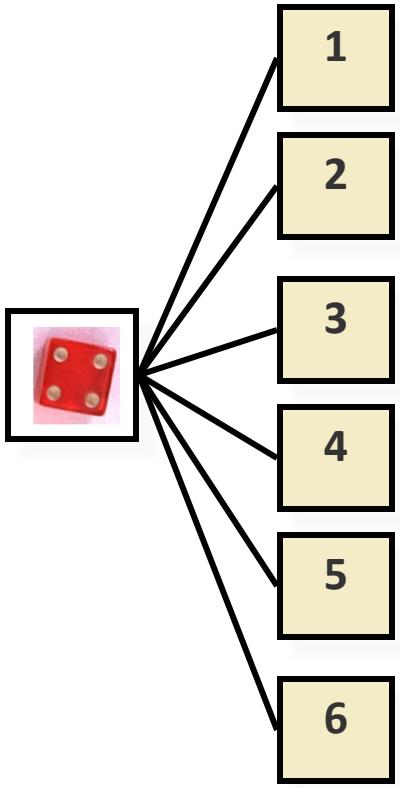
For **disjoint sets** A_n, $n \geq 1$, we have

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

Example

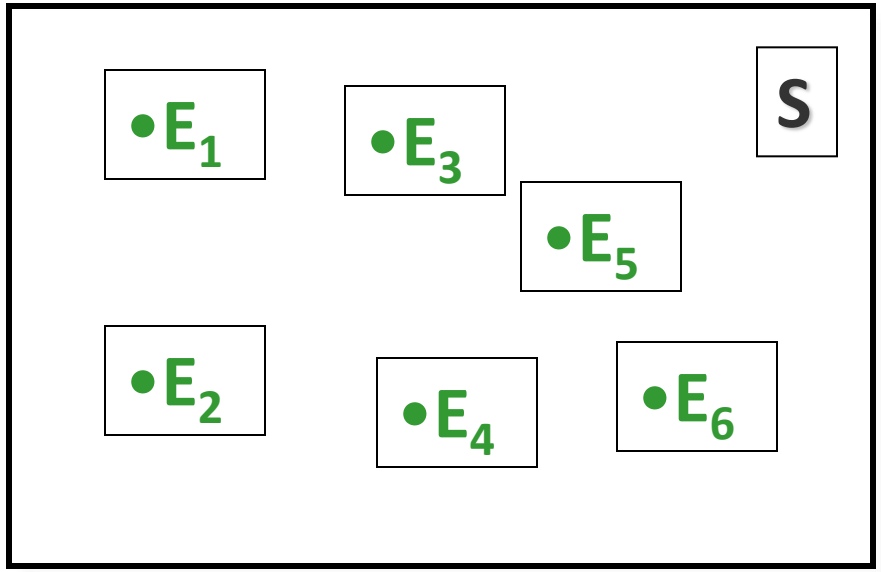
The die toss:

Simple Events



Sample Space

$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$

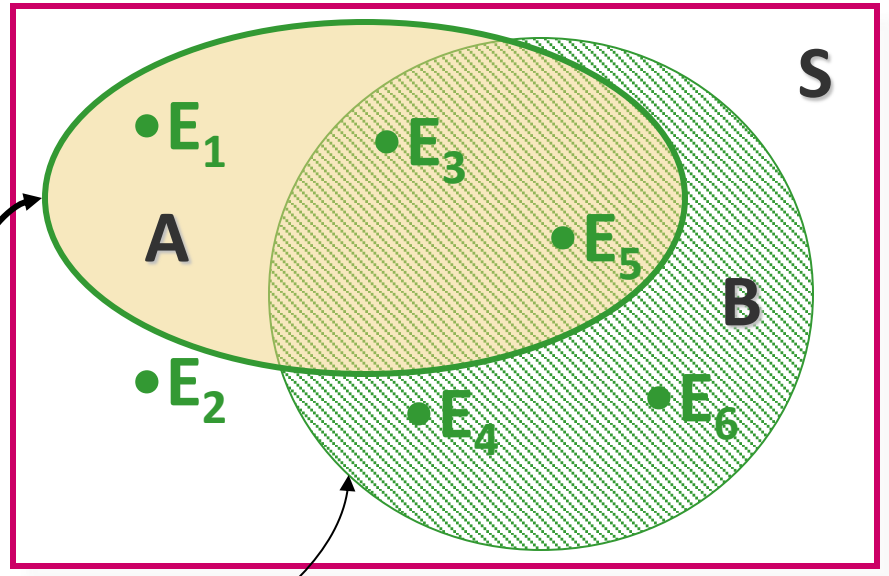


Basic Concepts

An **event** is a collection of one or more **simple events.**

The die toss:

- A: an odd number
- B: a number > 2



$A = \{E_1, E_3, E_5\}$

$B = \{E_3, E_4, E_5, E_6\}$

The Probability of an Event

The probability of an event A measures “how often” A will occur. We write **P(A).**

Suppose that an experiment is performed *n* times.

The relative frequency for an event A is

$$\frac{\text{Number of times A occurs}}{n} = \frac{f}{n}$$

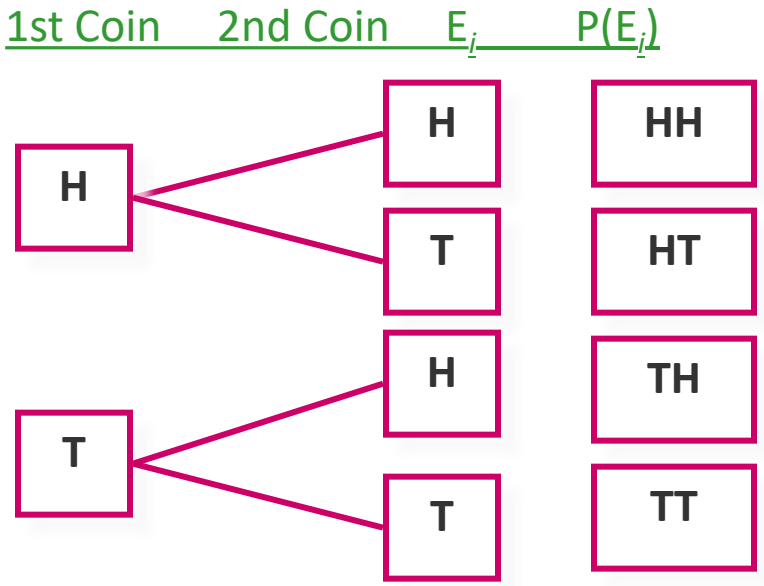
P(A) must be between 0 and 1.

- If event A can never occur, P(A) = 0.
- If event A always occurs when the experiment is performed, P(A) =1.

The sum of the probabilities for all simple events in S equals 1.

Example

Toss a fair coin twice. What is the probability of observing at least one head?



P(at least 1 head)

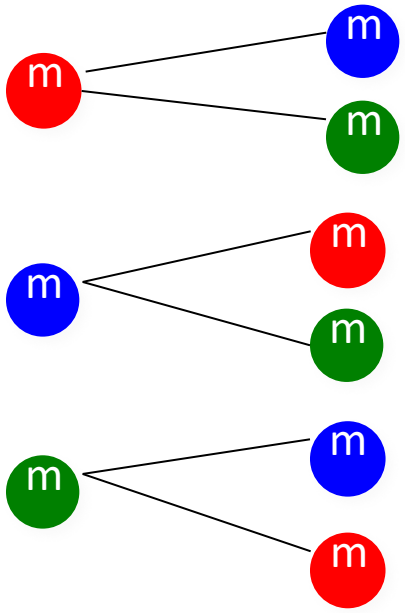
$= P(E_1) + P(E_2) + P(E_3)$

$= 1/4 + 1/4 + 1/4 = 3/4$

Example

A bowl contains three M&Ms®, one red, one blue and one green. A child selects two M&Ms at random. What is the probability that at least one is red?

1st M&M	2nd M&M	E_i	$P(E_i)$
m	m	RB	1/6
m	m	RG	1/6
m	m	BR	1/6
m	m	BG	1/6
m	m	GB	1/6
m	m	GR	



P(at least 1 red)

= P(RB) + P(BR)+ P(RG) + P(GR)

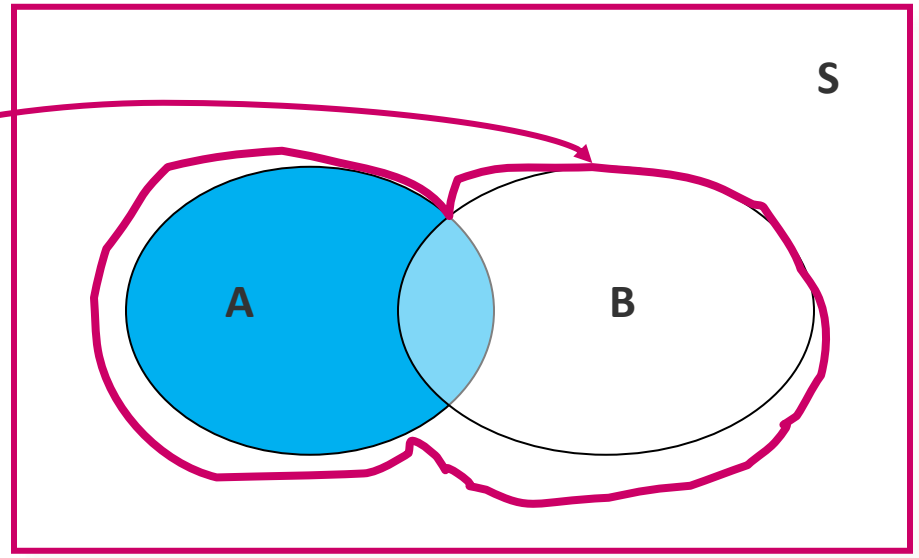
= 4/6 = 2/3

Or and And operation

The beauty of using events, rather than simple events, is that we can **combine** events to make other events using logical operations: and, or and not.

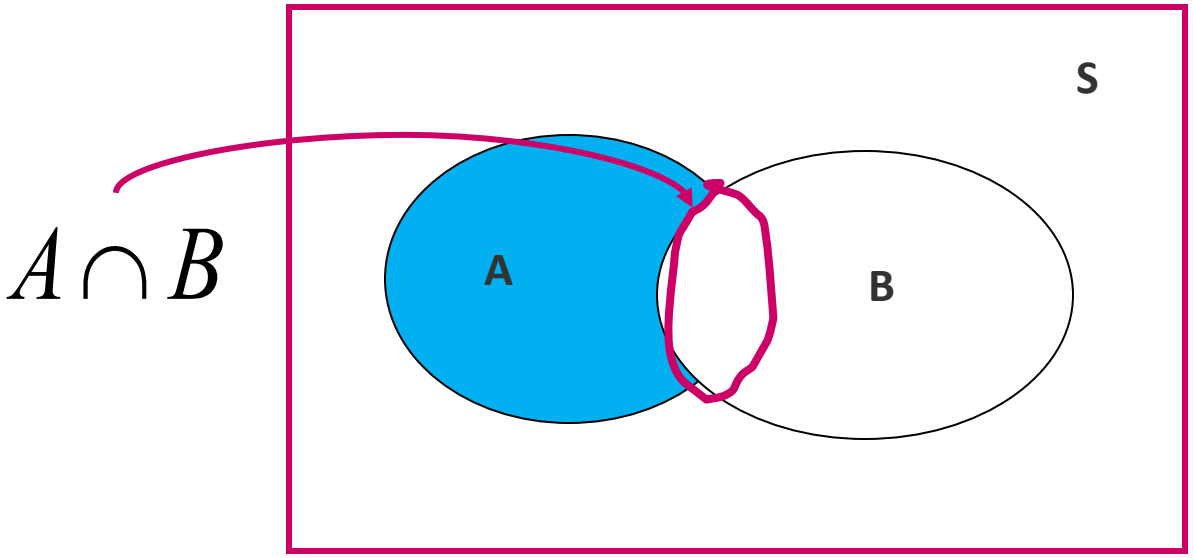
The **union** of two events, **A** and **B**, is the event that either **A or B or both** occur when the experiment is performed. We write

$$A \cup B$$



Or and And Operation

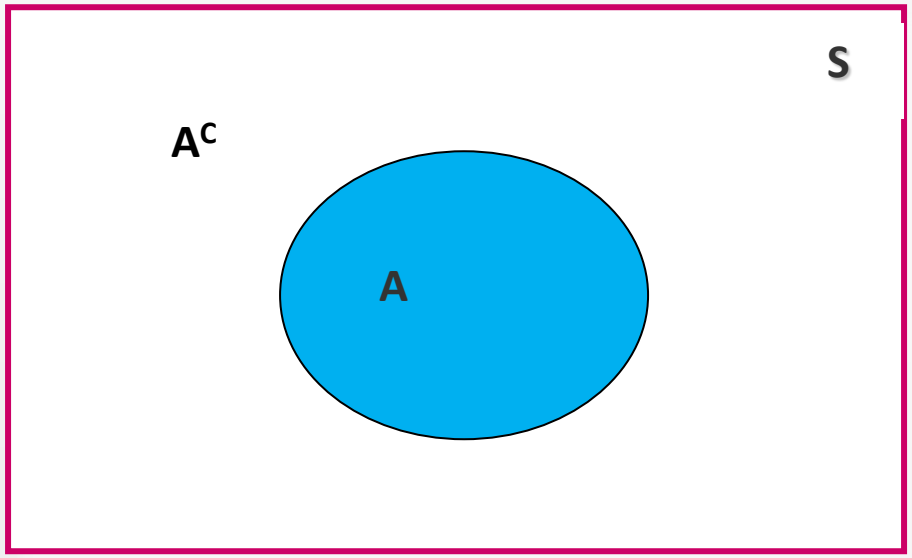
The **intersection** of two events, **A** and **B**, is the event that both A **and** B occur when the experiment is performed. We write **$A \cap B$.**



- If two events A and B are **mutually exclusive**, then **$P(A \cap B) = 0$.**

Complement

The **complement** of an event **A** consists of all outcomes of the experiment that do not result in event **A**. We write A^C.



Example

Example: Suppose that there were 120 students in the classroom, and that they could be classified as follows

A: brown hair

P(A) = 50/120

B: female

P(B) = 60/120

	Brown	Not Brown
Male	20	40
Female	30	30

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= 50/120 + 60/120 - 30/120$$
$$= 80/120 = 2/3$$

Calculating Probabilities for Complements

We know that for any event **A:**

- $P(A \cap A^C) = 0$

Since either **A** or A^C must occur,

$P(A \cup A^C) = 1$

so that $P(A \cup A^C) = P(A) + P(A^C) = 1$

$$P(A^C) = 1 - P(A)$$

Example

Select a student at random from the classroom. Define:

A: male

P(A) = 60/120

B: female

P(B) = ?

	Brown	Not Brown
Male	20	40
Female	30	30

A and B are complementary, so that

P(B) = 1- P(A)
= 1- 60/120 = 60/120

Conditional Probabilities

The probability that A occurs, given that event B has occurred is called the **conditional probability** of A given B and is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ if } P(B) \neq 0$$

"given"

The Multiplicative Rule for Intersections

For any two events, **A** and **B**, the probability that both **A** and **B** occur is

$$P(A \cap B) = P(A)\,P(B \text{ given that } A \text{ occurred})$$
$$= P(A)P(B|A)$$

If the events **A** and **B** are independent, then the probability that both **A** and **B** occur is

$$P(A \cap B) = P(A)\,P(B)$$

Example

In a certain population, 10% of the people can be

classified as being high risk for a heart attack. Three people are randomly selected from this population. What is the probability that exactly one of the three are high risk?

H: high risk

N: not high risk

$$P(\text{exactly one high risk}) = P(HNN) + P(NHN) + P(NNH)$$

$$= P(H)P(N)P(N) + P(N)P(H)P(N) + P(N)P(N)P(H)$$

$$= (.1)(.9)(.9) + (.9)(.1)(.9) + (.9)(.9)(.1) = 3(.1)(.9)^2 = .243$$

Example

Suppose we have additional information in the previous example. We know that only 49% of the population are female. Also, of the female patients, 8% are high risk. A single person is selected at random. What is the probability that it is a high risk female?

H: high risk

F: female

From the example, $P(F) = .49$ and $P(H|F) = .08$.

Use the Multiplicative Rule:

$$P(\text{high risk female}) = P(H \cap F)$$

$$= P(H|F)\,P(F) = (.08)\,.49 = .0392$$

Bayes' Rule

Let S_1 , S_2 , S_3 ,..., S_k be mutually exclusive and exhaustive events with prior probabilities $P(S_1)$, $P(S_2)$,...,$P(S_k)$. If an event A occurs, the posterior probability of S_i, given that A occurred is

$$P(S_i \mid A) = \frac{P(S_i)P(A \mid S_i)}{\sum P(S_i)P(A \mid S_i)} \text{ for } i = 1, 2, \ldots k$$

Example

From a previous example, we know that 49% of the population are female. Of the female patients, 8% are high risk for heart attack, while 12% of the male patients are high risk. A single person is selected at random and found to be high risk. What is the probability that it is a male?

Define H: high risk F: female M: male

We know:
P(F) =
P(M) =
P(H|F) =
P(H|M) =

$$P(M|H) = \frac{P(M)P(H|M)}{P(M)P(H|M) + P(F)P(H|F)}$$

$$= \frac{.51(.12)}{.51(.12) + .49(.08)} = .61$$

Example

Suppose a rare disease infects one out of every 1000 people in a population. And suppose that there is a good, but not perfect, test for this disease: if a person has the disease, the test comes back positive 99% of the time. On the other hand, the test also produces some false positives: 2% of uninfected people are also test positive. And someone just tested positive. What are his chances of having this disease?

Define A: has the disease B: test positive

We know:
P(A) = .001
P(A^c) =.999
P(B|A) = .99
P(B|A^c) =.02

We want to know P(A|B)=?

$$P(A \mid B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A^C)P(B|A^C)}$$

$$= \frac{.001 \times .99}{.001 \times .99 + .999 \times .02} = .0472$$

Example

Assume you have an AI program that is trying to estimate the posterior probability that you happy or sad, given that the AI program has observed whether you are

- Watching Game of Thrones (w)
- Sleeping (s)
- Crying (c)
- Facebooking (f)

Let the unknown state be

X = h if you are happy

X = s if you are sad

Let Y denote the observation, which can be w,s,c, or f

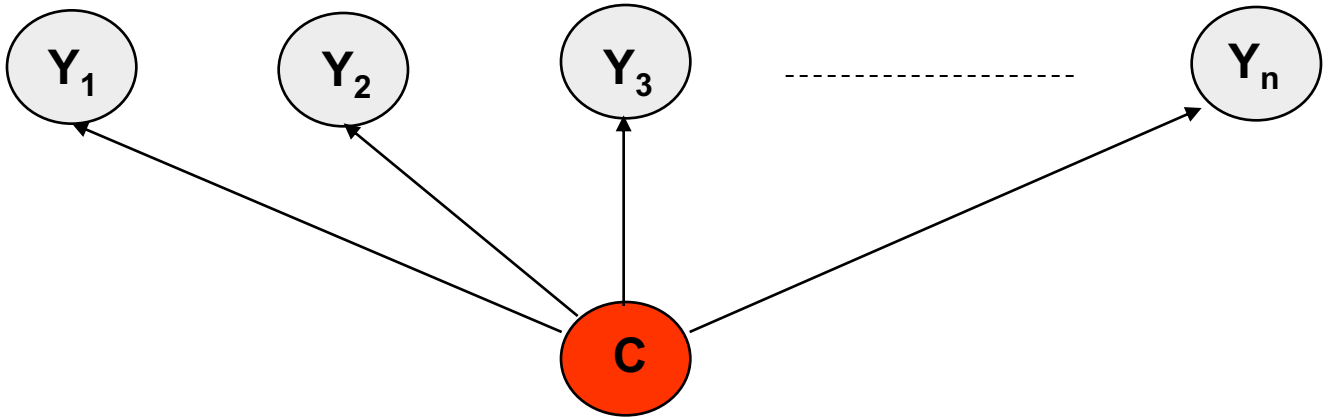
P(X=h | Y=f) =?

P(X=s | Y=c) = ?

Dynamic model

In general, we assume we have an initial distribution P(x), a transition model P(xt|xt-1), and an observation model P (Yt|Xt)

Naïve Bayes Model



$$P(C \mid Y_1,\ldots Y_n) = \alpha \prod P(Y_i \mid C)\ P(C)$$

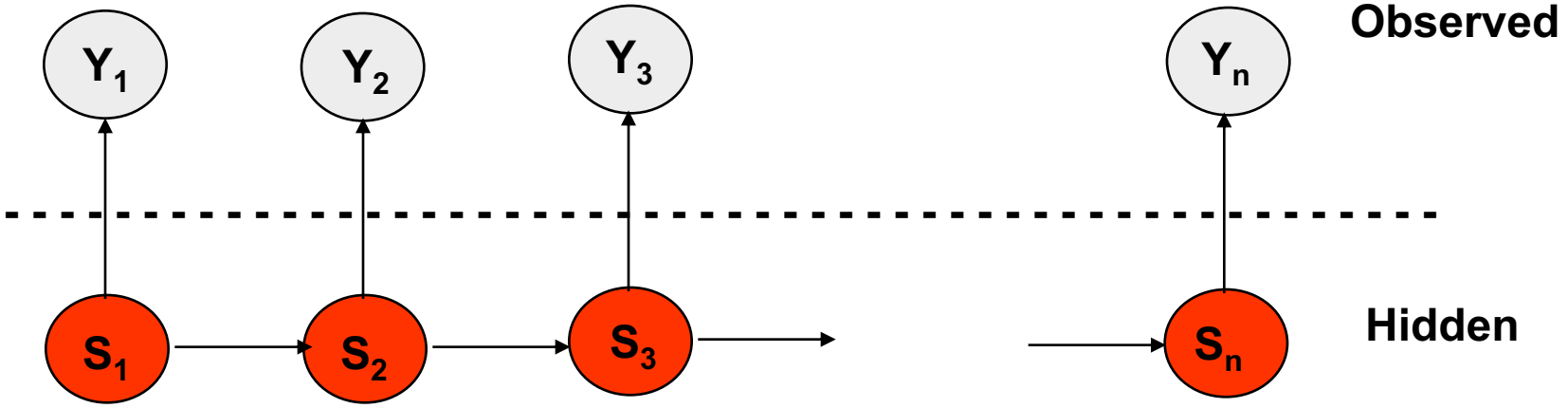
Features Y are conditionally independent given the class variable C

Widely used in machine learning

e.g., spam email classification: Y's = counts of words in emails

Conditional probabilities P(Yi | C) can easily be estimated from labeled data

Hidden Markov Model (HMM)



Two key assumptions:

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

Widely used in speech recognition, protein sequence models

Since this is a Bayesian network polytree, inference is linear in n

Markov Property

P (weather(today) | weather (all previous days)) = p (weather (today) | weather (yesterday))

P(word(n) | word (n-1), …word(1)) = p(word(n) | word(n-1))

In general" states"

State at time t: s(t)

p(s(t) | s(t-1), s(t-2), …, s(0)) = p(s(t)|s(t-1))

Using chain role of probability:

P(s4,s3,s2,s1) = p(s4|s3,s2,s1)p(s3,s2,s1)

= p(s4|s3,s2,s1)p(s3|s2,s1)p(s2,s1)

= p(s4|s3,s2,s1)p(s3|s2,s1)p(s2|s1)p(s1)

= p(s4|s3)p(s3|s2)p(s2|s1)p(s1)

Terminology

If we have only p(s(t)|s(t-1)) → First order Markov (*)

p (s(t)|s(t-1),s(t-2)) → Second order Markov

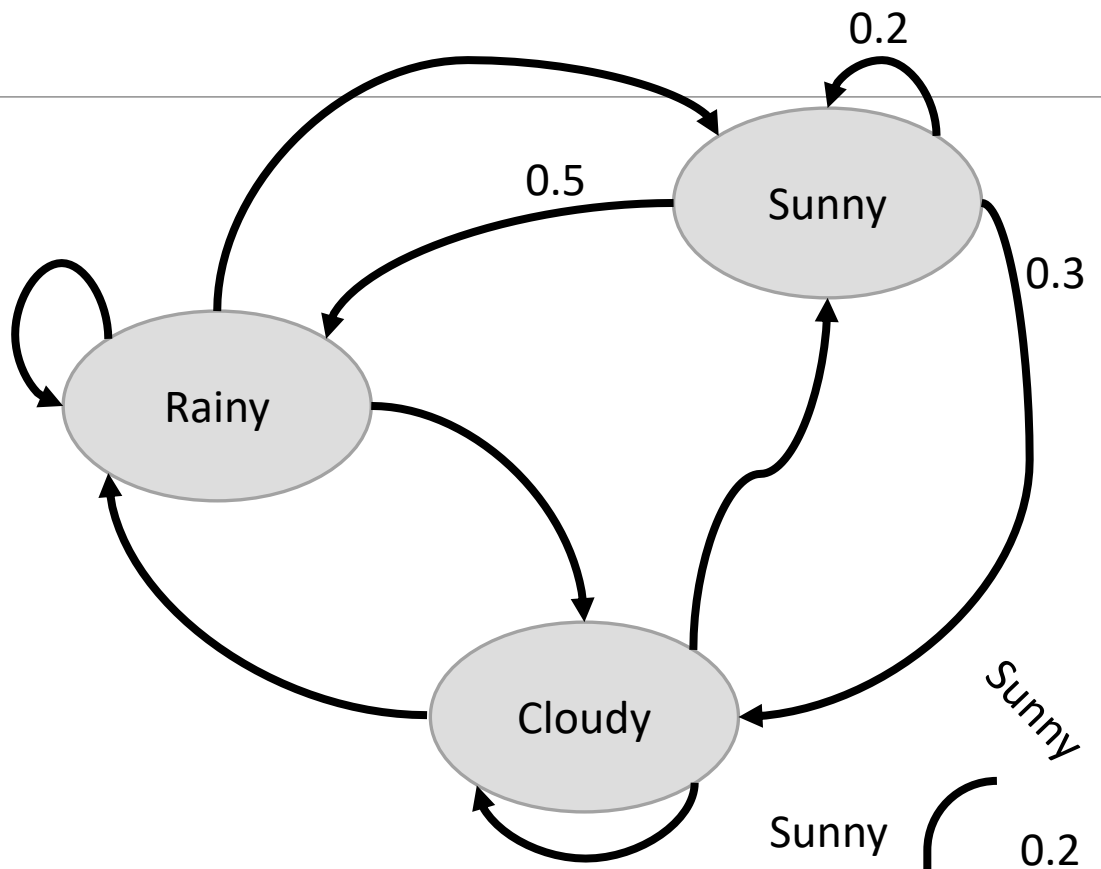
p (s(t)|s(t-1),s(t-2), s(t-3)) → Third order Markov

Markov Models

Weather example: 3 states : sunny, rainy, cloudy

How many weights?

Each state can go to each state, including itself



	Sunny	Rainy	Cloudy
Sunny	0.2	0.5	0.3
Rainy	0.6	0.1	0.3
Cloudy	0.2	0.3	0.5

M states → M*M weights

A → MxM matrix

A(i,j) = Probability to go to state j from state i

A(i,j) = p(s(t) = j | s(t-1) = i)

Constraints: A(i, :) must sum to 1, i=1,…M

In general:

p(s(0),s(1),s(2), s(t)) = p(s(0)).p(s(1)|s(0)).p(s(t)|s(t-1))

Training a Markov Model

Maximum likelihood! Very simple:

I like dogs

I like cats

I love kangaroos

6 states: 0=I, 1=like, 2=love, 3=dogs, 4=cats, 5=kangaroos

s(0) = [1,0,0,0,0,0]

p(like|I) = 2/3

p(love | I) = 1/3

p(dogs | like) = ½

p(cats | like) = ½

p(kangaroos | love) =1