Markov Chain Model

DP를 이용하는 방법,

ML에서 많이 활용된 방법 - HMM

Hidden Markov Model

질환 / 정상 구별 feature

→ 'serial 한 개념' : 문장

signal processing

주식, DNA, → 앞 정보로 뒷 정보 예측

An example of Markov chain model

Given an observation sequence of states, is this biased dice or not?

$x=\underbrace{123324}_{f}\underbrace{45612}_{b}\underbrace{3535}_{f}$

두개의 주사위, 어디서부터 어디까지?

fair / bias → fragmentation 문제

정상 주사위 / 유리한 주사위

앞뒤 순서 고려 (series data grouping)

- 모델 주어짐 → Supervised
- 모델 파라미터 모름 / No label → EM

Given an observation sequence of states, is it a year of good harvest?

x=RRRSSRSRSSSRSR

R = Rainy, S = Sunny

날씨의 변화만 보고 풍년/흉년 예측하는 모델 만듦 → series 패턴 리게 중요

An example of Markov chain model

Given an observation sequence of states, is this gene or non-gene region?

가 있을때 결정해야되는 문제

x=CGCCGTACGTGT

DNA에서 1%인 gene이 어떤 부분에 있는지 알아내는 모델

(유전자 영역 찾기)

Probability

많은 probability 에는 iid가 assumption 으로 들어가는데 이것을 Markov Chain으로 Relax

모든 것은 독립적으로 시행된다 이전상태 정보 이용X

- Assumption of independent and identically distributed (iid) may lose information (correlation between observations that are close in the sequence) in sequential data

앞에 어떻게 나왔느냐에 따라서
확률이 달라지는 문제이므로

- eg. weather forecast
- Relax iid assumption by using Markov chain

" 이전상태" 정보 이용

앞뒤 관계 고려

Reasoning over time or space

- Need to reason about a sequence of observations
  - speech recognition
  - stock market
  - DNA sequence information

series의 정보가 들어가는 것

→ need to introduce time (order) into the models

Chain rule and Markov chain

- From the chain rule, every joint distribution over X_1, X_2, X_3, X_4 is

$$P(X_1, X_2, X_3, X_4) = P(X_1)\, P(X_2 | X_1)\, P(X_3 | X_1, X_2)\, P(X_4 | X_1, X_2, X_3)$$

given

독립

X2가 주어졌을 때

X1이 X3의 probability와 independent 하다면

Independency를

- If we assume that

assumption으로 줌

$$X_3 \perp\!\!\!\perp X_1 \mid X_2 \quad \text{and} \quad X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$$

independent

X3이 주어지면 X1, X2에 대해선 Independent

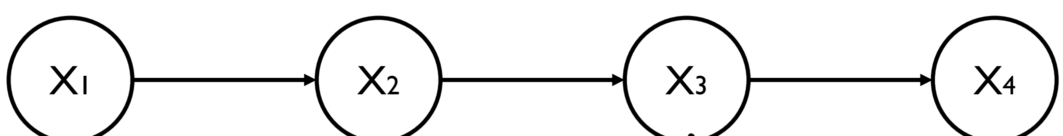
the joint probability is

이전것만 고려

$$P(X_1, X_2, X_3, X_4) = P(X_1)\, P(X_2 | X_1)\, P(X_3 | X_2)\, P(X_4 | X_3)$$

X3 given X2

Conditional prob. → 연속



바로 전것만 고려

X2가 주어지면 X1에 대해서 Independent하다

Overview of Markov chain

- The state at any time is denoted by X_t
- A sequence of length T is denoted by $X = \{X_1, X_2, \ldots, X_T\}$, $X_i \in \{w_1, w_2, w_3\}$

$$X = X_1\ X_2\ X_3\ \ldots\ X_T$$

- A sequence $X = \{w_1, w_3, w_3, w_2, w_1, w_2\}$
- $X = w_1\ w_3\ w_3\ w_2\ w_1\ w_2$

Conditional probability를
Markov chain 모델에서 만들어주는 파라미터

- Parameters in Markov chain are transition probabilities

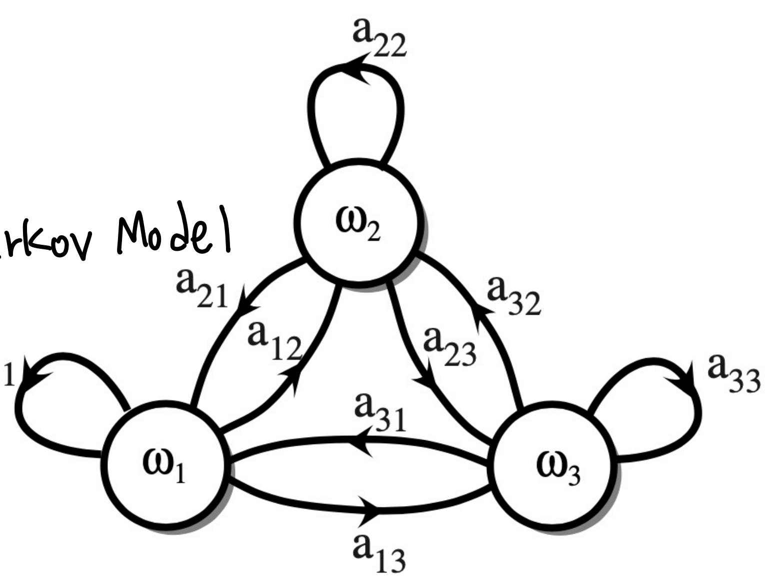
Sequence의 probability 계산

$P(X_{t+1} = w_j \mid X_t = w_i) = a_{ij}$ → $S = w_i\ w_j$

Conditional probability

the probability of having state w_j at step t+1 given that the state at time t was w_i

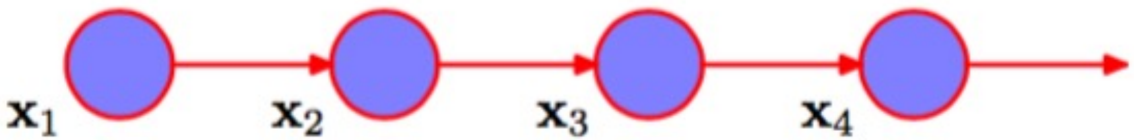
Markov Model



1. Markov 모델이 뭔지
2. 왜 transition probability가 필요한지

Markov chain of order n

- First-order Markov chain of observations
  바로 전 상태만 이용
  - allows earlier observation to have an influence
  - $p(x_n \mid x_{n-1})$, the distribution of a particular observation x_n is conditioned on the value of the previous observation x_{n-1}



$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n | \mathbf{x}_{n-1})$$

Markov chain of order n

- Second-order Markov chain of observations

  두 개의 상태 이용

  - the conditional distribution of x_n given x_{n-1} and x_{n-2} is independent of all other observation $x_1, x_2, \ldots, x_{n-3}$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1) \prod_{n=3}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2})$$

Markov chain of order n

- *n*th order Markov chain: the process in which the current state depends on n previous states

$$P(x_i \mid x_{i-n} \dots x_{i-2} x_{i-1})$$

n개의 전 상태 고려

전상태를 많이 고려하면

order를 높이면 오버피팅의 단점이 있음

An example of nth-order Markov chain

1st order MC

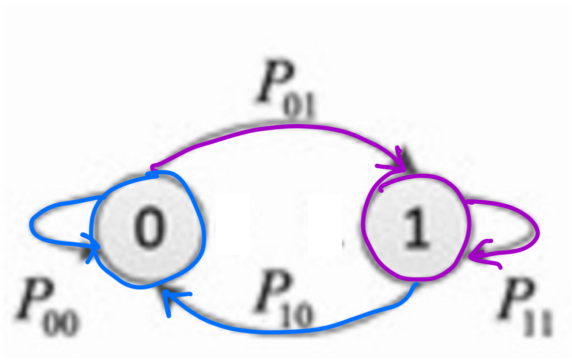
State X = {0, 1}

Sequence 에 들어오는 Symbol 두개

An example of nth-order Markov chain

1st order MC

State X = {0, 1}



An example of nth-order Markov chain

2nd order MC

S = 00101100111

0으로 오는 방법

00 → 0
01 → 0
10 → 0
11 → 0

An example of nth-order Markov chain

2nd order MC

S = 00101100111

State X = {00, 01, 10, 11}

000
010
100
110
001
011
101
111

00	1
01	2
10	3
11	4

An example of nth-order Markov chain

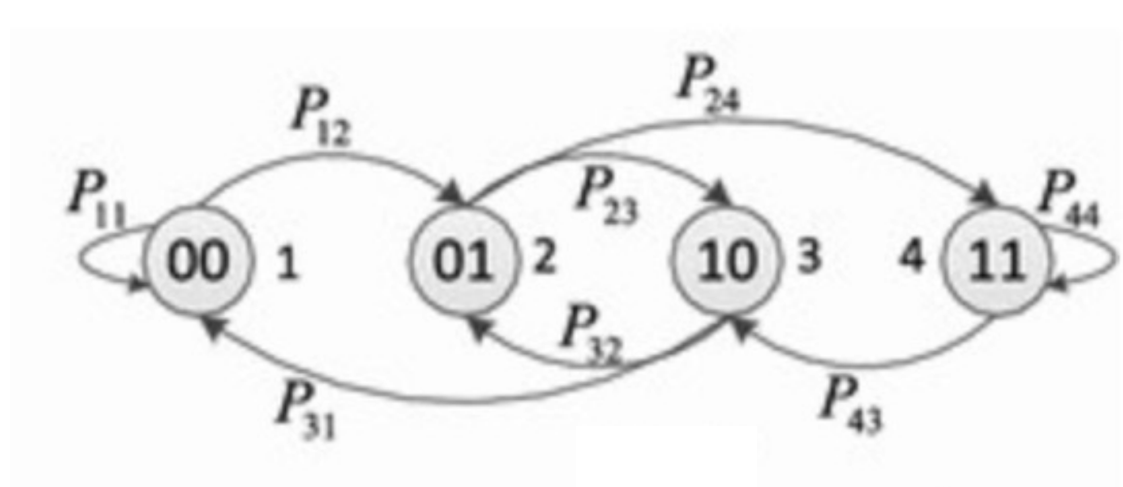
2nd order MC

S = 00101100111

State X = {00, 01, 10, 11}

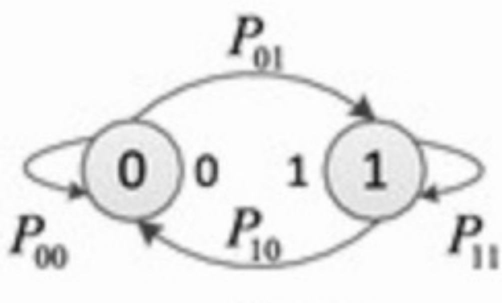
000
010
100
110
001
011
101
111

00	1
01	2
10	3
11	4

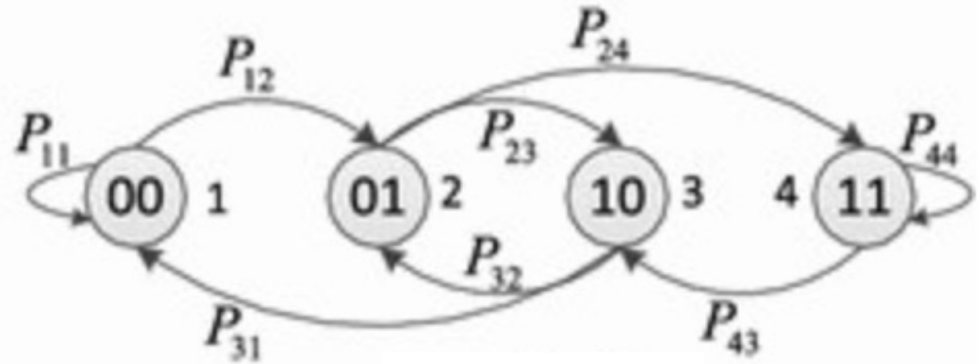


파라미터 개수: 8개 (2+2+2+2)

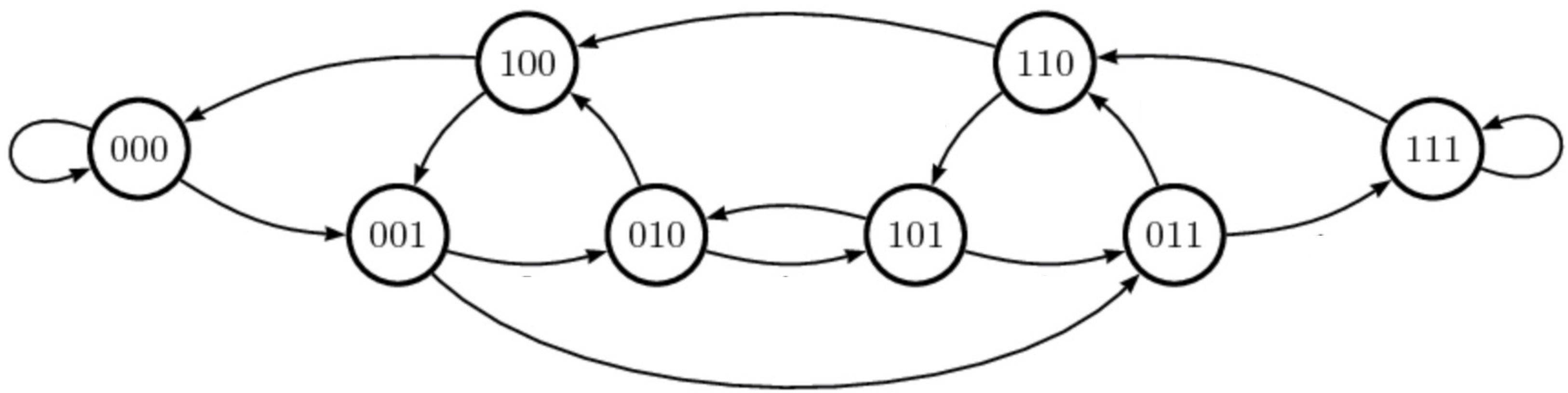
An example of nth-order Markov chain



1st order MC



2nd order MC

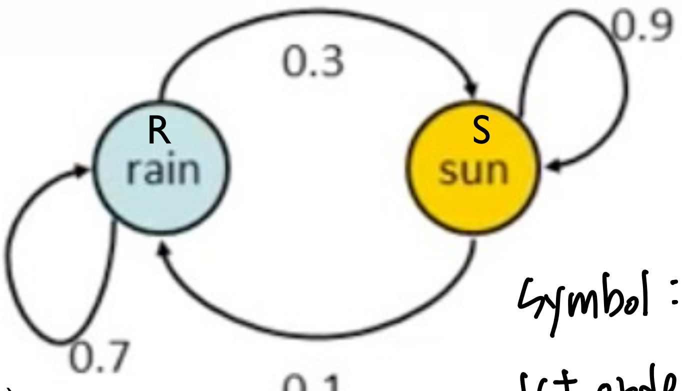


3rd order MC

Markov chain of weather

P(sun|rain) = a_{RS}

P(sun|sun) = a_{SS}



P(rain|rain) = a_{RR}

P(rain|sun) = = a_{SR}

Symbol : 2개

1st order

→ transition probability : 4개

independent parameter : 2개

states X = {rain, sun}

initial distribution: $P(X_1 = \text{sun}) = 1$

Conditionaly Probability Table

CPT $P(X_t \mid X_{t-1})$

X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	rain	0.7

$$P(X_2) = \sum_{X_1} P(X_2, X_1) = \sum_{X_1} P(X_2 \mid X_1)\, P(X_1)$$

↳ initial prob.

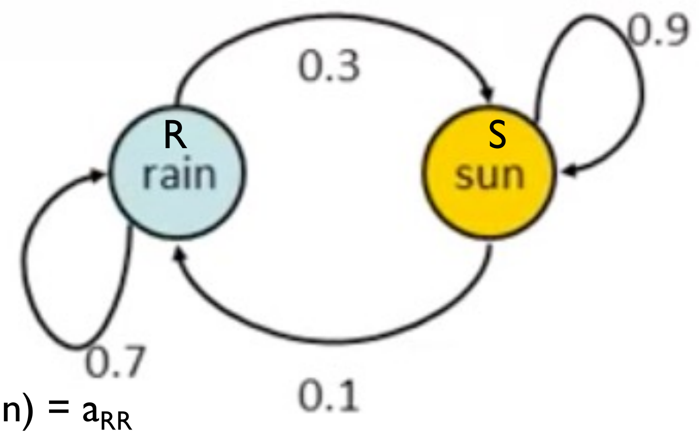
$$P(X_2 = \text{sun}) = P(X_2 = \text{sun} \mid X_1 = \text{sun})P(X_1 = \text{sun}) + P(X_2 = \text{sun} \mid X_1 = \text{rain})P(X_1 = \text{rain})$$

$$= 0.9\ 1.0 + 0.3\ 0.0 = 0.9$$

$$P(X_2 = \text{rain}) = 0.1 = 1-0.9$$

Markov chain of weather

$P(sun|rain) = a_{RS}$ $P(sun|sun) = a_{SS}$



$P(rain|rain) = a_{RR}$

$P(rain|sun) = = a_{SR}$

states $X = \{rain, sun\}$

initial distribution: $P(X_1 = sun) = 1$

CPT $P(X_t \mid X_{t-1})$

X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	rain	0.7

$P(X_2 = sun) = 0.9$, $P(X_2 = rain) = 0.1$ ← 둘째날만 고려

$$P(X_3 = sun) = P(X_3 = sun \mid X_2 = sun)P(X_2 = sun) + P(X_3 = sun \mid X_2 = rain)P(X_2 = rain)$$

$$= 0.9\ 0.9 + 0.1\ 0.3 = 0.84$$

$$P(X_3 = rain) = 0.16$$

An example of Markov chain

What is the P(X) on some day t?

$P(X_1) = p$

$P(X_t) = \sum_{X_{t-1}} P(X_{t-1}, X_t)$ length t 일때 transition prob. 다 곱함

$= \sum_{X_{t-1}} P(X_t \mid X_{t-1}) P(X_{t-1})$

An example of Markov chain

$$P(X_n = sun) = P(X_n = sun \mid X_{n-1} = sun)P(X_{n-1} = sun) + P(X_n = sun \mid X_{n-1} = rain)P(X_{n-1} = rain)$$

$$P(X_0) = (1.0, 0.0)$$

$$T = \begin{array}{cc} \begin{vmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{vmatrix} & \begin{array}{l} S_{n-1} = sun \\ S_{n-1} = rain \end{array} \\ \begin{array}{cc} S_n = sun & S_n = rain \end{array} & \end{array}$$

칼렴 : sunny 한 확률

전날

$$P(X_1) = P(X_0)\,T = (1.0, 0.0) \begin{vmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{vmatrix} = (0.9, 0.1)$$

$$P(X_2) = P(X_1)\,T = (0.9, 0.1) \begin{vmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{vmatrix} = (0.9\times0.9 + 0.1\times0.3\ ,\ 0.9\times0.1+0.1\times0.7)$$

$$= (0.84, 0.16)$$

$$P(X_n) = P(X_{n-1})\,T = P(X_0)\,T^n$$

An example of Markov model

- Influence of initial distribution gets less and less over time

  → the distribution is independent of the initial distribution

- Stationary distribution

  → the distribution of X_t after a sufficiently long time that the distribution of X_t does not change any longer.

$$P_\infty(X) = P_{\infty+1}(X) = \sum P_\infty(X)\, P_{t+1}(X_{t+1} \mid X_t)$$

An example of Markov chain model

Given an observation sequence of states, is it a year of good harvest?

x=RRRSSRSRSSSRSR

R = Rainy, S = Sunny

흉년 모델, 풍년 모델

두 모델 필요

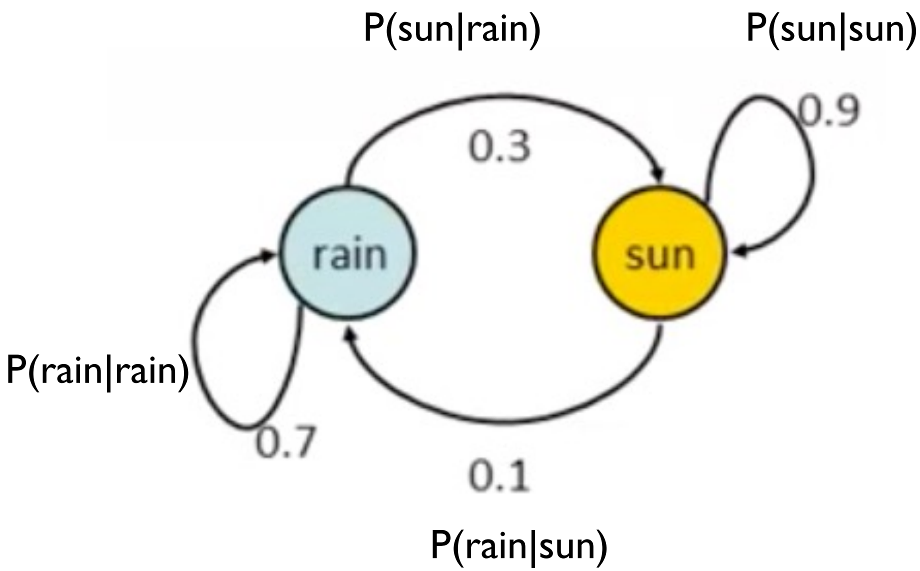
→ Markov Chain

전 영향을 받는 정보를 충분히 받음

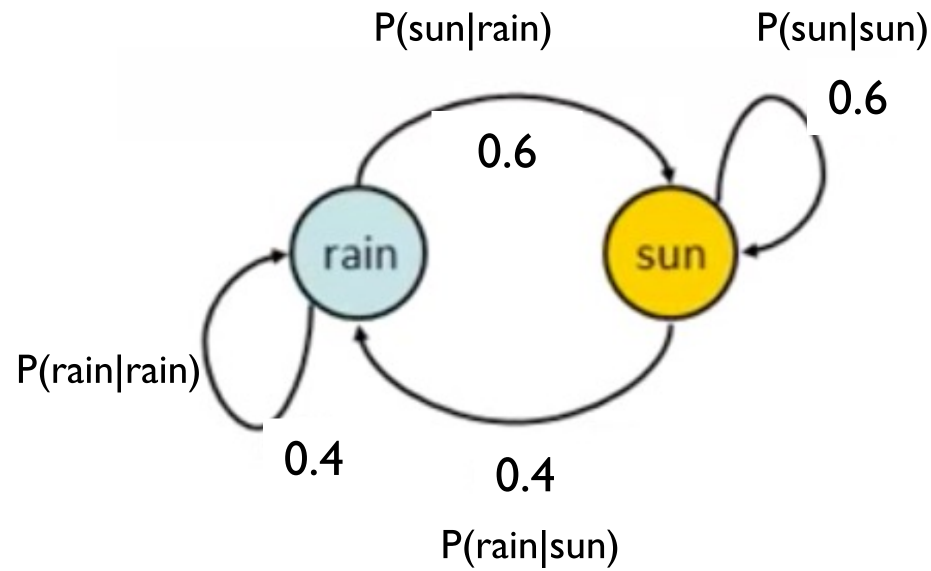
→ Transition Probability 계산

Prediction using Markov models

θ_1 for good harvest



θ_2 for bad harvest



When a sequence of the weather is RRSRRS, do you expect to have good harvest?

$P(RRSRRS|\theta_1) > P(RRSRRS|\theta_2)$?

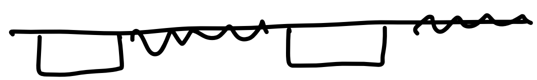
Model
에서의

Prob.

An example of Markov chain model

Given an observation sequence of states, is this gene or non-gene region?

x=CGCCGTACGTGT



→ MC로는 해결X

이게 주어졌을때 유전자인지 아닌지

→ fragmentation 문제

→ Hidden Markov

Markov chain of DNA sequence

- 1st order Markov chain: the current state depends only on the previous state

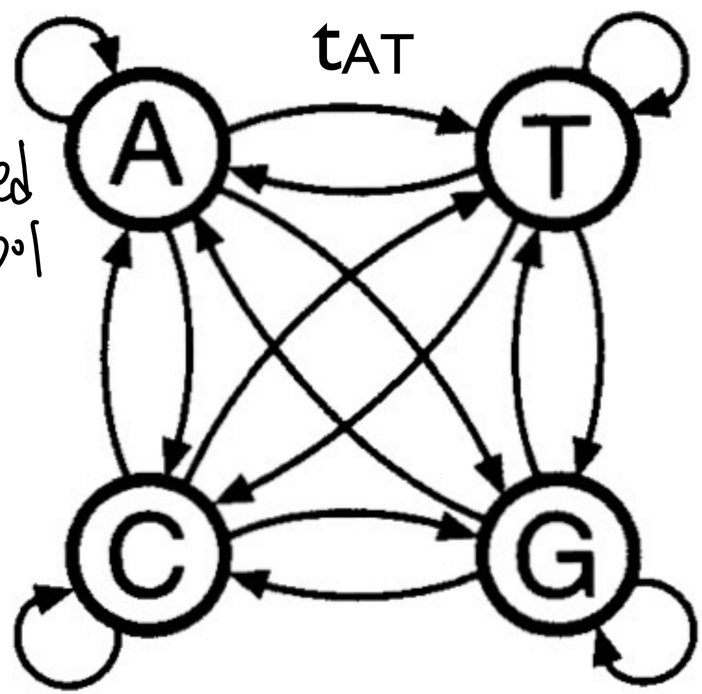
$$P(x_i \mid x_{i-1})$$

Markov chain of DNA sequence

- 1st order Markov chain: the current state depends only on the previous state

$$P(x_i \mid x_{i-1})$$

observed symbol

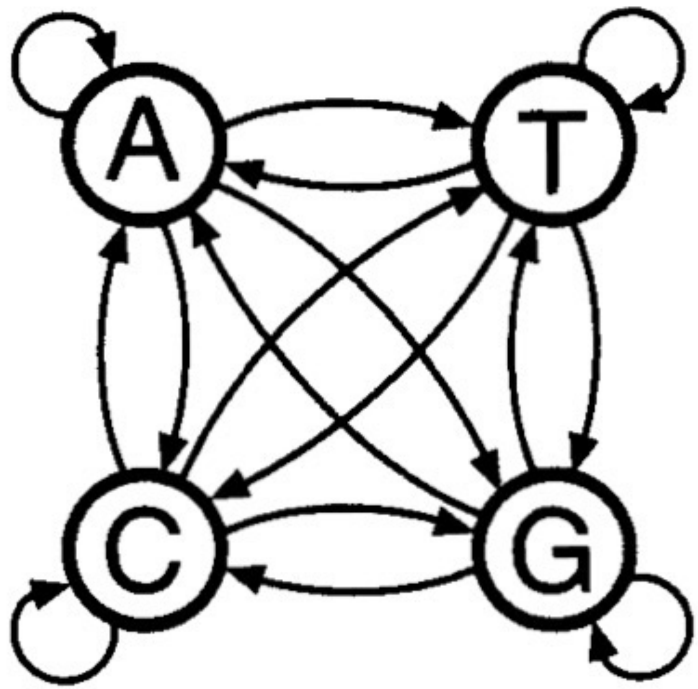


state가 각각의 심볼 나타냄
edge가 transition prob.

모델 파라미터 : running 대상
→ parameter : 4×4개

states: A, C, G, T
transitions: $t_{st} = \Pr(x_i = t \mid x_{i-1} = s)$

Markov chain of DNA sequence



$$\Pr(cggt) = \Pr(c)\Pr(g \mid c)\Pr(g \mid g)\Pr(t \mid g)$$

transition prob.의 곱

Markov chain of DNA sequence

- 2nd order Markov chain for DNA can be treated as a 1st order Markov chain over alphabet

Markov chain of DNA sequence

- 2nd order Markov chain for DNA can be treated as a 1st order Markov chain over alphabet A = {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}

state : $4 \times 4 = 16$가지

parameter : $16 \times 4 = 64$

Markov chain of DNA sequence

- 2nd order Markov chain for DNA can be treated as a 1st order Markov chain over alphabet A = {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}

seq = ACGCTA

Transition prob.

P(ACGCTA) = P(AC)P(G|AC)P(C|CG)P(T|GC)P(A|CT)

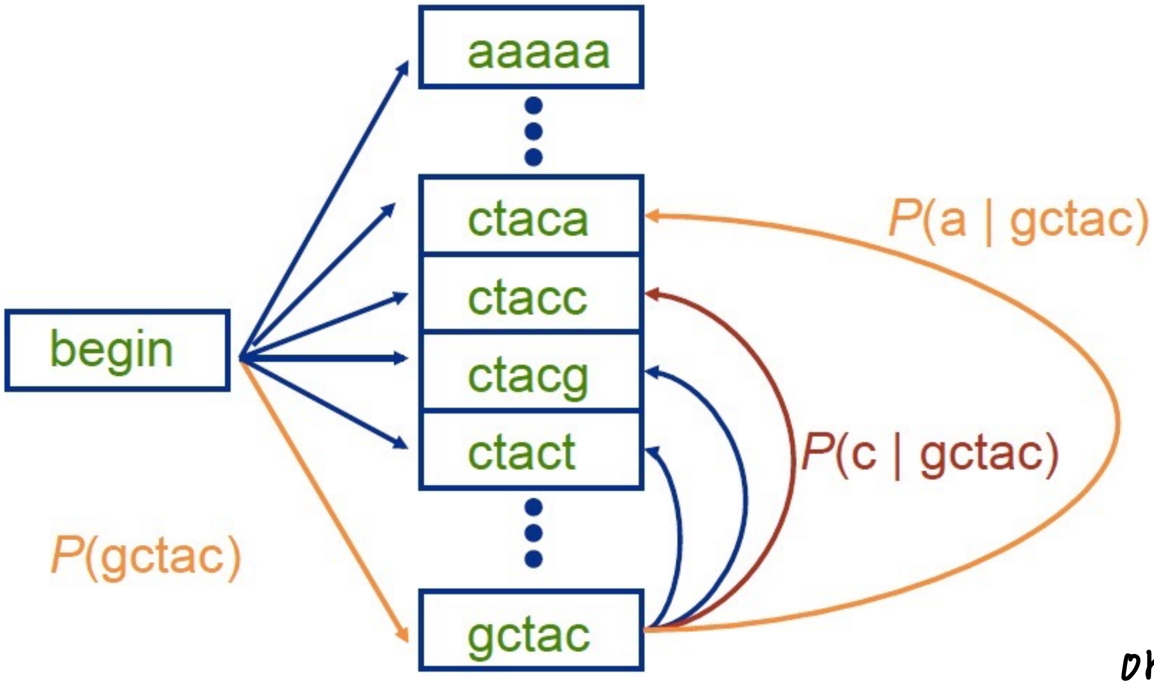
initial

$$\longrightarrow \boxed{AC} \longrightarrow \boxed{CG} \longrightarrow \boxed{GC} \longrightarrow \boxed{CT} \longrightarrow \boxed{TA}$$

주어진 시퀀스 계산 → 유전자 예측

Markov chain of DNA sequence

- 5th order Markov chain



order를 높인다
→ 상세한 state
→ 파라미터수↑

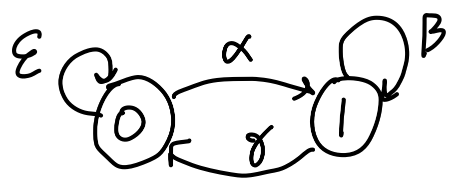
$$P(gctaca) = P(gctac)P(a \mid gctac)$$

한 state에서 나가는 transition probability: 4개

Learning transition probabilities

- Maximum Likelihood Estimators (MLE) for the transition probabilities are the frequencies of the transitions observed in the training data

- For example, a sequence ACGTCGCA in the training data, we can count the number of CG and CX for P(G|C)



$\alpha + \beta \neq 1$

$$a_{st} = \frac{c_{st}}{\sum_{t'} c_{st'}}$$

$P(a|\alpha) + P(b|\alpha) = 1$

(Alpha에서 갈 수 있는 state가 a,b 두 개 라면)

$\alpha + \varepsilon = 1$

C_{st} is the number of times that letter t followed letter s
t' is any symbol in the set

$$t_{01} = \frac{\overset{3}{\#01}}{\#01 + \#00}$$

(ratio Prob.)

3+4

11 000111
01 000111

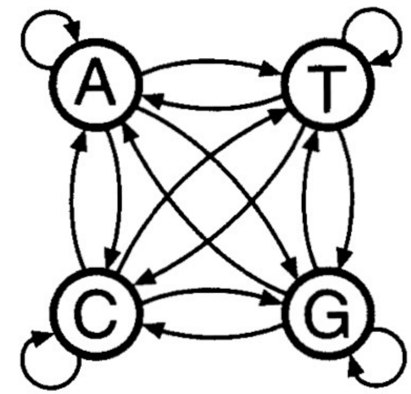
Learning transition probabilities

Sequences for learning

GCCGCGCTTG

GCTTGGTGGC

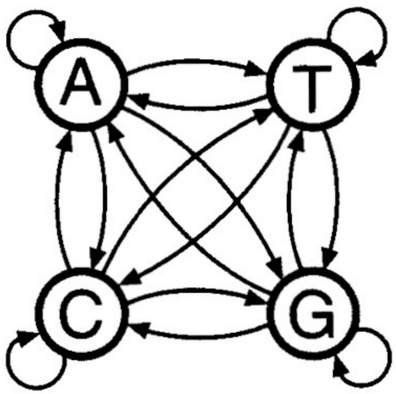
TGGCCGTTGC



P(C|G) = ?

$$\frac{\#GC}{\#GC + \#GA + \#GG + \#GT} = \frac{7}{7+3+2} = \frac{7}{12}$$

learning transition probabilities



Sequences for learning

GCCGCGCTTG

GCTTGGTGGC

TGGCCGTTGC

P(A|G) = 0/12 = 0

P(C|G) = P(GC)/(P(GA)+P(GC)+P(GG)+P(GT) = 7/12 = 0.58

p(G|G) = 3/12 = 0.25

p(T|G) = 2/12 = 0.17

CpG island: an example

- CG dinucleotides are rarer in eukaryotic genomes than expected given the probabilities of C and G
- CG dinucleotides are richer in the upstream of genes
- In CpG island, methylation process is suppressed

CG 영역

Gene 앞 부분에는 CG가 굉장히 많이 나타나는 영역이 있다
-> features가 됨 -> Gene의 signal을 찾기 위한 모델이 됨

CpG island: an example

CATTCCGCCTTCTCTCCCGAGGTGGCGCGTGGGA
GGTGTTTTGCTCGGGTTCTGTAAGAATAGGCCAGG
CAGCTTCCCGCGGGATGCGCTCATCCCCTCTCGG
GGTTCCGCTCCCACCGCGCCCGCGTTCGGCCGGTT
CCGCCTGCGAGATGTTTTCCGACGGACAATGATTC
CACTCTCGGCGCCTCCCATGTTGATCCCAGCTCCT
CTGCGGGCGTCAGGACCCCTGGGCCCCGCCCCG
CTCCACTCAGTCAATCTTTTGTCCCCGTATAAGGCG
GATTATCGGGGTGGCTGGGGGCGGCTGATTCCGA
CGAATGCCCTTGGGGGTCACCCGGGAGGGAACTC
CGGGCTCCGGCTTTGGCCAGCCCGCACCCCTGGT
TGAGCCGGCCCGAGGGCCACCAGGGGGCGCTCG
ATGTTCCTGCAGCCCCCCGCAGCAGCCCCACTCC
CCGGCTCACCCTACGATTGGCTGGCCGCCCCGAG
CTCTGTGCTGTGATTGGTCACAGCCCGTGTCCGTC
GCGGGCGCCGGGGCGGATACGAGGTGACGCGCA
GAGGCCCAGCTCGGGGCGGTGTCCCGCGCCGGC
GACTGCGGGCGGAGTTTCGCGAGGGCCGAAGCG
GGGCAGTGTGACGGCAGCGGTCCTGGGAGGCGC
CCGCGCGCGTCGGAGCAGCTCCCCGTCCTCCGCA
GCCGTCACCGCCCGGCCGTCGCCGCGCCCTGGCC

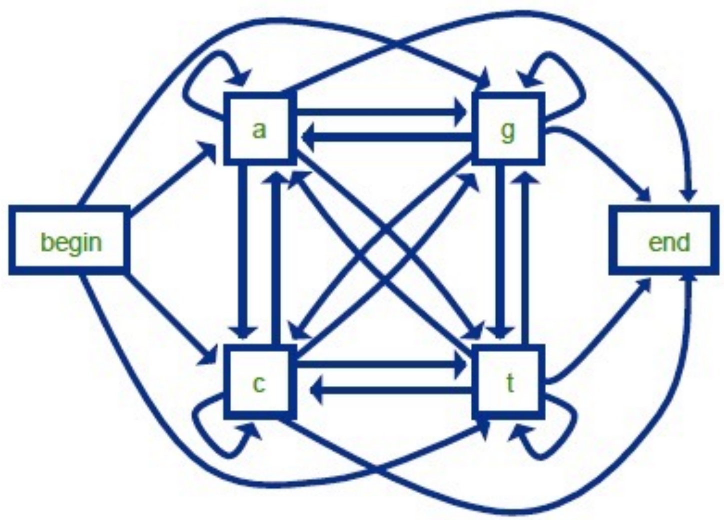
유전자

CTCTTAGTTTTGGGTGCATTTGTCTGGTCTTCCAAA
CTAGATTGAAAGCTCTGAAAAAAAAAACTATCTTGT
GTTTCTATCTGTTGAGCTCATAGTAGGTATCCAGGA
AGTAGTAGGGTTGACTGCATTGATTTGGGACTACAC
TGGGAGTTTTCTTCGCCATCTCCCTTTAGTTTTCCT
TTTTTTCTTTCTTTCTTTTCTTTTTTTTTCTTTTTTTTT
TTGAGATGTCGTCTTGCTCAGTCCCCCAGGCTGGA
GTGCAGTGGTGCGATCTTGGCTCACTGTAGCCTCC
ACCTCCCAGGTTCAAGCAATTCTACTGCCTTAGCCT
CCCGAGTAGCTGGGATTACAAGCACCCGCCACCAT
TCCTGGCTAATTTTTTTTTTTTGTATTTTTAGTTGAGA
CAGGGTTTCACCATGTTGGTGATGCTGGTCTCAGA
CTCCTGGGGCCTAGCGATCCCCCTGCCTCAGCCT
CCCAGAGTGTTAGGATTACAGGCATGAGCCACTGT
ACCCGGCCTCTCTCCAGTTTCCAGTTGGAATCCAA
GGGAAGTAAGTTTAAGATAAAGTTACGATTTTGAAAT
CTTTGGATTCAGAAGAATTTGTCACCTTTAACACCT
AGAGTTGAACGTTCATACCTGGAGAGCCTTAACATT
AAGCCCTAGCCAGCCTCCAGCAAGTGGACATTGGT
CAGGTTTGGCAGGATTCGTCCCCTGAAGTGGACT
GAGAGCCACACCCTGGCCTGTCACCATACCCATCC

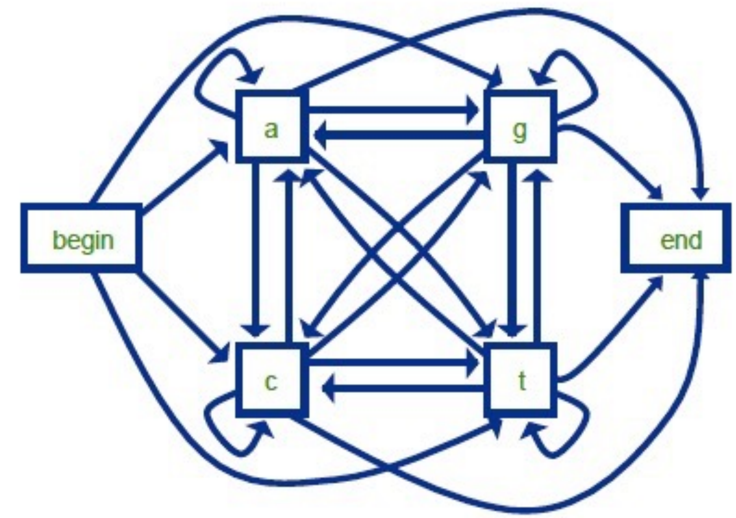
CpG island: an example

An example of 1st order Markov chain

CpG island



Random genomic region (null)



+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

P(AG) = P(G|A)
A: given

−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322 +	0.298 +	0.078	+0.302 = 1
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

condition (given)

CpG island: an example

x=CGCCGTACGTGT

$$P(CpG \mid x) = \frac{P(x \mid CpG)P(CpG)}{P(x)}$$ (Bayesian rule)

x가 주어졌을때 모델의 확률

특정 모델에서의 x의 Prob.

$$= \frac{P(x \mid CpG)P(CpG)}{P(x \mid CpG)P(CpG) + P(x \mid null)P(null)}$$ Normalization

= 원하는 특정 모델에서의 x probability / x가 가능한 모든 모델에서의 x probability의 합

we need P(x|CpG) and P(x|null), assuming P(CpG) = P(null)

$$S(x) = \log \frac{P(x \mid \text{model}^+)}{P(x \mid \text{model}^-)}$$

For example, if S(x) > 0, x is in the CpG island