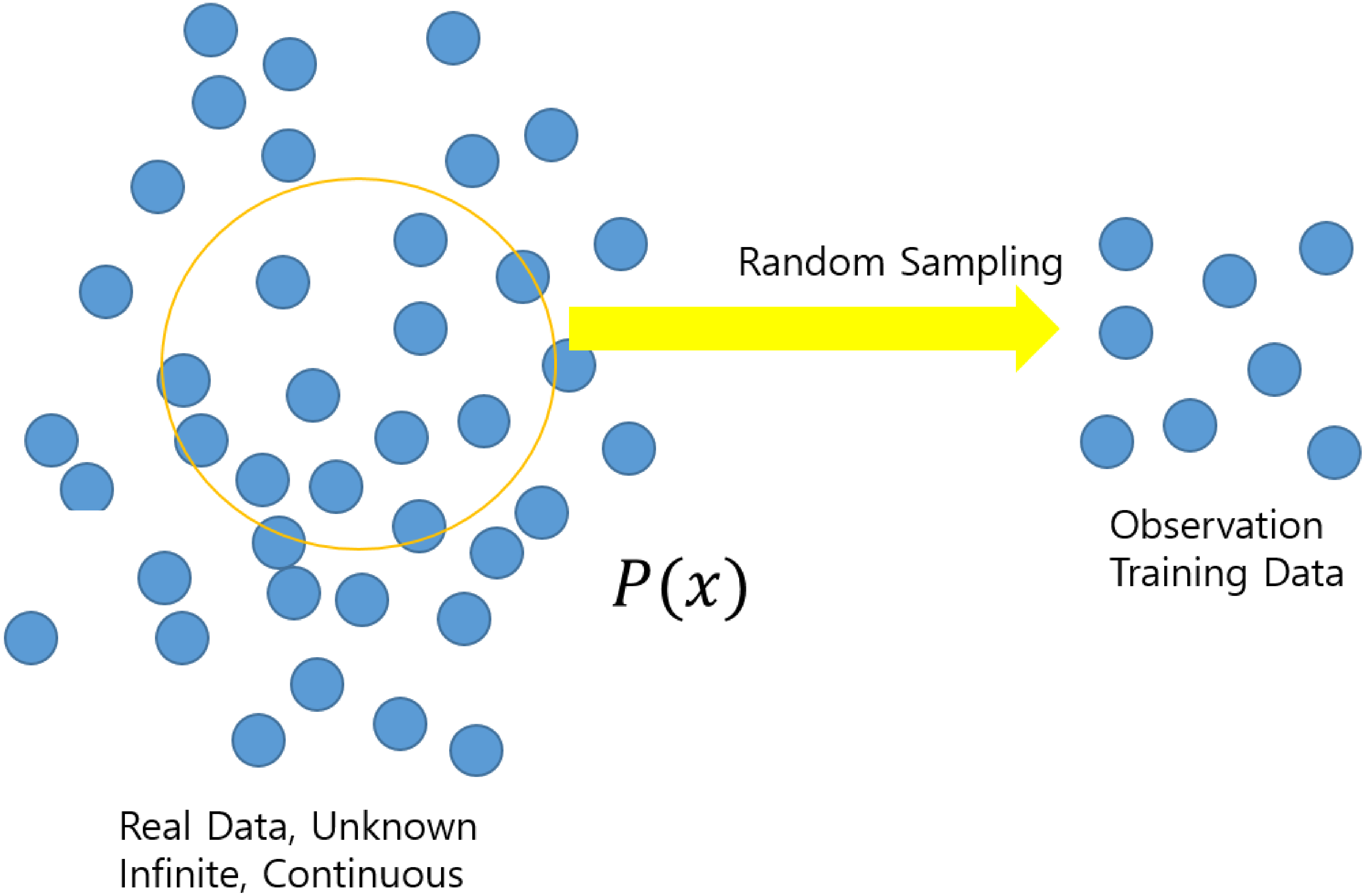
Structured probabilistic Graphical model for deep learning

MMI DEEPLEARNING SEMINA

2019.1.24 | 구범혁



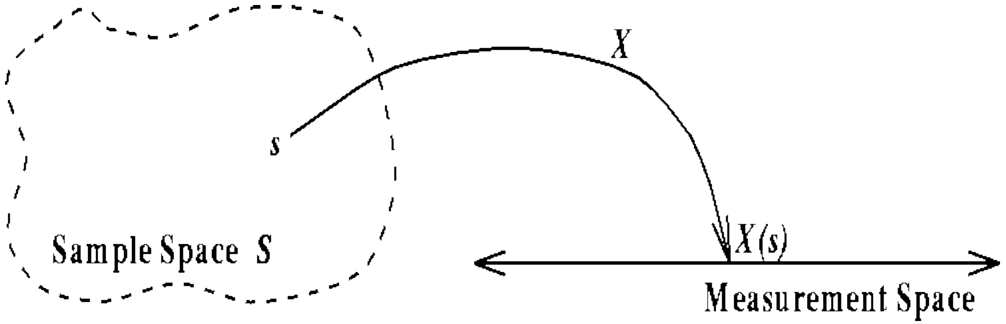
Random Sampling
$P(x)$
Observation
Training Data
Real Data, Unknown
Infinite, Continuous

$P(x)$: 데이터 자체의 분포

$P(x|c)$: $Class$에 속해 있는 데이터의 분포$(Likelihood)$

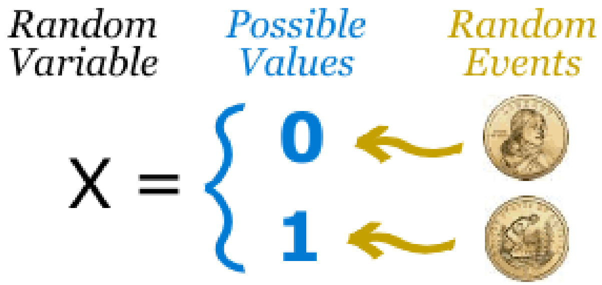
$P(c|x)$: 데이터가 들어왔을 때 $Class$에 $mapping$ 되는 확률 분포

$P(c)$: 어떤 $Class$가 어떻게 있을지를 결정하는 사전 지식



이러한 확률 분포들을 유명한 분포로 가정하고,
확률 분포의 파라미터를 구하는 것이 목적

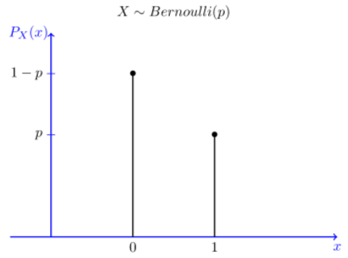
다른 해석 “함수라고 생각하기”
함수의 x값(Random variable)이 있고, 함수값 f(x)는 확률



Q1. 동전의 앞면 또는 뒷면이 나올 확률은?

음.. 일단 모르니까 확률 분포를 먼저 생각하면, 동전 던지기는 1회
시행이니까 Bernoulli distribution으로 정의하자.
그리고 확률의 합은 1이니까 앞면일 경우의 확률을 파라미터로 정의하자.

$$p(x|\mu) = \mu^x(1-\mu)^{1-x}$$



$$p(x|\mu) = \mu^x (1-\mu)^{1-x}$$

이러한 확률 분포들을 유명한 분포로 가정하고,
확률 분포의 파라미터를 구하는 것이 목적

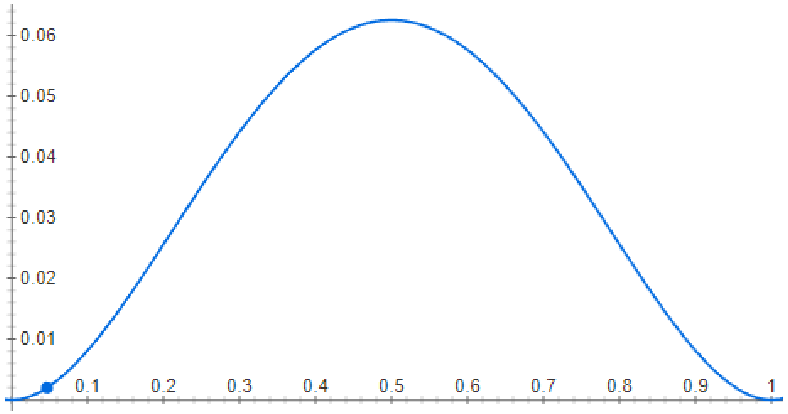
베르누이 분포의 파라미터 μ를 구하자

관측된 데이터
앞면 3번, 뒷면 3번

직관적으로 생각하면 $\mu = 0.5$
하지만 수학적으로 풀어봐야 논리적이고 납득이된다.
질문을 약간 틀어서..

“μ가 얼마일 때 관측된 데이터가 가장 잘 설명될까?
즉 $P(data|\mu)$가 어떨때 가장 큰 값을 나타낼까?”

$$p(data|\mu) = \mu^3 (1-\mu)^3$$



$$p(data|\mu) = \mu^3(1-\mu)^3$$

"μ가 얼마일 때 관측된 데이터가 가장 잘 설명될까?
즉 $P(data|\mu)$가 어떨때 가장 큰 값을 나타낼까?"

Likelihood

MLE(Maximum Likelihood Estimation)

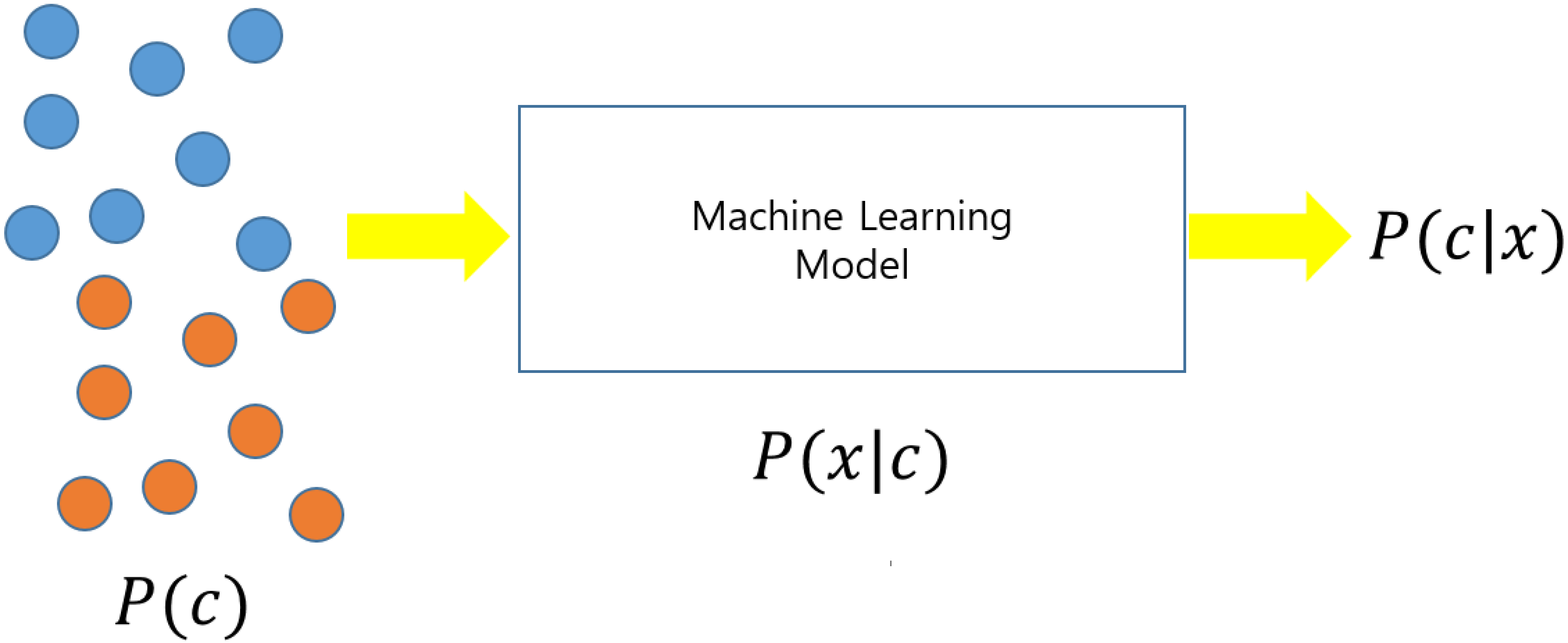
$$\mu = argmax_\mu P_{Bernoulli}(Observation|\mu)$$

$$Likelihood = P(x_1, x_2, \dots, x_n|\mu) = \prod_{n=1}^{N} P(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$Log - Likelihood = \log(\mu)\sum x_n + \log(1-\mu)\sum(1-x_n)$$

$$\mu = \frac{1}{N}\sum x_n$$

"Sample에서 전체 중 앞면이 나온 횟수"



Machine Learning
Model
$P(c|x)$
$P(x|c)$
$P(c)$

Graphical model

From Wikipedia, the free encyclopedia

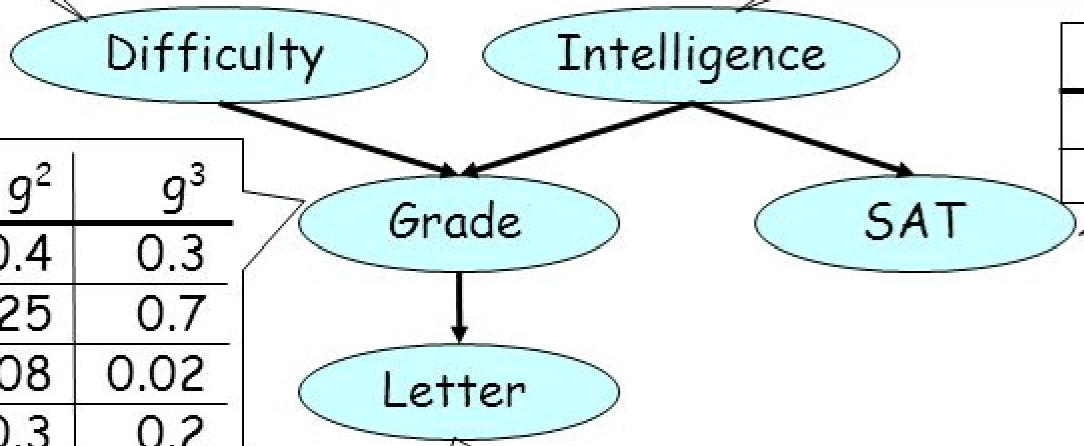
> This article includes a list of references, but **its sources remain unclear** because it has **insufficient inline citations**. Please help to improve this article by introducing more precise citations. *(May 2017)* *(Learn how and when to remove this template message)*

A **graphical model** or **probabilistic graphical model** (**PGM**) or **structured probabilistic model** is a probabilistic model for which a graph expresses the conditional dependence structure between random variables. They are commonly used in probability theory, statistics—particularly Bayesian statistics—and machine learning.

The Student Network

d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

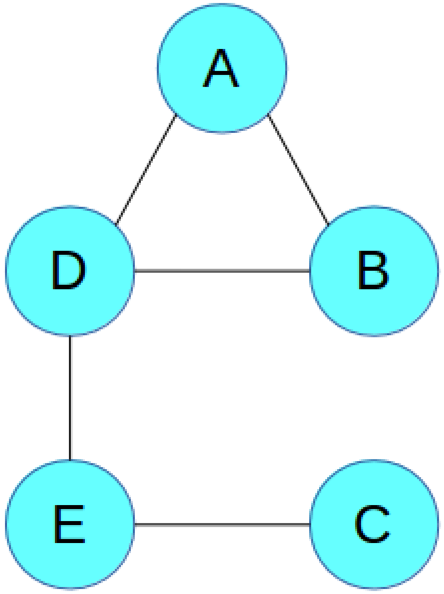


	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

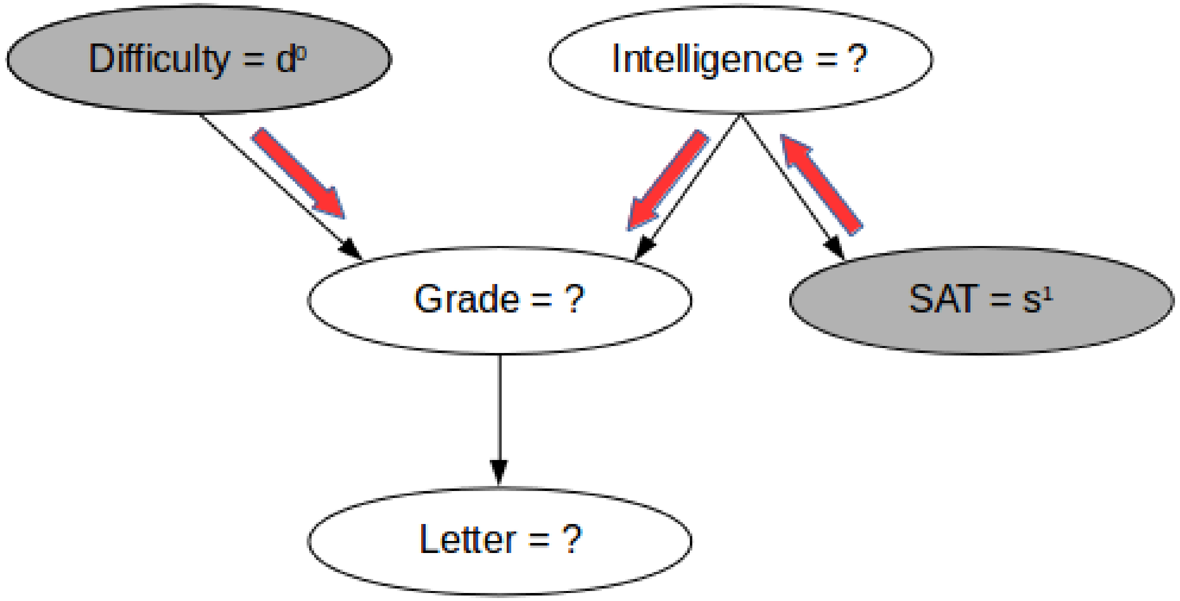
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Daphne Koller

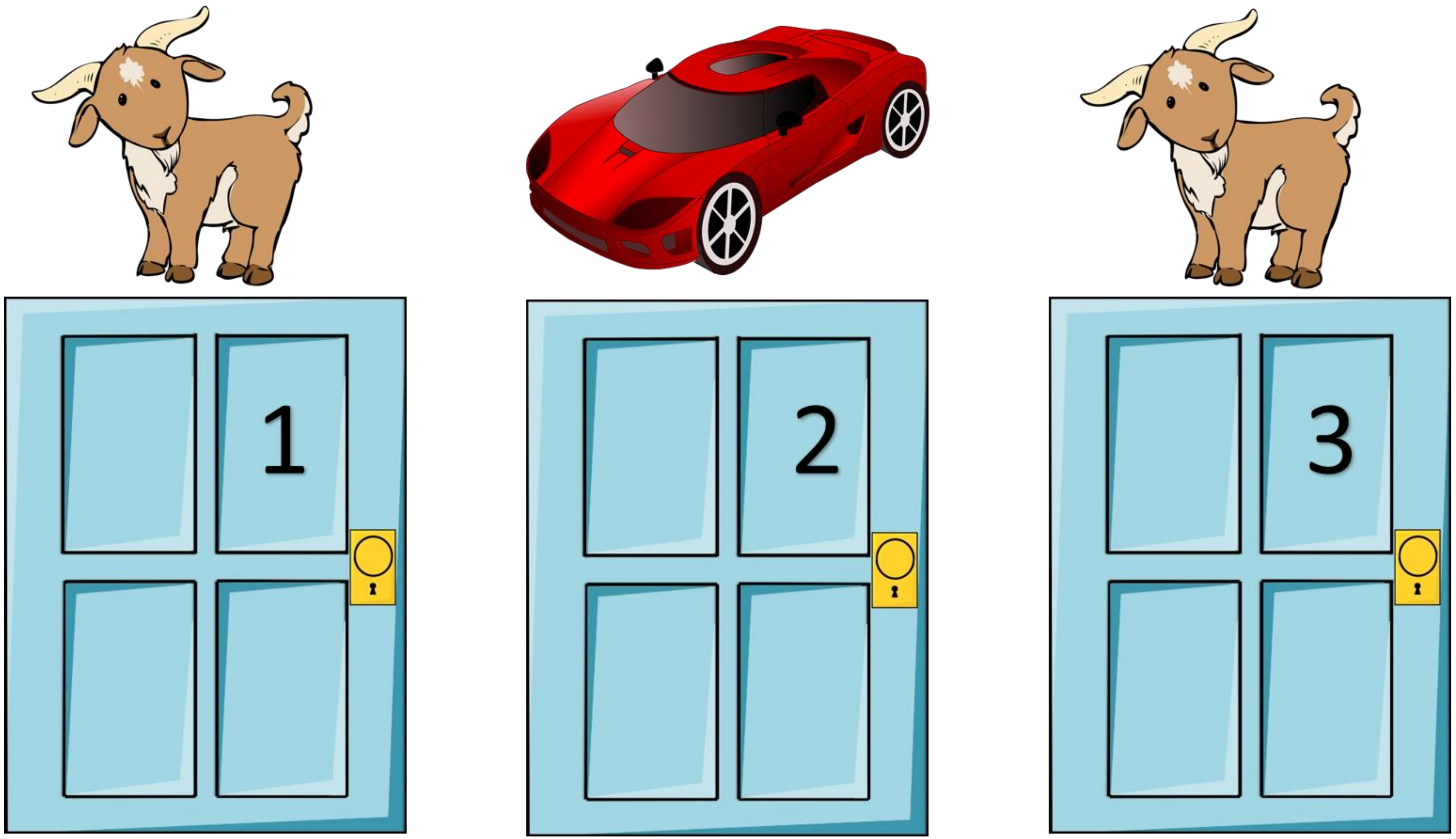


A	B	C	ϕ(A, B, C)
0	0	0	10
0	0	1	1
0	1	0	1
0	1	1	10
1	0	0	1
1	0	1	10
1	1	0	10
1	1	1	1



1. 학생이 똑똑하다는 것을 알면(Intelligence is observed), SAT 점수가 낮아도 좋은 Grade를 기대할 수 있다. 왜냐하면 똑똑하다는 사실을 알기 때문이다. => If intelligence is observed, then SAT and Grade are independent.

2. 학생이 똑똑하다는 것을 알아도 difficulty는 알 수 없다. 반면에 학생이 bad grade인 경우(grade is observed), Cource는 어려웠고(difficult) 따라서 똑똑한 학생은 나쁜 Grade를 받았음을 알 수 있다. => If Grade is unobserved, then intelligence and difficulty are independent.

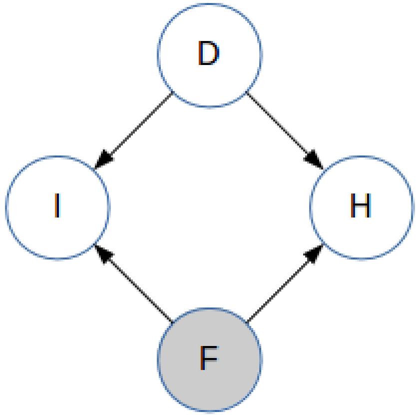

1
2
3

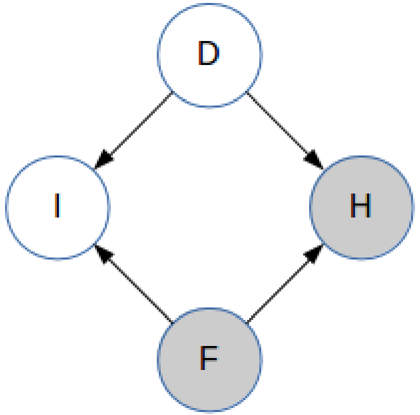
D : The door with the car.

F : Your first choice

H : The door opened by the host.

I : Is F = D?





“H is observed”

p(D)

1	2	3
1/3	1/3	1/3

p(F)

1	2	3
1/3	1/3	1/3

$$p(I|F=1) = \frac{p(I, F=1)}{p(F=1)} = \frac{\sum_D p(I|F=1, D)p(D)}{p(F=1)}$$

$$p(D|F=1) = \frac{p(D, F=1)}{p(F=1)} = \frac{p(D)p(F=1)}{p(F=1)}$$

p(I | D, F)

	0	1
D=1, F=1	0	1
D=1, F=2	1	0
D=1, F=3	1	0
D=2, F=1	1	0
D=2, F=2	0	1
D=2, F=3	1	0
D=3, F=1	1	0
D=3, F=2	1	0
D=3, F=3	0	1

p(H | D, F)

	1	2	3
D=1, F=1	0	1/2	1/2
D=1, F=2	0	0	1
D=1, F=3	0	1	0
D=2, F=1	0	0	1
D=2, F=2	1/2	0	1/2
D=2, F=3	1	0	0
D=3, F=1	0	1	0
D=3, F=2	1	0	0
D=3, F=3	1/2	1/2	0

p(I | F=1)

0	1
2/3	1/3

p(D | F=1)

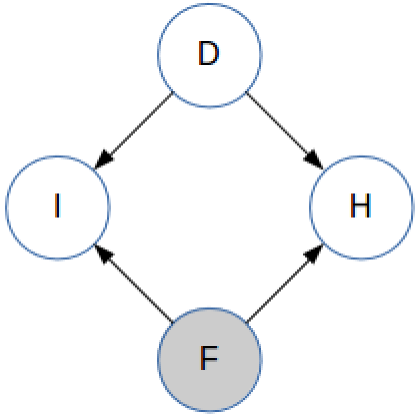
1	2	3
1/3	1/3	1/3

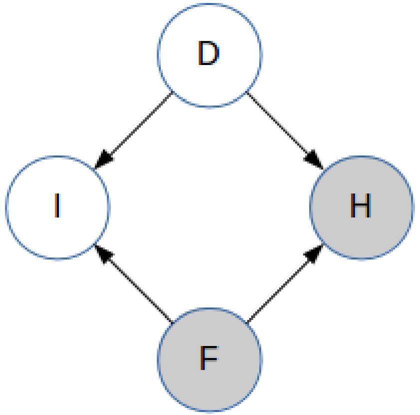
D : The door with the car.

F : Your first choice

H : The door opened by the host.

I : Is F = D?





"H is observed"

p(D)

1	2	3
1/3	1/3	1/3

p(F)

1	2	3
1/3	1/3	1/3

p(I | D, F)

	0	1
D=1, F=1	0	1
D=1, F=2	1	0
D=1, F=3	1	0
D=2, F=1	1	0
D=2, F=2	0	1
D=2, F=3	1	0
D=3, F=1	1	0
D=3, F=2	1	0
D=3, F=3	0	1

p(H | D, F)

	1	2	3
D=1, F=1	0	1/2	1/2
D=1, F=2	0	0	1
D=1, F=3	0	1	0
D=2, F=1	0	0	1
D=2, F=2	1/2	0	1/2
D=2, F=3	1	0	0
D=3, F=1	0	1	0
D=3, F=2	1	0	0
D=3, F=3	1/2	1/2	0

$$p(D|F=1,H=2) = \frac{p(D,F=1,H=2)}{p(F=1,H=2)}$$

$$= \frac{p(D)p(F=1)p(H=2|D,F=1)}{\sum_D p(D)p(F=1)p(H=2|D,F=1)}$$

$$p(D|F=1,H=2) = \frac{p(D,F=1,H=2)}{p(F=1,H=2)}$$

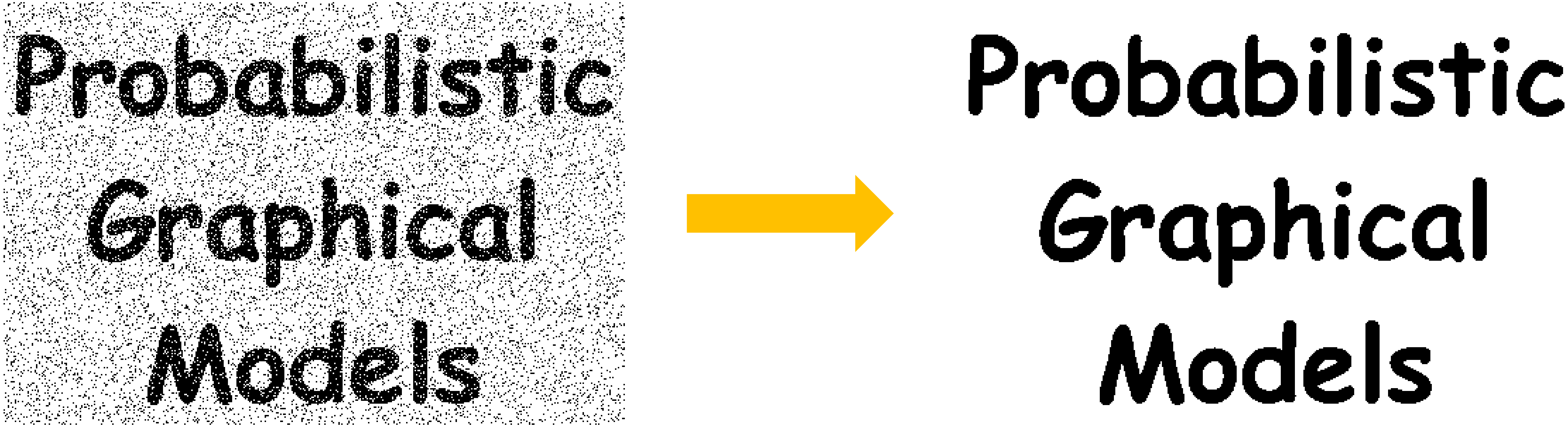
$$= \frac{p(D)p(F=1)p(H=2|D,F=1)}{\sum_D p(D)p(F=1)p(H=2|D,F=1)}$$

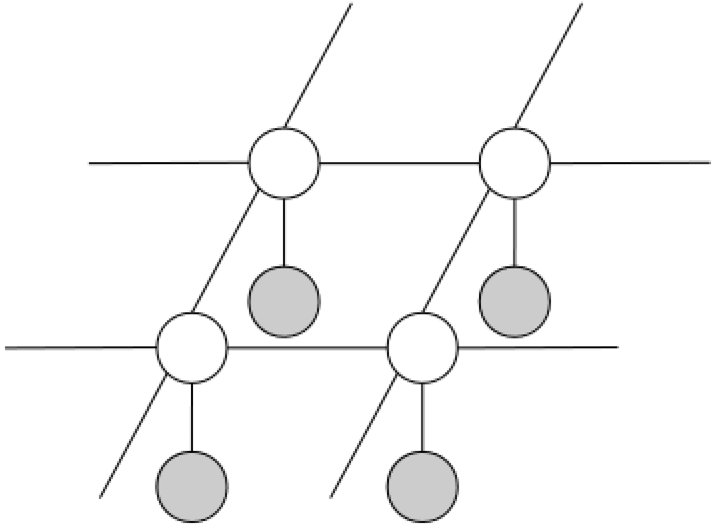
p(I | F=1, H=2)

0	1
2/3	1/3

p(D | F=1, H=2)

1	2	3
1/3	0	2/3





$$
\begin{aligned}
Y^* &= \arg\max_Y P(Y|X) \\
&= \arg\max_Y \log P(Y|X) \\
&= \arg\max_Y [\log P(X,Y) - \log P(X)] \\
&= \arg\max_Y \log P(X,Y)
\end{aligned}
$$

$$p(X,Y) = \frac{1}{Z} \prod_{ij} \phi(X_{ij}, Y_{ij}) \prod_{(ij,kl)} \phi(Y_{ij}, Y_{kl})$$

$$Y^* = \arg\max_Y \log p(X,Y) = \arg\max_Y \sum_{ij} w_e X_{ij} Y_{ij} + \sum_{(ij,kl)} w_s Y_{ij} Y_{kl}$$