Mathematics for Artificial Intelligence

6강: 확률론 맛보기

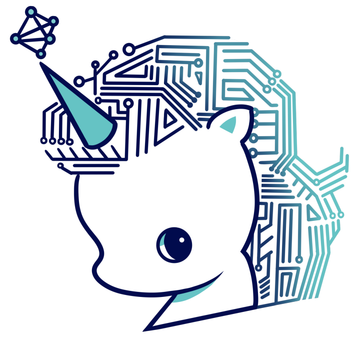
임성빈 UNIST 인공지능대학원 & 산업공학과
Learning Intelligent Machine Lab

UNIST LIM Lab
Learning Intelligent Machine

× boostcamp AI Tech

딥러닝에서 확률론이 왜 필요한가요?

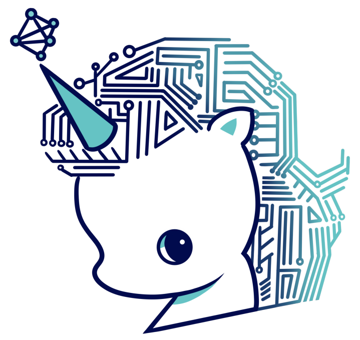
- 딥러닝은 확률론 기반의 기계학습 이론에 바탕을 두고 있습니다
- 기계학습에서 사용되는 손실함수(loss function)들의 작동 원리는 데이터 공간을 통계적으로 해석해서 유도하게 됩니다



예측이 틀릴 위험(risk)을 최소화하도록 데이터를 학습하는 원리는 통계적 기계학습의 기본 원리이다

딥러닝에서 확률론이 왜 필요한가요?

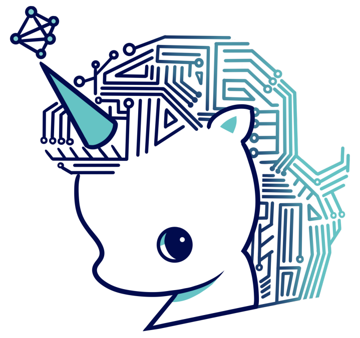
- 딥러닝은 확률론 기반의 기계학습 이론에 바탕을 두고 있습니다
- 기계학습에서 사용되는 손실함수(loss function)들의 작동 원리는 데이터 공간을 통계적으로 해석해서 유도하게 됩니다
- 회귀 분석에서 손실함수로 사용되는 L_2-노름은 예측오차의 분산을 가장 최소화하는 방향으로 학습하도록 유도합니다
- 분류 문제에서 사용되는 교차엔트로피(cross-entropy)는 모델 예측의 불확실성을 최소화하는 방향으로 학습하도록 유도합니다



교차 엔트로피는 다음 강의에서 소개할 예정입니다

딥러닝에서 확률론이 왜 필요한가요?

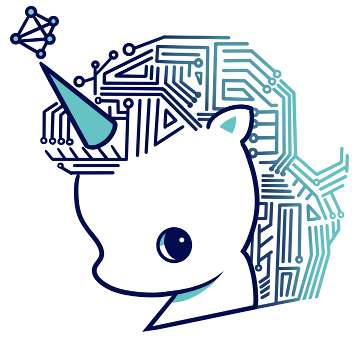
- 딥러닝은 확률론 기반의 기계학습 이론에 바탕을 두고 있습니다
- 기계학습에서 사용되는 손실함수(loss function)들의 작동 원리는 데이터 공간을 통계적으로 해석해서 유도하게 됩니다
- 회귀 분석에서 손실함수로 사용되는 L_2-노름은 예측오차의 분산을 가장 최소화하는 방향으로 학습하도록 유도합니다
- 분류 문제에서 사용되는 교차엔트로피(cross-entropy)는 모델 예측의 불확실성을 최소화하는 방향으로 학습하도록 유도합니다
- 분산 및 불확실성을 최소화하기 위해서는 측정하는 방법을 알아야 합니다



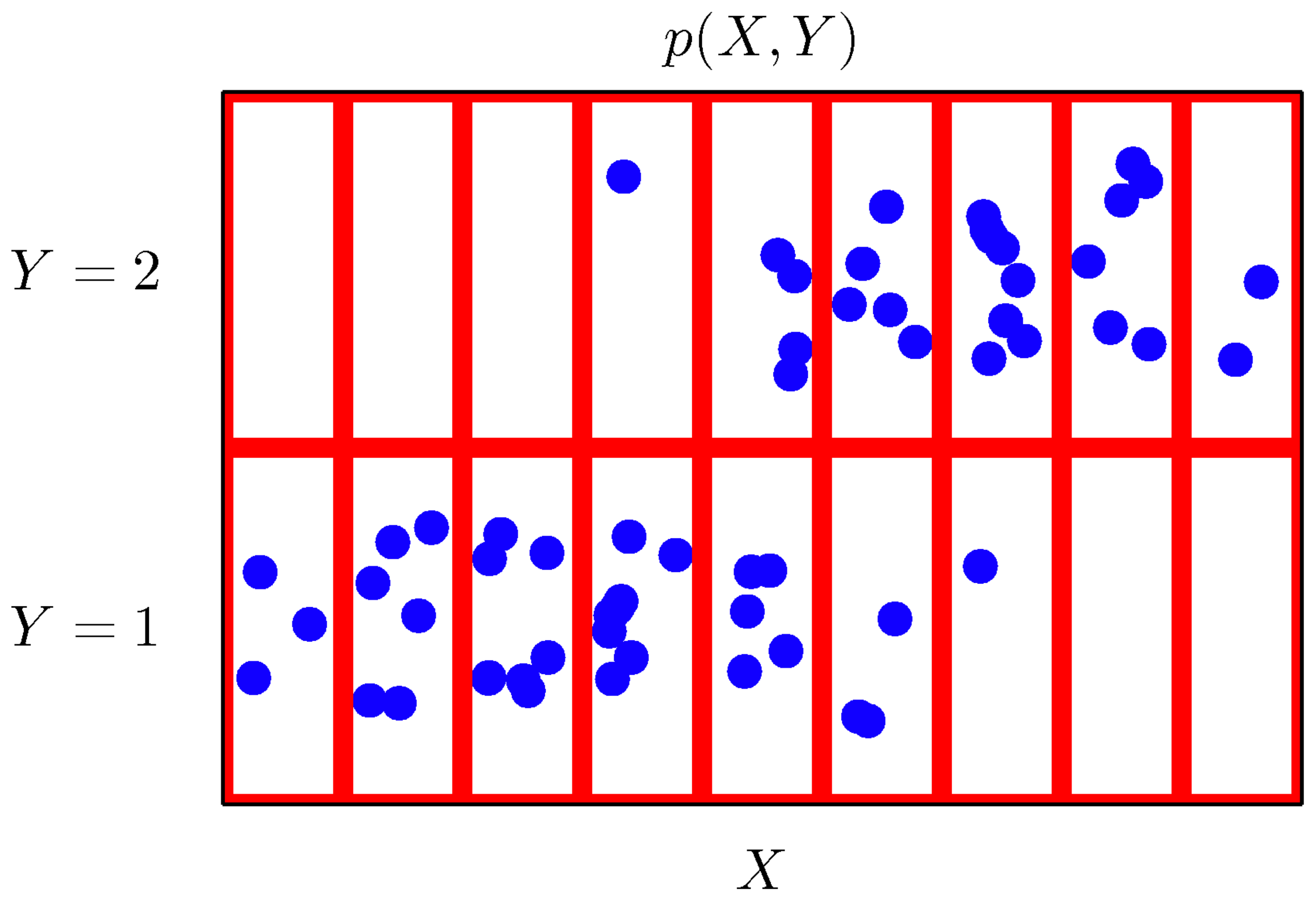
두 대상을 측정하는 방법을 통계학에서 제공하기 때문에
기계학습을 이해하려면 확률론의 기본 개념을 알아야 합니다

확률분포는 데이터의 초상화

- 데이터공간을 $\mathscr{X} \times \mathscr{Y}$ 라 표기하고 $\mathscr{D}$ 는 데이터공간에서 데이터를 추출하는 분포입니다



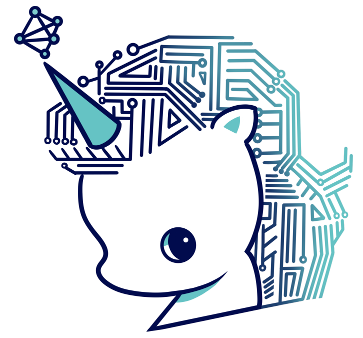
이 수업에선 데이터가 정답 레이블을 항상 가진 지도학습을 상정합니다



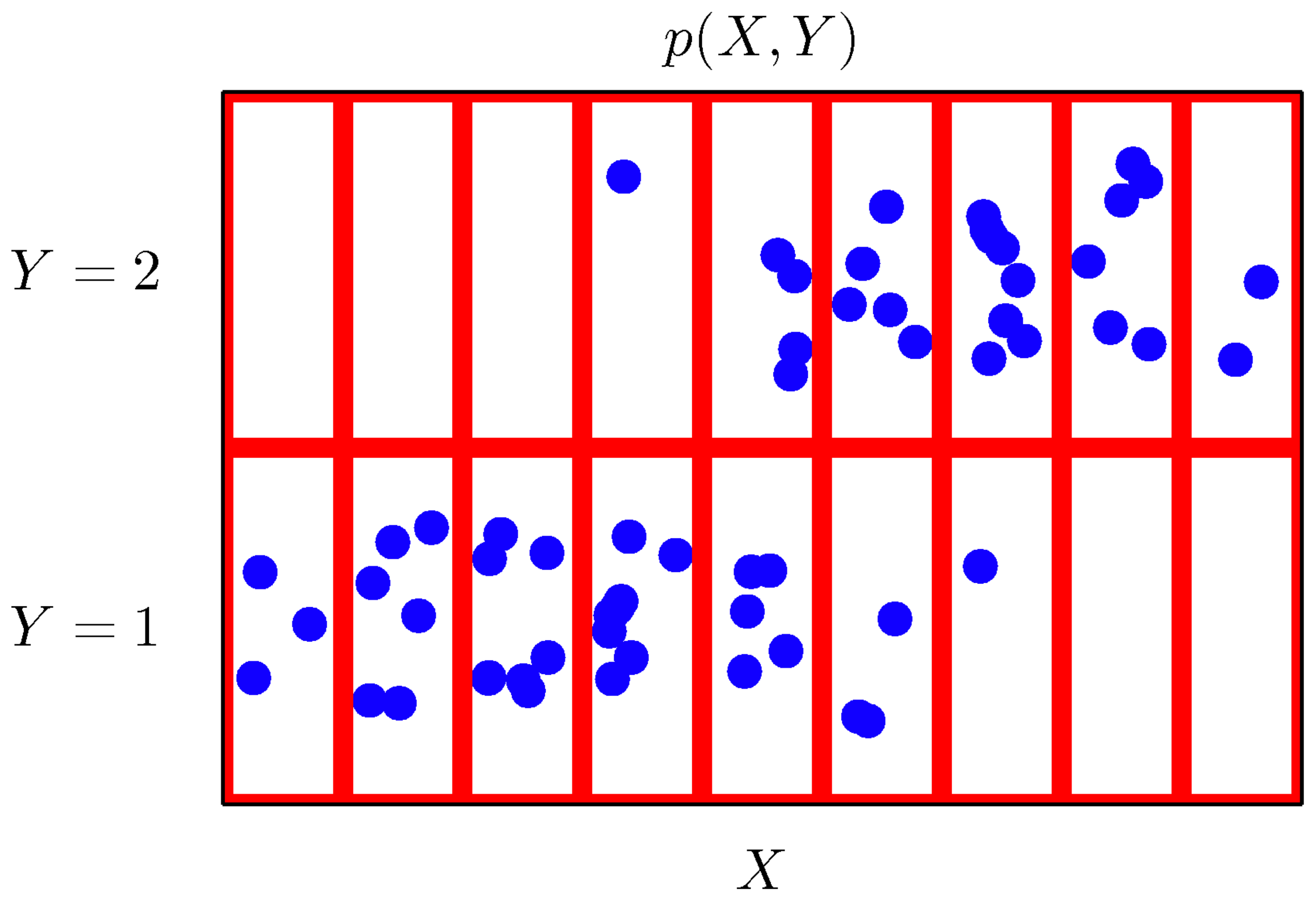
(출처: *Pattern Recognition and Machine Learning, Bishop*)

확률분포는 데이터의 초상화

- 데이터공간을 $\mathscr{X} \times \mathscr{Y}$ 라 표기하고 $\mathscr{D}$ 는 데이터공간에서 데이터를 추출하는 분포입니다
- 데이터는 확률변수로 $(\mathbf{x}, y) \sim \mathscr{D}$ 라 표기



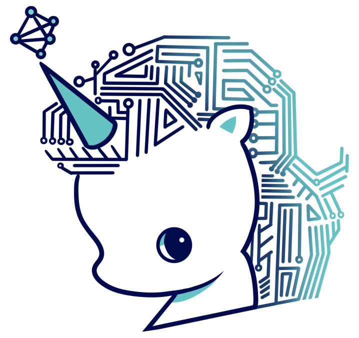
$(\mathbf{x}, y) \in \mathscr{X} \times \mathscr{Y}$ 는 데이터공간 상의 **관측가능한 데이터**에 해당합니다



(출처: *Pattern Recognition and Machine Learning, Bishop*)

이산확률변수 vs 연속확률변수

- 확률변수는 확률분포 $\mathscr{D}$ 에 따라 이산형(discrete)과 연속형(continuous) 확률변수로 구분하게 됩니다

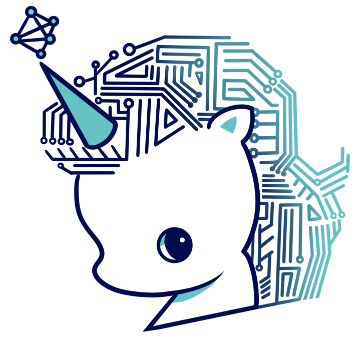


데이터공간 $\mathscr{X} \times \mathscr{Y}$ 에 의해 결정되는 것으로 오해를 하지만 $\mathscr{D}$ 에 의해 결정된다

이산확률변수 vs 연속확률변수

- 확률변수는 확률분포 $\mathscr{D}$ 에 따라 이산형(discrete)과 연속형(continuous) 확률변수로 구분하게 됩니다
- 이산형 확률변수는 확률변수가 가질 수 있는 경우의 수를 모두 고려하여 확률을 더해서 모델링한다

$$\mathbb{P}(X \in A) = \sum_{\mathbf{x} \in A} P(X = \mathbf{x})$$



$P(X = \mathbf{x})$ 는 확률변수가 $\mathbf{x}$ 값을 가질 확률로 해석할 수 있다

이산확률변수 vs 연속확률변수

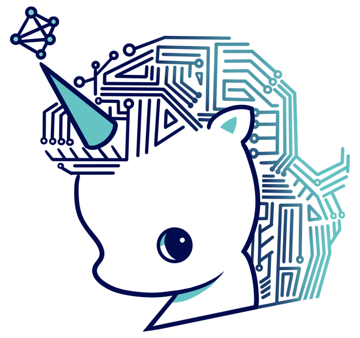
- 확률변수는 확률분포 $\mathscr{D}$ 에 따라 이산형(discrete)과 연속형(continuous) 확률변수로 구분하게 됩니다
- 이산형 확률변수는 확률변수가 가질 수 있는 경우의 수를 모두 고려하여 확률을 더해서 모델링한다

$$\mathbb{P}(X \in A) = \sum_{\mathbf{x} \in A} P(X = \mathbf{x})$$

- 연속형 확률변수는 데이터 공간에 정의된 확률변수의 밀도(density) 위에서의 적분을 통해 모델링한다

$$P(\mathbf{x}) = \lim_{h \to 0} \frac{\mathbb{P}(\mathbf{x} - h \le X \le \mathbf{x} + h)}{2h}$$

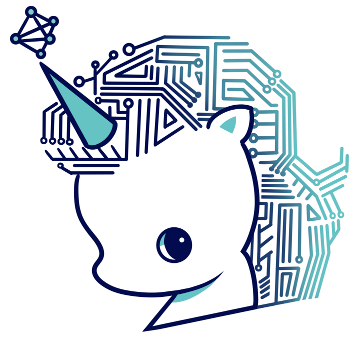
$$\mathbb{P}(X \in A) = \int_A P(\mathbf{x}) d\mathbf{x}$$



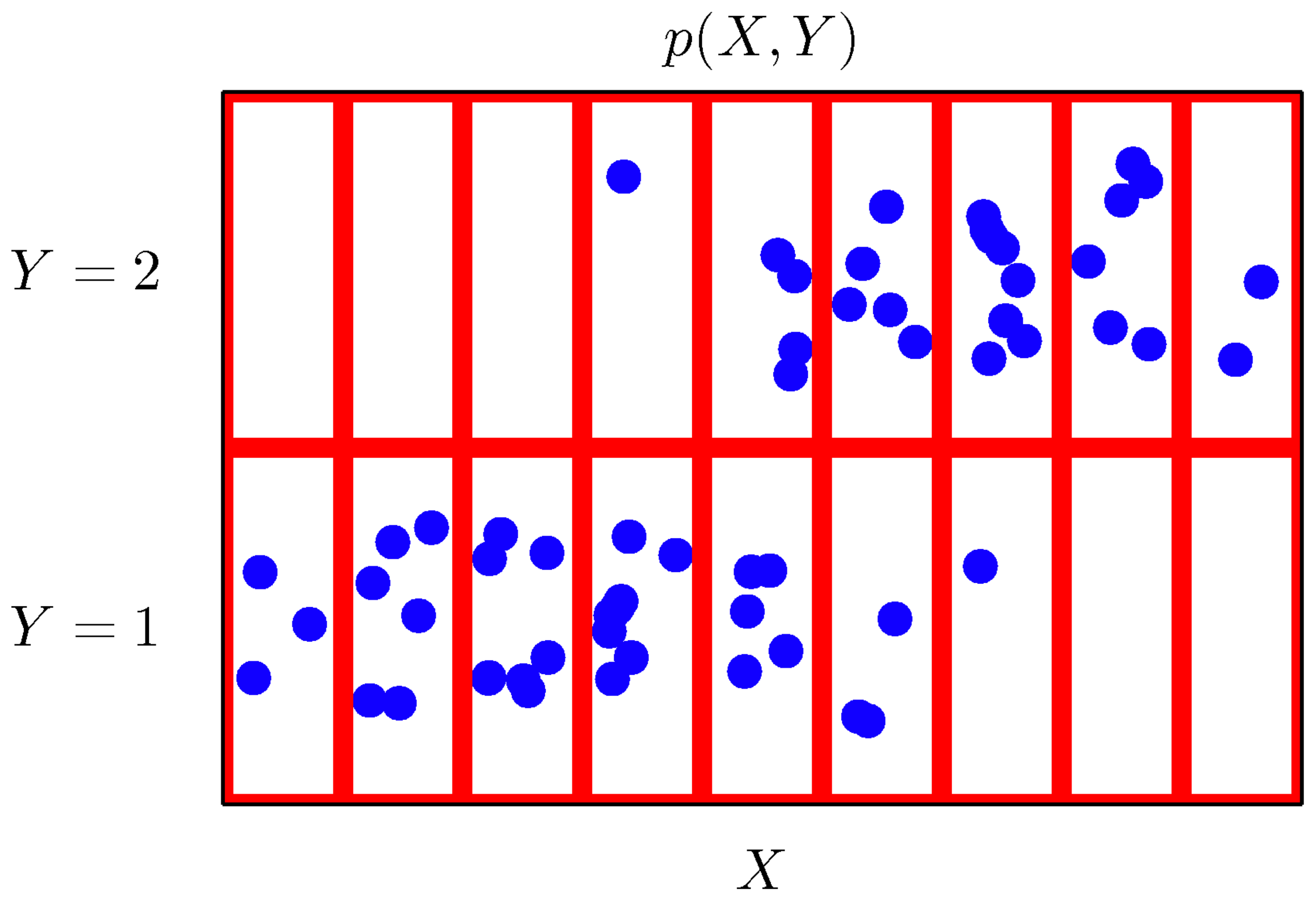
밀도는 누적확률분포의 변화율을 모델링하며 확률로 해석하면 안된다

확률분포는 데이터의 초상화

- 데이터공간을 $\mathscr{X} \times \mathscr{Y}$ 라 표기하고 $\mathscr{D}$ 는 데이터공간에서 데이터를 추출하는 분포입니다
- 데이터는 확률변수로 $(\mathbf{x}, y) \sim \mathscr{D}$ 라 표기
- 결합분포 $P(\mathbf{x}, y)$ 는 $\mathscr{D}$ 를 모델링합니다



$\mathscr{D}$ 는 이론적으로 존재하는 확률분포이기 때문에 사전에 알 수 없습니다

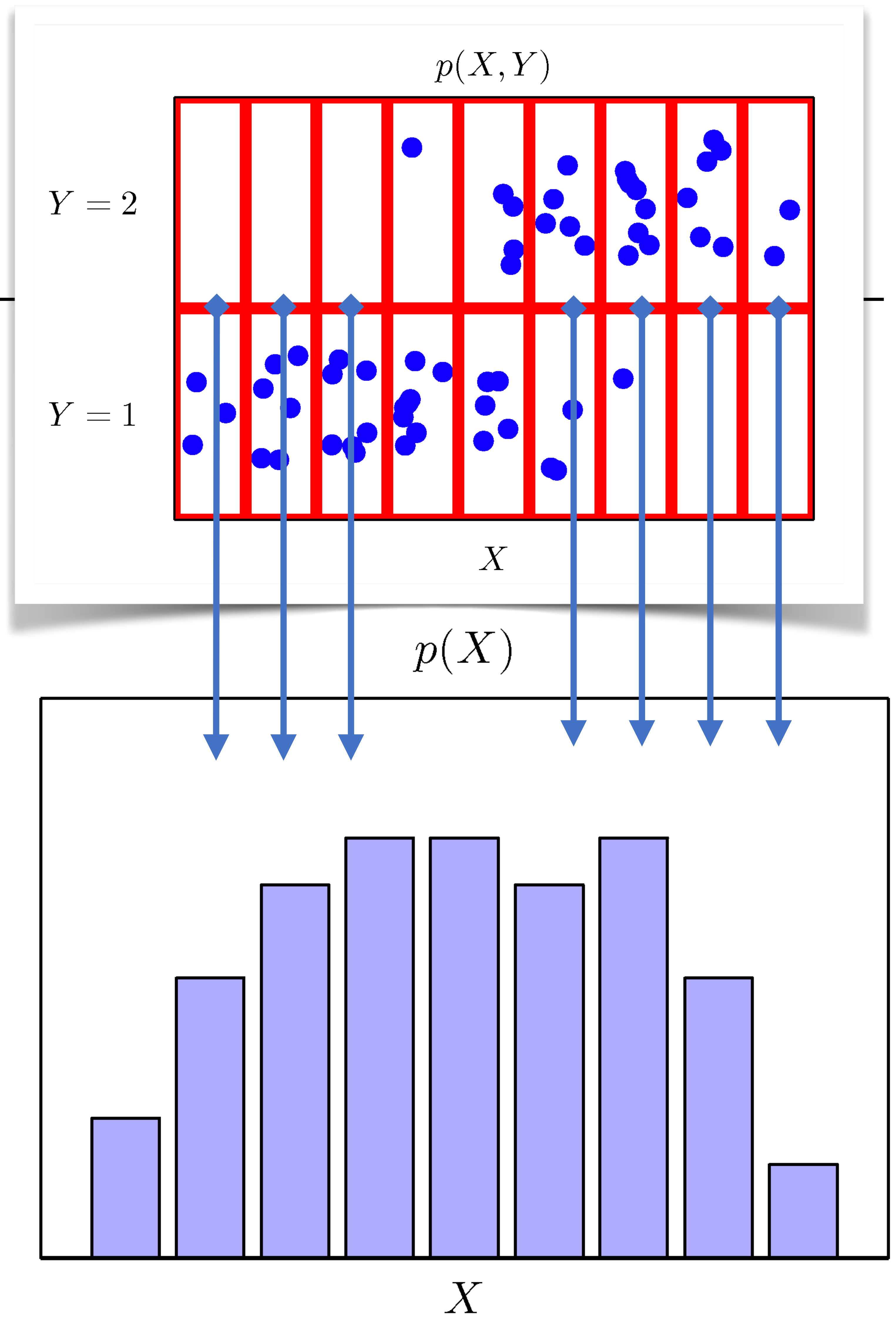


(출처: *Pattern Recognition and Machine Learning, Bishop*)

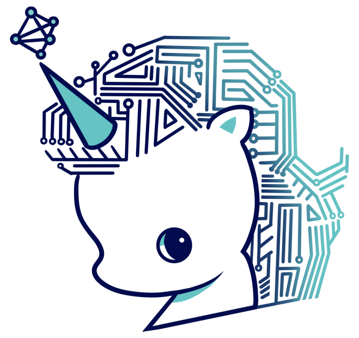
확률분포는 데이터의 초상화

- 데이터공간을 $\mathscr{X} \times \mathscr{Y}$ 라 표기하고 $\mathscr{D}$ 는 데이터공간에서 데이터를 추출하는 분포입니다
- 데이터는 확률변수로 $(\mathbf{x}, y) \sim \mathscr{D}$ 라 표기
- 결합분포 $P(\mathbf{x}, y)$ 는 $\mathscr{D}$ 를 모델링합니다
- $P(\mathbf{x})$ 는 입력 $\mathbf{x}$ 에 대한 주변확률분포로 y 에 대한 정보를 주진 않습니다

$$P(\mathbf{x}) = \sum_{y} P(\mathbf{x}, y) \quad P(\mathbf{x}) = \int_{y} P(\mathbf{x}, y)\mathrm{d}y$$



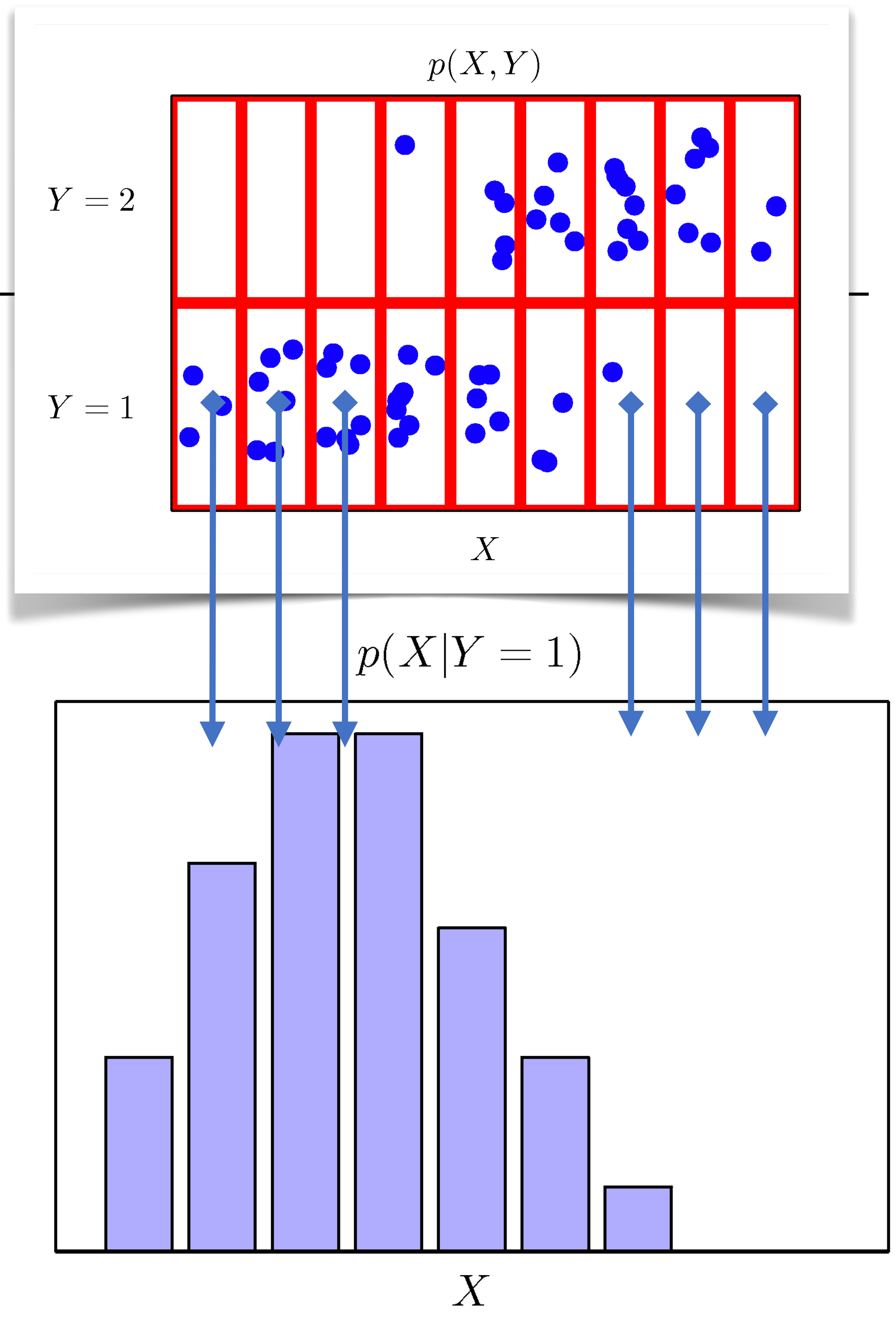
(출처: *Pattern Recognition and Machine Learning, Bishop*)



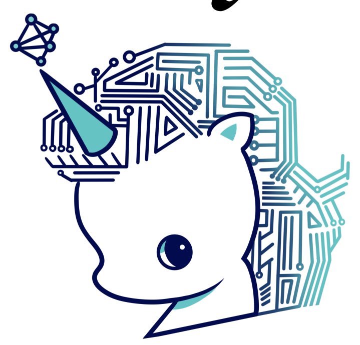
주변확률분포 $P(\mathbf{x})$ 는 결합분포 $P(\mathbf{x}, y)$ 에서 유도 가능합니다

확률분포는 데이터의 초상화

- 데이터공간을 $\mathscr{X} \times \mathscr{Y}$ 라 표기하고 $\mathscr{D}$ 는 데이터공간에서 데이터를 추출하는 분포입니다
- 데이터는 확률변수로 $(\mathbf{x}, y) \sim \mathscr{D}$ 라 표기
- 결합분포 $P(\mathbf{x}, y)$ 는 $\mathscr{D}$ 를 모델링합니다
- $P(\mathbf{x})$ 는 입력 $\mathbf{x}$ 에 대한 주변확률분포로 y 에 대한 정보를 주진 않습니다
- 조건부확률분포 $P(\mathbf{x} \mid y)$ 는 데이터 공간에서 입력 $\mathbf{x}$ 와 출력 y 사이의 관계를 모델링합니다



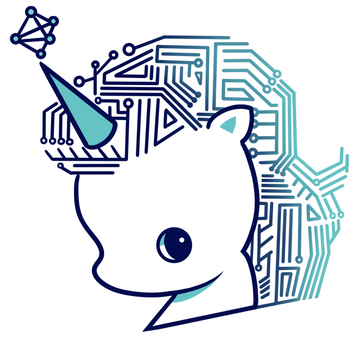
(출처: *Pattern Recognition and Machine Learning, Bishop*)



$P(\mathbf{x} \mid y)$ 는 특정 클래스가 주어진 조건에서 데이터의 확률분포를 보여줍니다

조건부확률과 기계학습

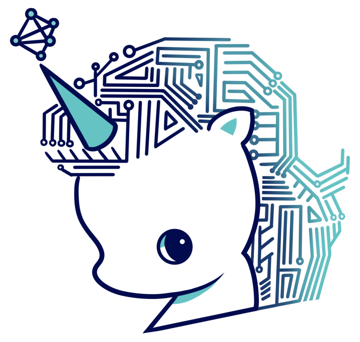
- 조건부확률 $P(y \mid \mathbf{x})$ 는 입력변수 $\mathbf{x}$ 에 대해 정답이 y 일 확률을 의미합니다



연속확률분포의 경우 $P(y \mid \mathbf{x})$ 는 확률이 아니고
밀도로 해석한다는 것을 주의하자

조건부확률과 기계학습

- 조건부확률 $P(y\,|\,\mathbf{x})$ 는 입력변수 $\mathbf{x}$ 에 대해 정답이 y 일 확률을 의미합니다
- 로지스틱 회귀에서 사용했던 선형모델과 소프트맥스 함수의 결합은 데이터에서 추출된 패턴을 기반으로 확률을 해석하는데 사용됩니다
- 분류 문제에서 softmax$(\mathbf{W}\phi + \mathbf{b})$은 데이터 $\mathbf{x}$ 로부터 추출된 특징패턴 $\phi(\mathbf{x})$ 과 가중치행렬 $\mathbf{W}$ 을 통해 조건부확률 $P(y\,|\,\mathbf{x})$ 을 계산합니다

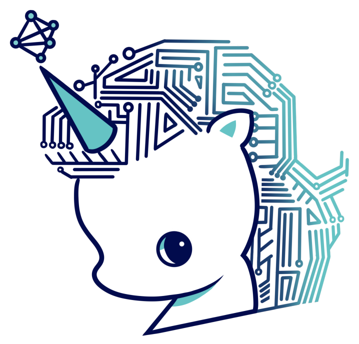


$P(y\,|\,\phi(\mathbf{x}))$ 이라 써도 된다

조건부확률과 기계학습

- 조건부확률 $P(y \mid \mathbf{x})$ 는 입력변수 $\mathbf{x}$ 에 대해 정답이 y 일 확률을 의미합니다
- 로지스틱 회귀에서 사용했던 선형모델과 소프트맥스 함수의 결합은 데이터에서 추출된 패턴을 기반으로 확률을 해석하는데 사용됩니다
- 분류 문제에서 softmax$(\mathbf{W}\phi + \mathbf{b})$은 데이터 $\mathbf{x}$ 로부터 추출된 특징패턴 $\phi(\mathbf{x})$ 과 가중치행렬 $\mathbf{W}$ 을 통해 조건부확률 $P(y \mid \mathbf{x})$ 을 계산합니다
- 회귀 문제의 경우 조건부기대값 $\mathbb{E}[y \mid \mathbf{x}]$ 을 추정합니다

$$\mathbb{E}_{y \sim P(y|\mathbf{x})}[y|\mathbf{x}] = \int_{y} yP(y|\mathbf{x})\mathrm{d}y$$

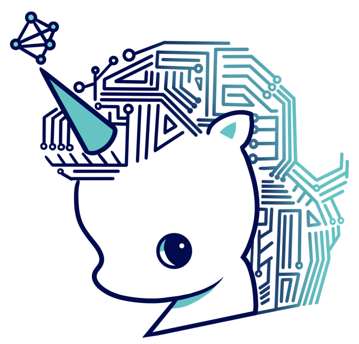


조건부기대값은 $\mathbb{E}\|y - f(\mathbf{x})\|_2$ 을 최소화하는 함수 $f(\mathbf{x})$ 와 일치한다

기대값이 뭔가요?

- 확률분포가 주어지면 데이터를 분석하는 데 사용 가능한 여러 종류의 통계적 범함수(statistical functional)를 계산할 수 있습니다
- 기대값(expectation)은 데이터를 대표하는 통계량이면서 동시에 확률분포를 통해 다른 통계적 범함수를 계산하는데 사용됩니다

$$\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[f(\mathbf{x})] = \int_{\mathcal{X}} f(\mathbf{x})P(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[f(\mathbf{x})] = \sum_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})P(\mathbf{x})$$



연속확률분포의 경우엔 적분을, 이산확률분포의 경우엔 급수를 사용한다

기대값이 뭔가요?

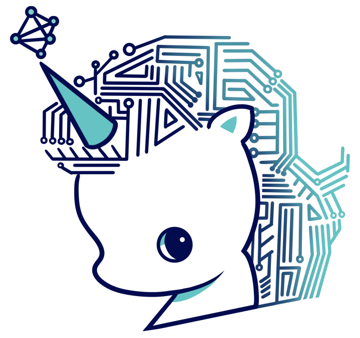
- 확률분포가 주어지면 데이터를 분석하는 데 사용 가능한 여러 종류의 통계적 범함수(statistical functional)를 계산할 수 있습니다
- 기대값(expectation)은 데이터를 대표하는 통계량이면서 동시에 확률분포를 통해 다른 통계적 범함수를 계산하는데 사용됩니다

$$\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[f(\mathbf{x})] = \int_{\mathcal{X}} f(\mathbf{x})P(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[f(\mathbf{x})] = \sum_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})P(\mathbf{x})$$

- 기대값을 이용해 분산, 첨도, 공분산 등 여러 통계량을 계산할 수 있습니다

$$\mathbb{V}(\mathbf{x}) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])^2] \qquad \text{Skewness}(\mathbf{x}) = \mathbb{E}\left[\left(\frac{\mathbf{x} - \mathbb{E}[\mathbf{x}]}{\sqrt{\mathbb{V}(\mathbf{x})}}\right)^3\right]$$

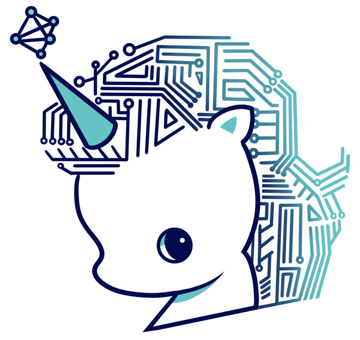
$$\text{Cov}(\mathbf{x}_1, \mathbf{x}_2) = \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2\sim P(\mathbf{x}_1,\mathbf{x}_2)}[(\mathbf{x}_1 - \mathbb{E}[\mathbf{x}_1])(\mathbf{x}_2 - \mathbb{E}[\mathbf{x}_2])]$$



위 수식에 f 대신 대입하면 통계량을 계산할 수 있다

조건부확률과 기계학습

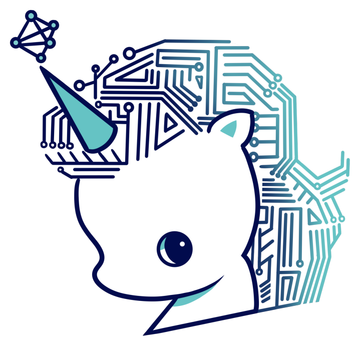
- 조건부확률 $P(y \mid \mathbf{x})$ 는 입력변수 $\mathbf{x}$ 에 대해 정답이 y 일 확률을 의미합니다
- 로지스틱 회귀에서 사용했던 선형모델과 소프트맥스 함수의 결합은 데이터에서 추출된 패턴을 기반으로 확률을 해석하는데 사용됩니다
- 분류 문제에서 softmax$(\mathbf{W}\phi + \mathbf{b})$은 데이터 $\mathbf{x}$ 로부터 추출된 특징패턴 $\phi(\mathbf{x})$ 과 가중치행렬 $\mathbf{W}$ 을 통해 조건부확률 $P(y \mid \mathbf{x})$ 을 계산합니다
- 회귀 문제의 경우 조건부기대값 $\mathbb{E}[y \mid \mathbf{x}]$ 을 추정합니다
- 딥러닝은 다층신경망을 사용하여 데이터로부터 특징패턴 ϕ 을 추출합니다



특징패턴을 학습하기 위해 어떤 손실함수를 사용할지는 기계학습 문제와 모델에 의해 결정된다

몬테카를로 샘플링

- 기계학습의 많은 문제들은 확률분포를 명시적으로 모를 때가 대부분이다
- 확률분포를 모를 때 데이터를 이용하여 기대값을 계산하려면 몬테카를로 (Monte Carlo) 샘플링 방법을 사용해야 한다



몬테카를로는 이산형이든
연속형이든 상관없이 성립한다

$$\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[f(\mathbf{x})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}^{(i)}), \qquad \mathbf{x}^{(i)} \overset{\text{i.i.d.}}{\sim} P(\mathbf{x})$$

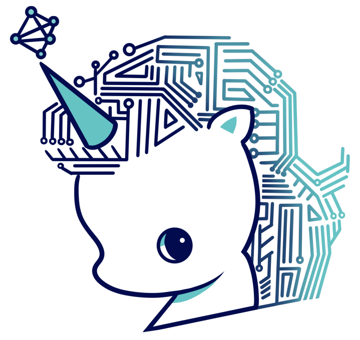
boostcamp AI Tech

몬테카를로 샘플링

- 기계학습의 많은 문제들은 확률분포를 명시적으로 모를 때가 대부분이다
- 확률분포를 모를 때 데이터를 이용하여 기대값을 계산하려면 몬테카를로 (Monte Carlo) 샘플링 방법을 사용해야 한다

$$\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[f(\mathbf{x})] \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}^{(i)}), \quad \mathbf{x}^{(i)} \overset{\text{i.i.d.}}{\sim} P(\mathbf{x})$$

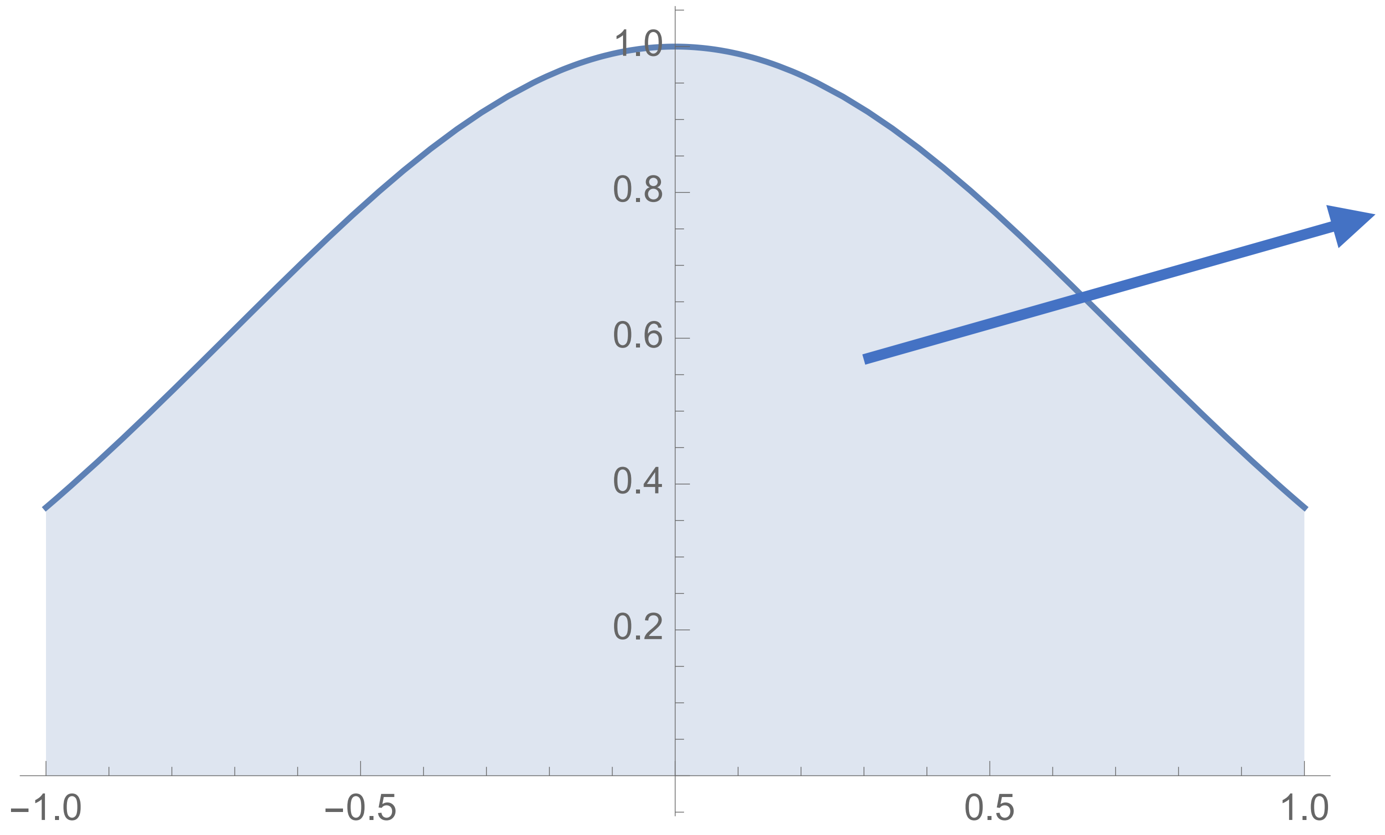
- 몬테카를로 샘플링은 독립추출만 보장된다면 대수의 법칙(law of large number)에 의해 수렴성을 보장한다



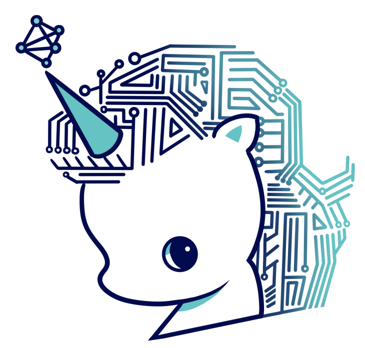
몬테카를로 샘플링은 기계학습에서
매우 다양하게 응용되는 방법입니다

몬테카를로 예제: 적분 계산하기

- 함수 $f(x) = e^{-x^2}$ 의 $[-1,1]$ 상에서 적분값을 어떻게 구할까?



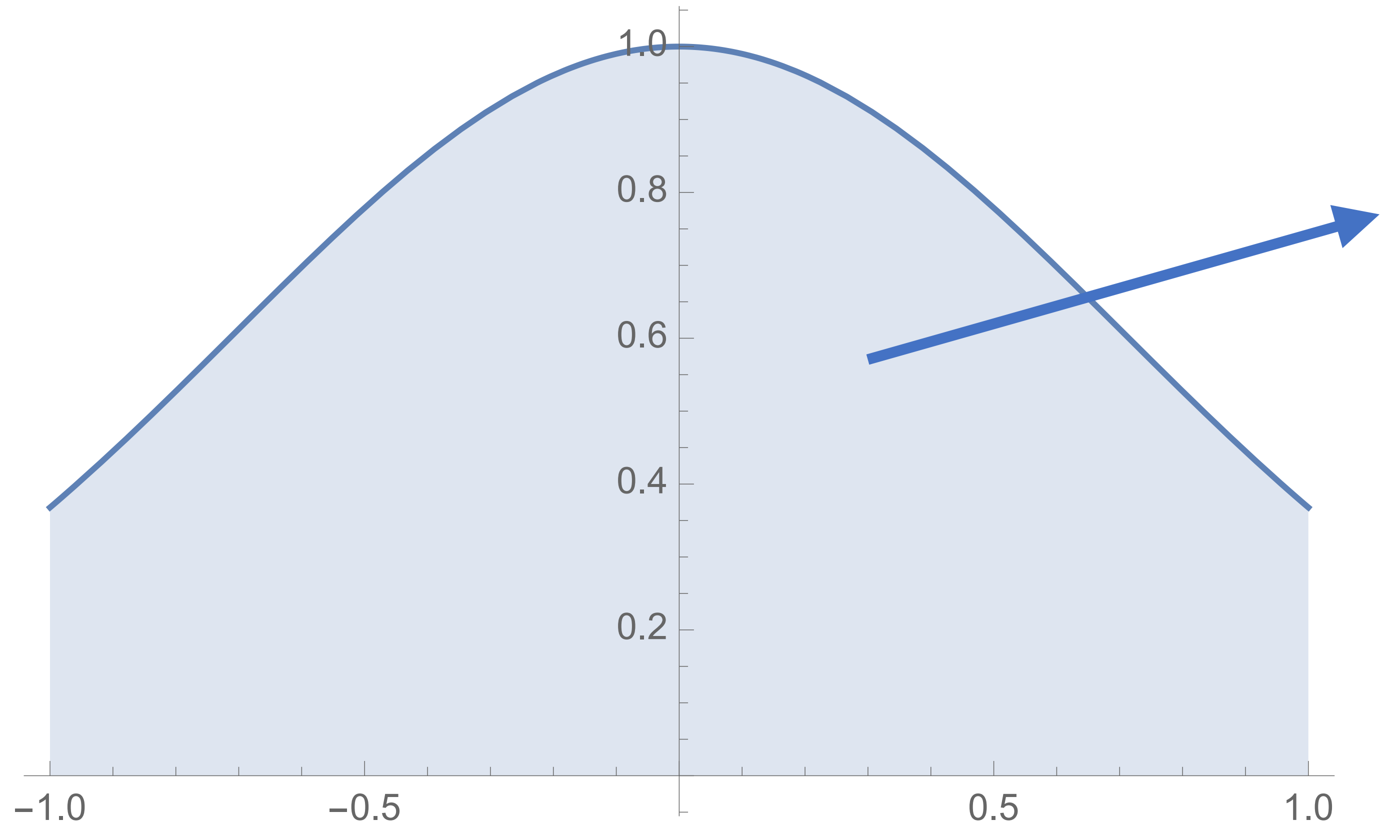
$$\int_{-1}^{1} e^{-x^2} \mathrm{d}x$$



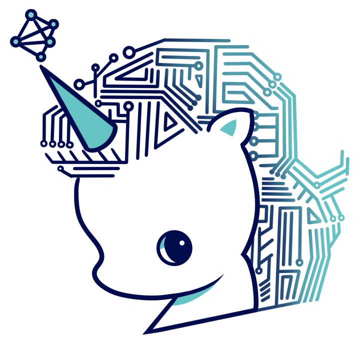
$f(x)$ 의 적분을 해석적으로 구하는 건 불가능하다

몬테카를로 예제: 적분 계산하기

- 함수 $f(x) = e^{-x^2}$ 의 $[-1,1]$ 상에서 적분값을 어떻게 구할까?



$$\int_{-1}^{1} e^{-x^2} \mathrm{d}x$$

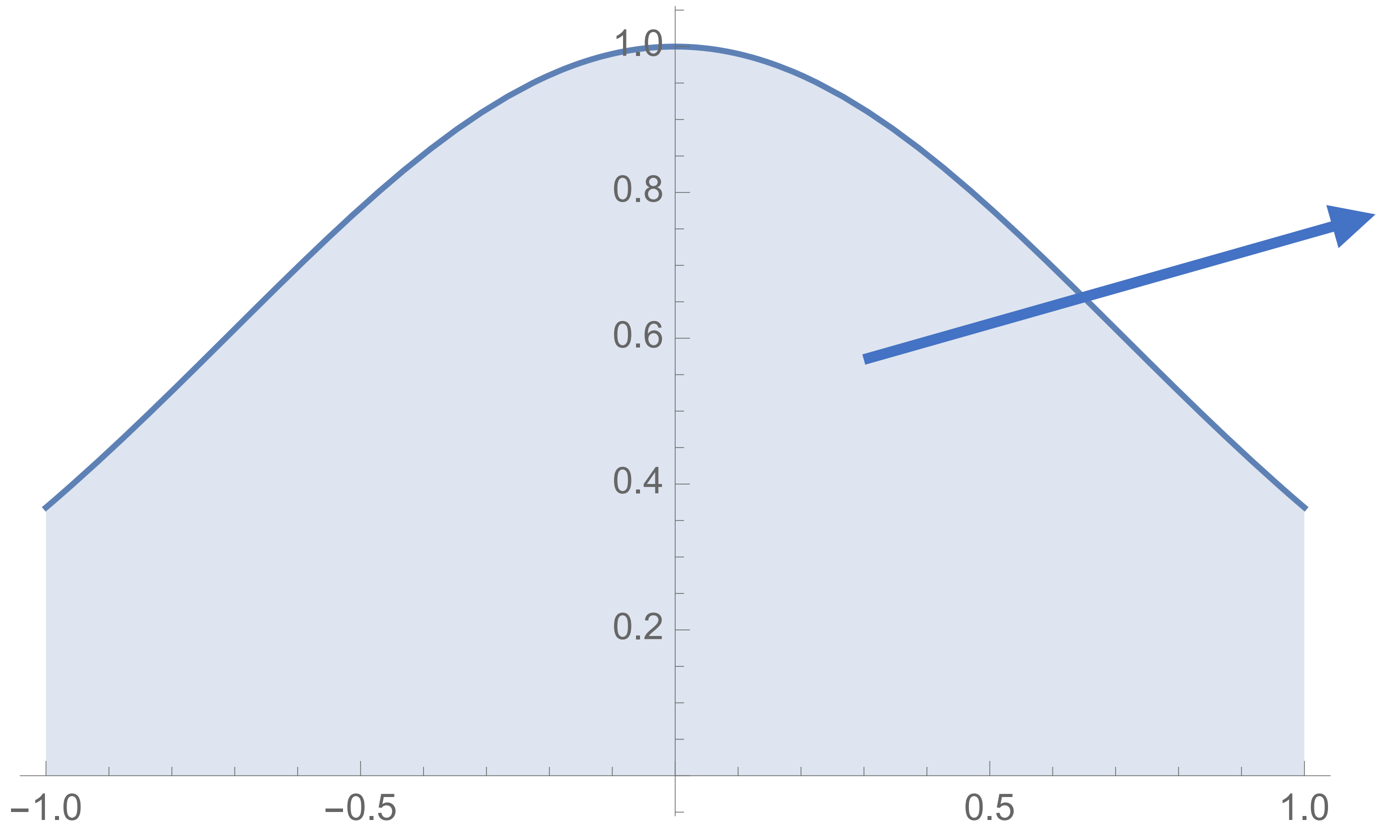


구간 $[-1,1]$ 의 길이는 2이므로 적분값을 2로 나누면 기대값을 계산하는 것과 같으므로 몬테카를로 방법을 사용할 수 있다

$$\frac{1}{2}\int_{-1}^{1} e^{-x^2} \mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}), \quad x^{(i)} \sim U(-1,1)$$

몬테카를로 예제: 적분 계산하기

- 함수 $f(x) = e^{-x^2}$ 의 $[-1,1]$ 상에서 적분값을 어떻게 구할까?



$$\int_{-1}^{1} e^{-x^2} \mathrm{d}x \approx 1.49364$$

$$\int_{-1}^{1} e^{-x^2} \mathrm{d}x \approx \frac{2}{N} \sum_{i=1}^{N} f(x^{(i)}), \quad x^{(i)} \sim U(-1,1)$$

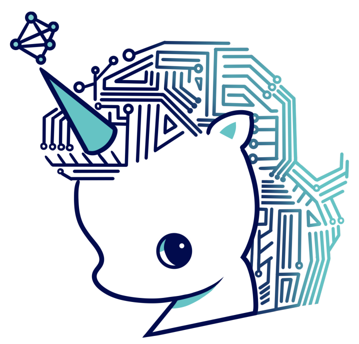
```python
import numpy as np

def mc_int(fun, low, high, sample_size=100, repeat=10):
    int_len = np.abs(high - low)
    stat = []
    for _ in range(repeat):
        x = np.random.uniform(low=low, high=high, size=sample_size)
        fun_x = fun(x)
        int_val = int_len * np.mean(fun_x)
        stat.append(int_val)
    return np.mean(stat), np.std(stat)

def f_x(x):
    return np.exp(-x**2)

print(mc_int(f_x, low=-1, high=1, sample_size=10000, repeat=100))
```

```
(1.4938754306231912, 0.00391398451303653)
```



1.49387±0.0039 이므로 오차 범위 안에 참값이 있다

boostcamp AI Tech

THE END

다음 시간에 보아요!