

Chapter 2. Basic tasks in ML

기계학습의 구분

- 지도학습 (Supervised Learning)
- 비지도학습 (Unsupervised Learning)
- 강화학습 (Reinforcement Learning)

지도학습

- 지도학습 (Supervised Learning)
- 입력데이터 x 와 출력데이터 y 사이의 관계를 파악하는 것이 목적
- x 에 대응되는 명확한 정답 y 가 존재함
- x 로부터 모델이 예측한 \hat{y} 가 실제 정답 y 와 같아지도록 함.
- 대표적인 태스크로 분류(classification)과 회귀(regression)으로 구분

정답이란?

- 분류 : 특정 클래스
 - 개, 고양이, 토끼 → One-hot vector 또는 정수 label로 변환
 - 정수 label 할당하는 경우 label 값의 증가는 입력 이미지와 관계 없음
 - 단순히 i-번째 클래스를 가리키는 것으로 사용
- 회귀 : 대응되는 실수값
 - 온도 → 38.6 도
 - 압력 → 1048hPa

비지도학습

- 비지도학습 (Unsupervised Learning)
- 주어진 데이터 x 들의 관계를 파악하는 데 중점
- x 에 대응하는 정답은 주어지지 않지만, x 들을 활용한 지표를 계산하고 이를 통해 관계를 학습함.
- 대표적인 태스크로 군집화(clustering)와 차원 축소 (dimension reduction)이 있음.

강화학습

- 강화학습 (Reinforcement Learning)
- 환경과 상호작용하며 현재 상황에 대한 최적의 의사결정을 학습하는데 중점
- 현재 관측을 바탕으로 기대 누적 보상 (expected cumulative reward)를 최대화
- 학습과정에서 현재 입력에 대응하는 보상 (reward)가 존재
- 단, 지도학습과 같이 명시적인 하나의 정답으로 볼 수 없음.
- e.g., 착한 행동을 한 아이에게 별사탕을 줄 것인가 혹은 막대사탕을 줄것인가?

지도학습의 두가지 태스크

- 분류(classification)
- 분류는 주어진 입력 데이터를 여러 개의 상호 배타적인 클래스 중 하나로 식별하거나 구분하는 작업을 의미
- 예시) 이메일을 ‘스팸’ 또는 ‘정상’으로 분류

지도학습의 두가지 태스크

- 회귀(regression)
- 회귀는 입력 데이터에 대응하는 연속적인 실수값을 예측하는 작업을 의미
- 예시) 주어진 집의 특성에 따라 가격을 예측하는 문제

지도학습

- 분류와 회귀 모두 입력 x 와 이에 대응하는 정답 y 를 가지고 있음.
- 지도학습의 공통적인 목적은 주어진 데이터로부터 $p(y|x)$ 를 정확하게 모델링하는 것
- y 의 분포의 유형에 따라 분류와 회귀 문제로 구분 가능.
- 분류: Bernoulli distribution, Categorical distribution
- 회귀: Gaussian distribution

Parametric distribution

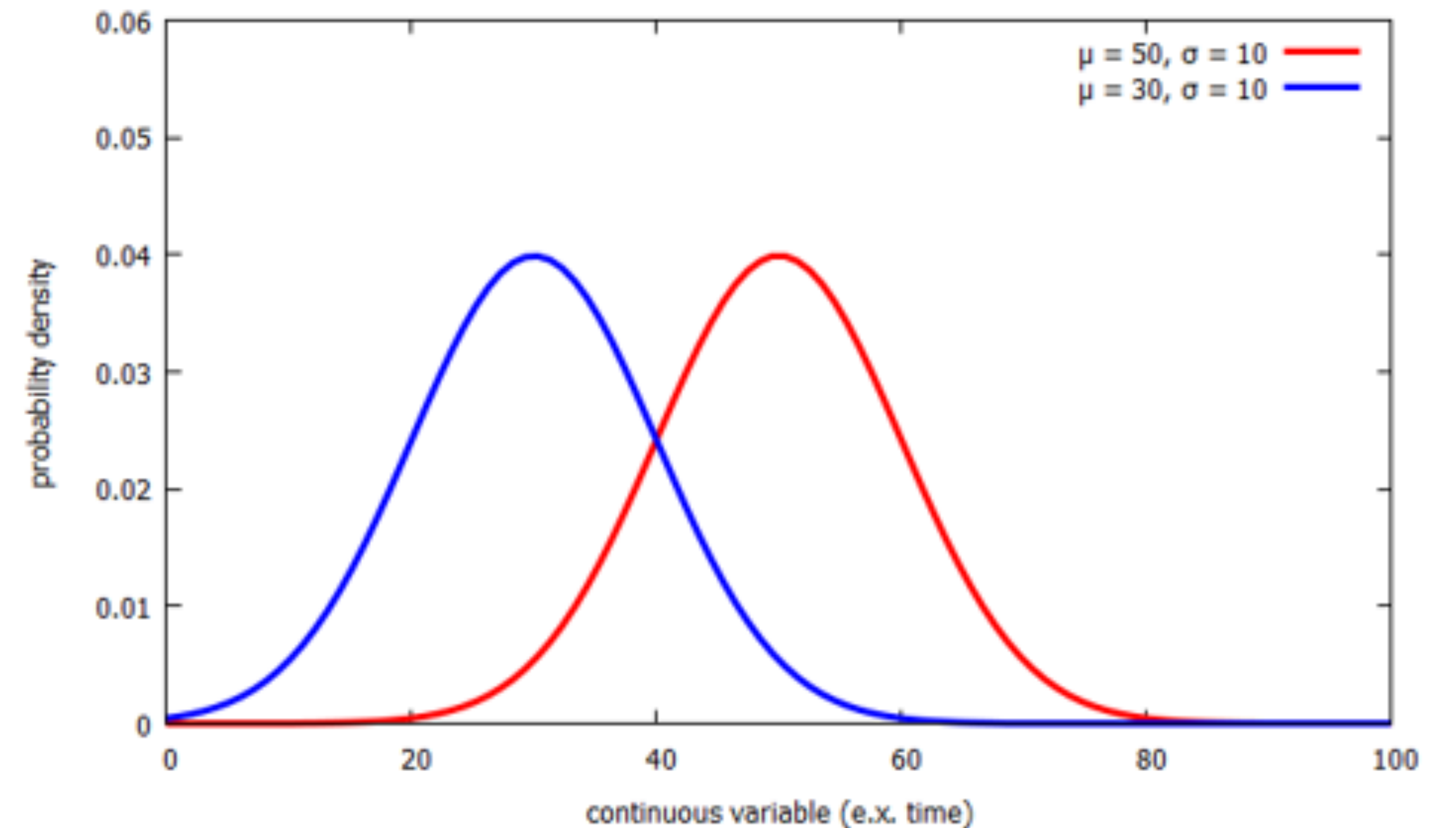
- Parametric distribution이란?
- 소수의 parameter로 전체 분포의 형태를 완전히 기술할 수 있는 확률분포
- **Bernoulli distribution:** $p \in [0,1]$, 성공 확률로 전체 분포가 결정됨
- **Categorical distribution:** p_1, p_2, \dots, p_k , $\sum p_i = 1$, 각 클래스에 대한 확률
- **Gaussian distribution:** μ, σ^2 , 평균과 분산으로 전체 분포가 결정됨.

Likelihood

- 우도(Likelihood)란?
- 랜덤변수 Y 가 파라미터 θ 로 정의되는 parametric distribution을 따르는 것으로 가정했을 때 (e.g., Gaussian, Categorical, Bernoulli...)
- 주어진 모델 파라미터 θ 에 대해 관측된 데이터가 발생할 확률
- $L(\theta) = p(y | \theta)$

Likelihood

- 두 개의 가우시안 분포를 가정
- $\theta_1 = (\mu_1, \sigma_1), \theta_2 = (\mu_2, \sigma_2)$
- $L(\theta_1) > L(\theta_2)$ 의 의미는?
- y 를 관측할 확률이 θ_1 이 더 크다.
- 즉 θ_1 이 θ_2 보다 확률 분포를 더 잘 설명한다.



Maximum Likelihood

- 수많은 θ 중 어떤 값을 선택해야 할까?
- $\theta^* = \operatorname{argmax}_{\theta} L(\theta) \Rightarrow$ maximum likelihood estimation

관측 데이터와 우도

- 지도학습에서 사용되는 데이터는 입력과 정답데이터 쌍들의 집합임.

- $D = \{(y_i, x_i)\}_{i=1}^N$

- 특정 샘플 (y_i, x_i) 에 대한 우도 $L_i(\theta) = p(y_i | x_i, \theta)$

- 각 샘플은 독립적이라 가정했을 때 전체 데이터셋에 대한 우도 $L_D(\theta)$ 는?

- $p(y_1 | x_1, \theta) \cdot p(y_2 | x_2, \theta) \cdot \dots \cdot p(y_N | x_N, \theta) = \prod_i p(y_i, x_i, \theta) = p(D | \theta)$

Negative log likelihood

- 주어진 데이터셋 D 에 대해 높은 likelihood를 갖는 θ^* 를 찾아야 함.
- $\theta^* = \operatorname{argmax}_{\theta} p(D | \theta)$
- Negative log 연산을 적용하여 곱셈을 덧셈의 형태로 변환하고 일반적인 최소화 기반 최적화 방법 적용
- 즉, Maximum likelihood estimation 문제가 Negative log likelihood의 최소화 문제로 변형

$$-\sum_i \log(p(y_i | x_i, \theta))$$

학습 모델과 손실함수

- 머신러닝/딥러닝 모델 = 함수 $f(x)$
- 단 함수의 형태를 결정하는 학습가능한 가중치 w 에 기반함.
- $f(x) \rightarrow f_w(x)$
- 어떤 가중치를 얻어야할까?
- 보통 손실함수를 최소화하는 가중치가 필요함.

Negative log likelihood와 loss function

- Parametric distribution에 대한 log likelihood: target = θ
- 학습가능한 파라미터를 갖는 머신러닝/딥러닝 모델: target = w
- 두 요소를 어떻게 연결 할 수 있을까?
- 바로 머신러닝/딥러닝 모델이 출력변수 y 에 대한 parametric distribution의 θ 를 예측하도록 함.

Negative log likelihood와 loss function

$$\theta = f_w(x)$$

$$\sum_i -\log(p(y_i | x_i, \theta)) = \sum_i -\log(p(y_i | x_i, w)) = \underline{-\log(p(D | w))}$$

Loss function

- Parametric distribution에 대한 Maximum likelihood estimation
- Negative log likelihood에 대한 최소화
- 머신러닝/딥러닝 모델의 loss function 최소화

주요 손실함수와 확률분포 - Bernoulli

- 베르누이 분포의 파라미터는 p , p 의 확률로 true, $(1 - p)$ 의 확률로 false
- $p = \hat{y} = f_w(x)$, 이때 likelihood는?
- $p(y = \text{true} | p) = p$, 파라미터가 p 일때 관측 y 가 true일 확률은? p
- $p(y = \text{false} | p) = 1 - p$, 파라미터가 p 일때 관측 y 가 false일 확률은? $1-p$

$$\sum_i -\log(p(y_i | x_i, \theta)) = \sum_{i \in \text{true}} -\log(p) + \sum_{i \in \text{false}} -\log(1 - p)$$

$$\sum_i -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

주요 손실함수와 확률분포 - categorical

- 카테고리 분포의 파라미터는 각 카테고리에 대한 확률들
- K개의 카테고리가 있다고 가정하면 카테고리 분포의 파라미터 $\theta = \{p_1, p_2, \dots, p_K\}$, $\sum p_k = 1$
- $p(y = c_1 | \theta) = p_1$, 주어진 파라미터에 대해 관측된 y가 1번 클래스일 확률은? p_1
- $p(y = c_k | \theta) = p_k$, 주어진 파라미터에 대해 관측된 y가 k번째 클래스일 확률은? p_k

$$\sum_i -\log(p(y_i | x_i, \theta)) = \sum_i -\log(p_{ik}) \rightarrow \log \text{ loss}$$

- p_{ik} 는 데이터셋에서 i번째 샘플이 k번째 클래스의 샘플일때 확률 p_k 를 의미함.

주요 손실함수와 확률분포 - Gaussian

- 가우시안 분포의 파라미터는 평균 μ , 분산 σ^2

$$y = g(x) + \epsilon, \epsilon \sim N(0, \sigma^2)$$

$$p(y | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$$


$$-\log(p(y_i | \mu, \sigma)) = \frac{1}{2} \log(2\pi\sigma^2) + \frac{(y_i - \mu)^2}{2\sigma^2}$$

- σ 를 상수로 취급하고, $\mu = f_w(x)$ 로 하면

$$\sum_i -\log(p(y_i | x_i, \theta)) = \sum_i (y_i - f_w(x_i))^2 \rightarrow \text{Mean squared error}$$

지도학습의 태스크와 출력변수 분포

- 지도학습의 경우 결국 loss function의 최소화를 통해서 MLE를 수행함.
- 또한 지도학습 태스크 (regression, classification)의 구분은 출력 변수의 분포에 따라 구분되는 것

태스크 유형	출력값 분포 가정	모델이 예측하는 것	사용 손실 함수	
회귀 (Regression)	정규분포 $\mathcal{N}(\mu, \sigma^2)$	평균 $\mu = f_w(x)$	평균제곱오차 (MSE)	
이진 분류	베르누이 분포 $\text{Bernoulli}(p)$	성공 확률 $p = f_w(x)$	이진 크로스 엔트로피	
다중 분류	카테고리 분포 $\text{Categorical}(p_1, \dots, p_k)$	클래스 확률 벡터 \vec{p}	다중 크로스 엔트로피	