심층학습 3장. 확률론과 정보 이론

송선영

3.1 확률의 필요성

- **확률론**
  - 불확실한 명제 서술
  - 불확실성이 존재하는 상황에서 뭔가를 추론할 수 있음

- **정보 이론**
  - 주어진 확률분포에 존재하는 불확실성의 양을 추정

- **불확실성이 발생할 수 있는 세 가지**
  - 모형화할 시스템에 내재한 확률성
  - 불완전한 관측 가능성
  - 불완전한 모형화

- 복잡하지만 확실한 규칙보다는 간단하지만 불확실한 규칙을 사용하는 것이 더 실용적

- **확률의 종류**
  - 빈도론자 확률
  - 베이즈 확률

3.2 확률변수

- **확률 변수**
  - 여러 값을 무작위하게 가지는 변수
  - 그 변수가 가질 수 있는 상태들을 서술
  - 반드시 확률 분포와 결합되어야 함
  - 이산(descrete)/ 연속(continuous)

  - **이산 확률 변수**
    - 개수가 유한한 셀 수 있는 확률변수
    - 반드시 정수는 아님

  - **연속 확률 변수**
    - 수치로 간주할 수 없음
    - 실숫값들과 연관됨

  - **확률 분포**
    - 각 상태가 실제로 확률변수의 값이 될 가능성

  - 표기법
    - 확률변수: 보통 글꼴의 영문 소문자 x
    - 확률변수의 값: 이탤릭 영문 소문자 x_1
    - 벡터값 변수: 굵은 글꼴 $\mathbf{x}$
    - 가능한 값 x

3.3 확률분포

- **확률 분포**
  - 하나의 확률변수 또는 확률변수들의 집합이 각각의 상태를 가질 가능도를 정의
  - 확률변수가 이산인지, 연속인지에 따라 확률분포를 서술하는 방식이 다름

  - **확률변수가 이산일 때**
    - 이산 변수를 서술하는 방법은 **확률질량함수 *P***

    - 확률질량함수란?
      - 확률변수의 한 상태를, 변수가 그 상태를 가질 확률
      - x= x 일 확률을 $P(x)$
      - 학률질량함수가 여러 변수에 동시에 작용할 수도 있음
        -> 다수의 변수에 관한 확률분포는 결합확률분포(=결합분포) $P(x, y)$

    - 어떤 함수 P가 확률변수 x에 대한 하나의 확률질량함수가 되기 위해 충족해야 할 조건
      - P의 정의역은 x의 모든 가능한 상태의 집합
      - $\forall x \in$ x, $0 \leq P(x) \leq 1$
      - $\sum_{x\in \text{ x}} P(x) = 1$ -> 이 성질을 충족하게 만드는 것을 정규화 라고 부름

3.3 확률분포

- **확률 분포**
  - 하나의 확률변수 또는 확률변수들의 집합이 각각의 상태를 가질 가능도를 정의
  - 확률변수가 이산인지, 연속인지에 따라 확률분포를 서술하는 방식이 다름

  - **확률변수가 연속일 때**
    - 연속 변수를 서술하는 방법은 **확률밀도함수 P**

    - 확률밀도함수란?
      - 확률밀도함수는 특정 상태의 확률을 직접 돌려주지는 않음
      - 대신, 확률변수의 값이 부피가 δx인 무한소 영역 안에 있을 확률이 $p(x)\delta x$ 임을 말해줌

    - 어떤 함수 P가 확률변수 x에 대한 하나의 확률밀도함수가 되기 위해 충족해야 할 조건
      - P의 정의역은 x의 모든 가능한 상태의 집합
      - $\forall\, x \in \mathrm{x},\ \ p(x) \geq 0$
      - $\int p(x)dx = 1$

    - 밀도함수를 적분하면 점 집합의 실제 확률질량을 구할 수 있음
      - x가 구간 [a,b]에 있을 확률은 $\int_{[a,b]} p(x)dx$

3.4 주변확률

- 부분집합에 관한 확률분포를 **주변확률분포 (=주변 분포)**

- x, y가 이산 확률변수이고, 그 둘에 대한 확률질량함수 $P(\mathrm{x},\mathrm{y})$ 를 알고 있다고 가정
  - 확률의 합의 법칙을 이용하여 $P(\mathrm{x})$를 구할 수 있음
  - $\forall\, x \in \mathrm{x},\ \ P(\mathrm{x} = x) = \sum_y P(\ \mathrm{x} = x,\ \mathrm{y} = y\ )$

- 연속 변수에 대해서는 합산 대신 적분을 사용해야 함
  - $\mathrm{P}(x) = \int p(x,y)dy$

3.5 조건부 확률

- 어떤 사건이 발생했을 때, 다른 어떤 한 사건이 발생할 확률을 **조건부 확률**
  - $\mathrm{P}(\mathrm{y} = y, \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$
  - 조건부 확률은 $P(\mathrm{x} = x) > 0$ 일 때만 정의됨

3.6 조건부 확률의 연쇄 법칙

- 다수의 확률변수에 관한 임의의 결합확률분포를 각각 하나의 변수에 관한 조건부 분포들로 분해할 수 있음
  - $P\left(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}\right) = P\left(\mathrm{x}^{(1)}\right) \prod_{i=2}^{n} P\left(\mathrm{x}^{(i)} \middle| \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}\right)$
  - 이러한 관계를 확률의 연쇄 법칙 또는 곱의 법칙이라고 부름

$$
\begin{aligned}
P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\
P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \\
P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}).
\end{aligned}
$$

3.7 독립과 조건부 독립

- 두 확률변수 x와 y의 확률분포를 x만 관여하는 인수와 y만 관여하는 인수의 곱으로 표현 가능
  - 이 때 두 변수(x,y)는 서로 **독립**임

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y).$$

- 확률변수 z가 주어졌을 대, 두 확률변수 x와 y에 관한 조건부 확률분포를 z의 모든 값에 대해 인수분해 할 수 있음
  - 이 때 두 변수(x,y)는 **조건부 독립**임

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z}, p(\mathrm{x} = x, \mathrm{y} = y | \mathrm{z} = z) = p(\mathrm{x} = x | \mathrm{z} = z)p(\mathrm{y} = y | \mathrm{z} = z).$$

3.8 기댓값, 분산, 공분산

- 확률분포 P(x)에 대한 함수 f(x) 의 **기댓값**은 P에서 뽑은 x들에 대한 f 값들의 평균을 의미
  - 이산 변수의 경우

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x) f(x).$$

  - 연속 변수의 경우

$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x) f(x) dx.$$

- 기댓값은 선형적임
- 만약, α와 β가 x에 의존하지 않는다고 할 때 아래 식이 성립함

$$\mathbb{E}_x[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_x[f(x)] + \beta \mathbb{E}_x[g(x)],$$

3.8 기댓값, 분산, 공분산

- **분산**
  - 확률변수 x의 함수가 해당 확률분포에서 비롯한 x 의 여러 값들에 따라 어느 정도나 변하는지를 나타내는 측도

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right].$$

  - 분산이 작다는 것은 f(x) 값들이 기댓값 주변에 몰려 있다는 것을 의미
  - 분산의 제곱급을 표준편차라고 부름

3.8 기댓값, 분산, 공분산

- **공분산**
  - 두 값의 선형 관계가 어느 정도인지, 그 값들의 규모가 어느 정도인지 말해 주는 측도

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])\right].$$

  - 공분산의 절댓값이 크다는 것을 값들이 아주 크게 변하며, 둘 다 해당 평균에서 동시에 크게 벗어남을 의미

  - 공분산이 양수: 두 변수가 동시에 상대적으로 큰 값을 가지는 경향이 있음을 의미
  - 공분산이 음수: 한 변수가 상대적으로 큰 값일 때 다른 한 변수는 상대적으로 작은 값

  - 그 외의 측도 : **상관계수**
    - 개별 변수의 규모에는 영향을 받지 않고 변수들의 관계만 측정하기 위해 각 변수의 기여를 정규화 한 것

  - 독립인 두 변수의 공분산은 0
  - 공분산이 0이 아닌 두 변수는 종속, 하지만 두 변수가 종속이라고 해서 공분산이 0이 아닌 건 아님
    - 두 변수가 종속이여도 공분산은 0일 수 있음

  - 확률벡터의 공분산행렬은 아래 식을 만족하는 n x n 행렬임
  - 공분산행렬의 주대각 성분은 분산임

$$\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j). \qquad \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \mathrm{Var}(\mathrm{x}_i).$$

3.9 흔히 쓰이는 확률분포들

- **베르누이 분포**
  - 하나의 이진 확률변수에 관한 분포
  - 이 분포는 매개변수 ∅ ∈ [0,1]
    - 이 매개변수는 주어진 확률변수가 1일 확률을 결정함

$$P(\mathrm{x}=1)=\phi$$

$$P(\mathrm{x}=0)=1-\phi$$

$$P(\mathrm{x}=x)=\phi^{x}(1-\phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}]=\phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x})=\phi(1-\phi)$$

3.9 흔히 쓰이는 확률분포들

- **멀티누이 분포**
  - 서로 다른 상태가 k개인 하나의 이산 변수에 관한 분포
    - K는 유한한 값

  - 멀티누이 분포의 매개변수는 벡터 $\boldsymbol{p} \in [0,1]^{k-1}$, (p_i는 i번째 상태의 확률)

  - 마지막 상태, k번째 상태의 확률은 $1 - 1^T p$로 정의 (반드시 $1^T p \leq 1$ 이여야 함)

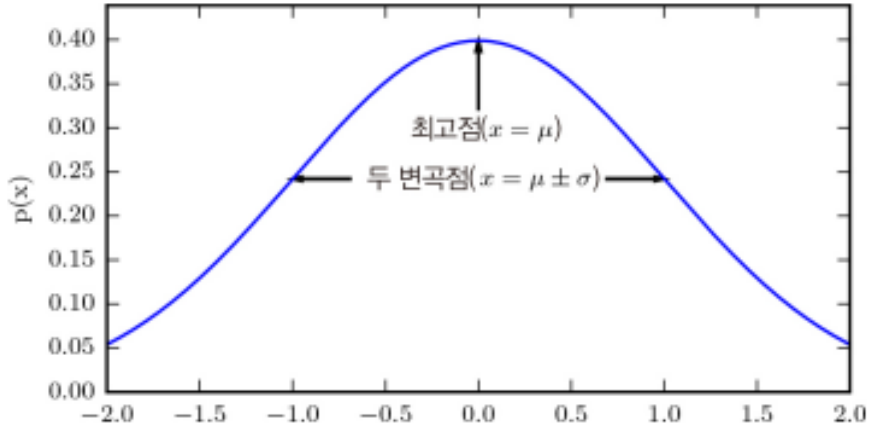
  - 대상들의 범주들에 관한 분포를 나타낼 때 자주 쓰임

  - 베르누이 분포와 멀티누이 분포는 모든 상태를 나열하는 것이 현실적으로 가능한 이산 변수들을 모형화 함
  - 반면, 연속 변수의 상태는 셀 수 없이 많으므로, 적은 수의 매개변수들로 서술하기에는 제약이 따름

3.9 흔히 쓰이는 확률분포들

- **가우스 분포**
  - 실수에 관한 분포 중 가장 흔히 쓰임
  - 정규 분포라고도 부름

$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\, exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$



3.9 흔히 쓰이는 확률분포들

- **가우스 분포**
  - 확률밀도함수를 평가하려면 σ의 제곱의 역수를 계산해야 함
  - 서로 다른 매개변수들로 확률밀도함수를 자주 평가해야 할 경우,
    - 분포의 분산의 역수에 해당하는 매개변수 $\beta \in (0, \infty)$를 이용해서 분포를 매개변수화하는 것이 효율적임
      -> 이 매개변수는 분포의 정밀도(precision)에 해당함

$$\mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} exp\left(-\frac{1}{2}\beta(x-\mu)^2\right).$$

- 기계 학습이 다루는 실수 수치들의 분포가 구체적으로 어떤 것인지 알지 못하는 상황에서, 기본적으로 정규분포를 선택하는 것이 바람직함

- 그 이유는?
  - 우리가 모형화하고자 하는 분포 중에는 정규 분포에 아주 가까운 것들이 많음
    - 중심극한정리에 따르면, 다수의 독립 확률변수들의 합은 근사적으로 정규분포를 따름
    - 중심극한정리: 동일한 확률분포를 가진 독립 변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다는 정리

  - 같은 분산을 가진 모든 가능한 분포 중에서 가장 많은 양의 불확실성을 부호화하는 것이 정규분포임
    - 즉, 정규분포는 모형에 주입하는 사전 지식의 양이 가장 적은 분포

3.9 흔히 쓰이는 확률분포들

- **지수분포와 라플라스 분포**
  - 심층학습에서는 최고점이 $x = 0$ 에 있는 확률분포를 사용하는 것이 바람직할 때가 많음
  - 이런 조건을 충족하는 방법이 지수분포를 사용하는 것임

  - 지수분포
    - 정의함수 $1_{x \geq 0}$ 을 이용해 x 의 모든 음수 값에 대해 확률 0을 배정

$$p(x;\lambda) = \lambda 1_{x \geq 0} \exp(-\lambda x).$$

  - 라플라스 분포
    - 지수분포와 밀접한 관계가 있는 분포
    - 확률질량의 최고점을 원하는 임의의 점 μ

$$\text{Laplace}(x;\mu,\gamma) = \frac{1}{2\gamma} exp\left(-\frac{|x-\mu|}{\gamma}\right).$$

3.9 흔히 쓰이는 확률분포들

- **디랙 분포와 경험분포**
  - 확률분포의 모든 질량이 한 점 주변에 몰려 있는 것이 바람직할 때가 있음
  - 이런 경우, 디랙 델타 함수 $\delta(x)$ 로 확률밀도함수를 정의하면 됨

$$p(x) = \delta(x - \mu).$$

  - 디랙 델타 함수
    - 0을 제외한 모든 곳에서는 값이 0이지만, 적분하면 1이 되는 함수
    - 디랙 델타 함수는 초함수임
      - 초함수: 적분할 때의 성질들에 의해 정의되는 또 다른 종류의 함수

    - 일반적으로 아래 식과 같은 경험분포의 한 구성요소로 쓰임

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x} - \boldsymbol{x}^{(i)}).$$

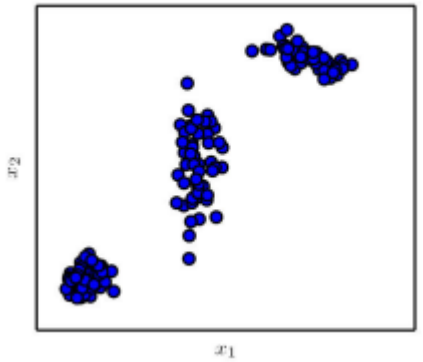
3.9 흔히 쓰이는 확률분포들

- **분포의 혼합**
  - 간단한 확률분포들을 결합해서 좀 더 복잡한 확률분포를 정의하는 경우도 많음
  - 이런 경우, 분포들을 결합할 때 사용하는 방법은 여러 분포로 하나의 **혼합분포**를 만드는 것임

  - **혼합분포**
    - 다수의 분포들로 구성
    - 각 시행에서는 멀티누이 분포에서 추출한 단위 성분으로 혼합분포의 표본을 생성함
      - $P(c)$는 단위 성분들에 관한 멀티누이 분포

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} | \mathrm{c} = i).$$

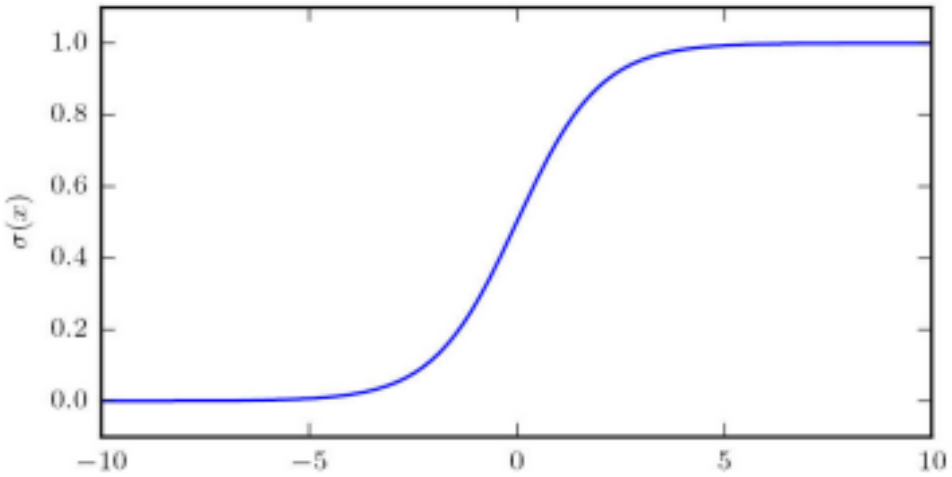
    - 흔히 쓰이는 강력한 혼합모형으로 **가우스 혼합모형**이 있음



3.10 흔히 쓰이는 함수들의 유용한 성질들

- 확률분포를 다룰 때, 자주 쓰는 함수들이 있음
  - 로그 S자형 함수 (logistic sigmoid)
    - 베르누이 분포의 매개변수를 산출할 때 쓰임

$$\sigma(x) = \frac{1}{1+\exp(-x)}.$$

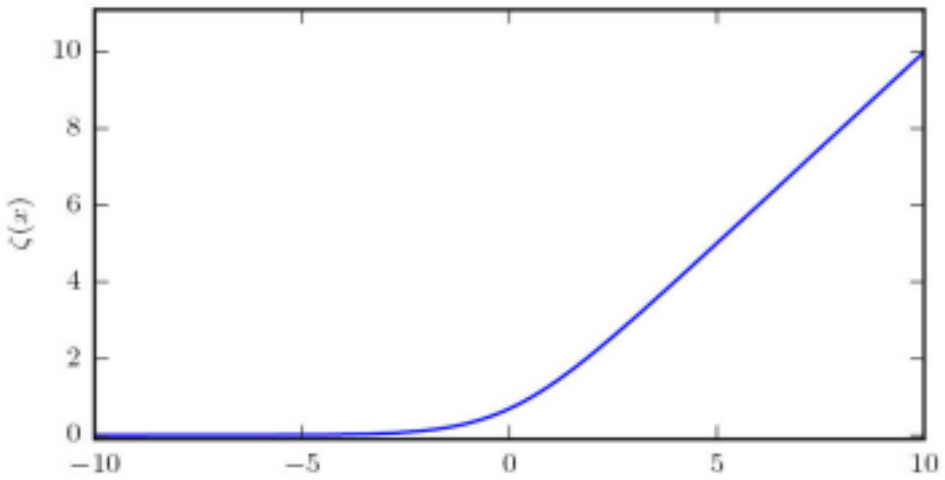


- 인수가 큰 음수나 큰 양수일 때, 함수는 평평한 수평선에 가까움
- 이는 입력이 변해도 함수의 값이 거의 변하지 않음을 뜻함
- 이런 상황을 S자형 함수가 포화되었다고 말함

3.10 흔히 쓰이는 함수들의 유용한 성질들

- 확률분포를 다룰 때, 자주 쓰는 함수들이 있음
  - 소프트플러스 함수
    - 정규분포의 매개변수를 산출할 때 사용

$$\zeta(x) = \log(1 + \exp(x)).$$



3.10 흔히 쓰이는 함수들의 유용한 성질들

- 확률분포를 다룰 때, 자주 쓰는 함수들이 있음
  - 소프트플러스 함수의 식

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + \exp(0)}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

$$1 - \sigma(x) = \sigma(-x)$$

$$\log \sigma(x) = -\zeta(-x)$$

$$\frac{d}{dx}\zeta(x) = \sigma(x)$$

$$\forall x \in (0,1), \sigma^{-1}(x) = \log\left(\frac{x}{1-x}\right)$$

$$\forall x > 0, \zeta^{-1}(x) = \log(\exp(x) - 1)$$

$$\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy$$

$$\zeta(x) - \zeta(-x) = x$$

Thank You

감사합니다.