Lecture 13: 목적함수와 정보이론 (교재 Chapter 11 대립학습모델)

<기계학습 개론> 2019 강의
서울대학교 컴퓨터공학부
장 병 탁

교재: 장교수의 딥러닝, 홍릉과학출판사, 2017.

Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University

BI
BIOINTELLIGENCE

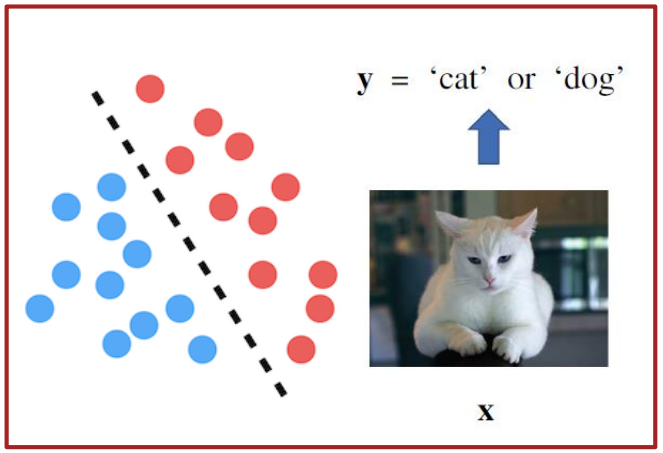
Version 20171122/20191030

목차

11.1 변별모델과 생성모델 (1/3)

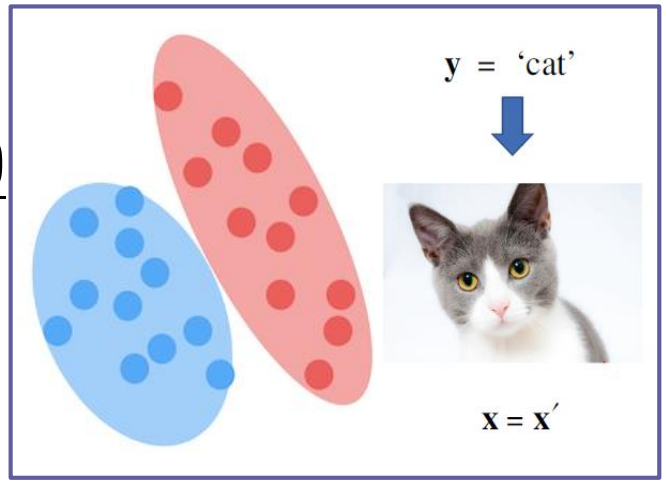
- 머신러닝 모델의 학습
  - 데이터: $D = \{(\mathbf{x}, \mathbf{y})\}$
  - 모델: $f(\mathbf{x}; \mathbf{w})$, 여기서 $\mathbf{w}$ 는 파라미터
- 변별모델 (분류모델)
  - 변별모델은 주어진 $\mathbf{x}$ 에 대한 $\mathbf{y}$ 의 조건부 확률을 학습
- 생성모델
  - 데이터의 확률 분포를 학습
  - $P(\mathbf{x}, \mathbf{y})$ 또는 $P(\mathbf{x})$ 모델링
  - $\mathbf{y}$ 를 샘플링 할 수 있음

$$f(\mathbf{x}; \mathbf{w}) = P(\mathbf{y}|\mathbf{x})$$



<변별모델>

$$P(\mathbf{y}|\mathbf{x}) = \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{x})}$$



<생성모델>

11.1 변별모델과 생성모델 (2/3)

- 변별모델의 목적 함수
  - 주어진 점 $\mathbf{x}$에 대한 $f(\mathbf{x};\mathbf{w}) = \mathbf{y}$’ 와 $\mathbf{y}$ 의 차이를 최소화

$$f : X \times W \rightarrow Y$$
$$f(\mathbf{x};\mathbf{w}) = \mathbf{y}'$$
$$\min_{\mathbf{w}} L(\mathbf{y},\mathbf{y}') = \min_{\mathbf{w}} L(\mathbf{y}, f(\mathbf{x};\mathbf{w}))$$

- 생성모델의 목적 함수
  - 주어진 점 $\mathbf{x}$ 에 대한 $f(\mathbf{x};\mathbf{w}) = \mathbf{x}$’와 $\mathbf{x}$ 의 차이를 최소화

$$f : X \times W \rightarrow X$$
$$f(\mathbf{x};\mathbf{w}) = \mathbf{x}'$$
$$\min_{\mathbf{w}} L(\mathbf{x},\mathbf{x}') = \min_{\mathbf{w}} L(\mathbf{x}, f(\mathbf{x};\mathbf{w}))$$

11.1 변별모델과 생성모델 (3/3)

변별모델과 생성모델의 장점을 결합하는 방법

- 변별모델(감독학습)을 위해 생성모델(무감독 학습)을 이용
  - 기존의 전통적인 머신러닝에서 많이 사용
  - 예) PCA로 특징을 추출한 후, SVM으로 학습
  - 예) DBN/DHN은 무감독 학습을 먼저 한 후, 감독 학습 수행
- 생성모델(무감독학습)을 위해 변별모델(감독 학습)을 이용
  - 생성대립넷(GAN)은 변별모델을 이용하여 생성모델을 학습
  - 임의의 샘플은 실제 데이터 분포를 따를 가능성이 낮음
  - 목적함수를 변별학습 문제로 형식화함으로써 해결
- GAN은 변별모델을 이용하여 생성모델을 학습하려는 시도
  - 게임이론을 도입하여 생성모델과 변별모델의 대립관계로 형식화
  - 생성모델의 성능이 변별모델의 평가를 받아서 계속 향상
  - 생성모델 기반 시행착오에 의해서 임의의 사진을 생성한 후 변별모델의 가이드에 기반한 감독학습 방식으로 계속 차이를 줄여나가면서 수정

11.2 정보이론과 상대 엔트로피 (1/6)

- 생성모델의 목적 함수
  - 생성모델 학습을 위해서는 정보이론에 기반한 목적함수가 중요
  - 생성모델의 목적은 실제 데이터의 확률 분포 $P(x)$의 학습
  - 실제 데이터의 확률 분포 $P\,(x)$와 모델의 확률 분포 $Q\,(x)$와의 차이를 정의
- 정보/정보량(Information)
  - 확률변수 X는 사건을 나타내고 이 사건의 결과로 나타난 값을 x 라고 정의
  - $P(x)$는 사건 x 의 확률
  - 사건을 관측할 정보량 $I(x)$은 확률에 반비례하도록 정의
    - log의 밑을 2로 한 것은 정보량을 bit수로 측정하기 위함
  - 확률이 적은 사건이 발생할수록 많은 정보를 제공

$$I(x) = \log_2 \frac{1}{P(x)}$$

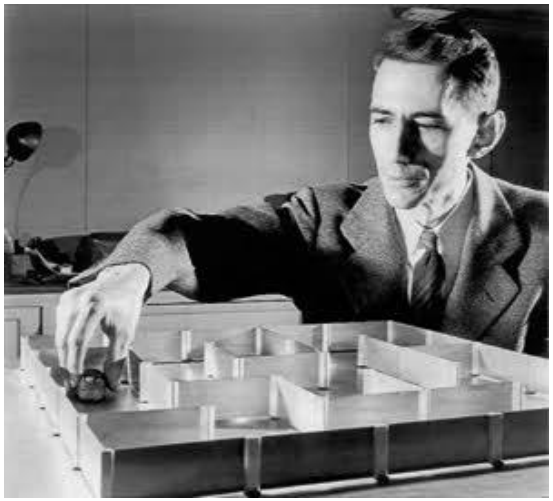
11.2 정보이론과 상대 엔트로피 (2/6)

- 평균 정보량(Average Information)
  - x_i는 $P_i = P(x_i)$ 의 확률로 생성될 때
  - 하나의 기호 수신을 통해서 얻는 평균 정보량

$$\langle I \rangle = \mathrm{E}_{x\sim P}\left[I(x)\right]$$

$$= \sum_{i=1}^{N} P(x_i) I(x_i) = \sum_{i=1}^{N} P(x_i) \log_2 \frac{1}{P(x_i)}$$

$$= -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i)$$



- 엔트로피(Entropy)
  - 물리 시스템의 혼돈(disorder)의 정도를 재는 척도
  - 정보원 X의 (정보) 엔트로피 $H(X)$는 그 정보원의 평균 정보량으로 정의

$$H(X) = \mathrm{E}_{x\sim P}[I(x)] = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i)$$

- 불확실성(Uncertainty)
  - 엔트로피는 정보원으로부터 생성되는 기호를 예측하는데 따르는 불확실성의 정도를 측정
  - 모든 기호들이 균일한 확률로 생성될 때, 즉 $P(x_i) = \frac{1}{N}$일 때 이 시스템은 가장 큰 엔트로피를 가짐.

11.2 정보이론과 상대 엔트로피 (3/6)

- 결합 엔트로피(Joint Entropy)
  - 두 개의 확률변수가 결합된 시스템에 대한 엔트로피로 정의

$$H(X,Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_i)\log_2 P(x_i, y_i)$$

  - 단일 변수의 엔트로피와 두 변수의 결합엔트로피 간에는 다음의 관계가 성립
  - 두 변수의 결합 엔트로피는 각 변수의 엔트로피의 단순 합보다는 항상 같거나 작음

$$H(X,Y) \le H(X) + H(Y)$$

- 상호 엔트로피(Mutual Entropy)
  - 상관 엔트로피(correlation entropy) 또는 상호정보량(mutual information)
  - 상호정보량은 결합 엔트로피가 $H(X, Y)$인 두 개의 확률변수 X와 Y가 공유하는 정보로 정의

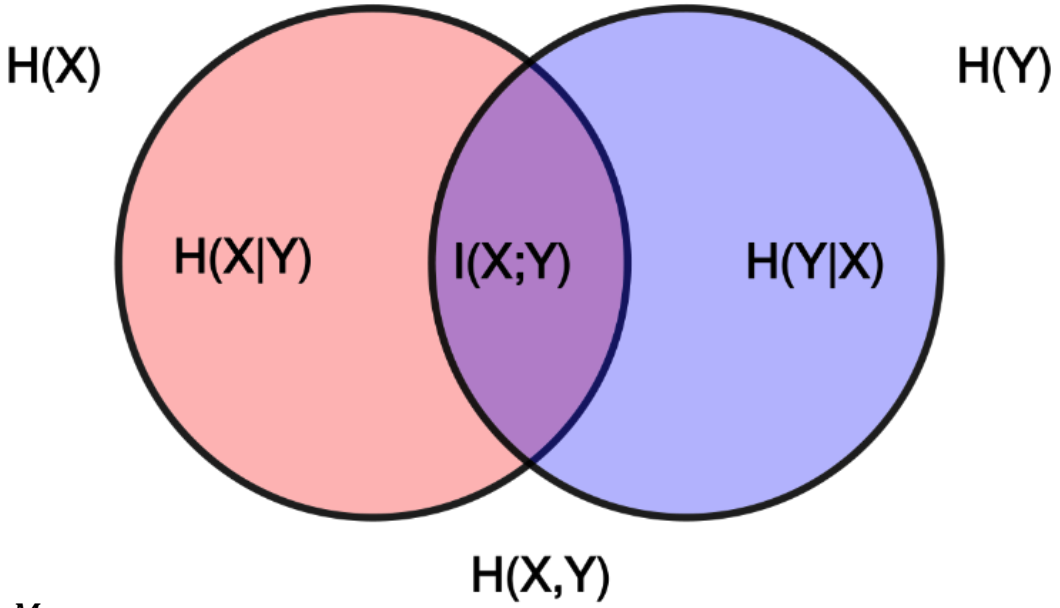
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

  - 상호정보량은 원래 엔트로피 $H(X)$와 조건부 엔트로피 $H(X|Y)$의 차이

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

11.2 정보이론과 상대 엔트로피 (4/6)

- 대칭성
  - $I(X;Y) = I(Y;X)$
  - 단, 조건부 엔트로피는 대칭이 아님
- 독립성/의존성
  - 상호 엔트로피는 두 확률변수가 독립인 경우에만 0



$$I(Y;X) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log_2 P(x_i; y_j)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j)\log_2 \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$$

11.2 정보이론과 상대 엔트로피 (5/6)

- 교차 엔트로피(Cross Entropy)
  - 교차 엔트로피는 한 확률 변수에 대한 두 개의 확률 분포에 대한 척도

$$H(P,Q) = -\sum_{i=1}^{N} P(x_i)\log_2 Q(x_i) \qquad H(P,Q) = -\sum_{i=1}^{N} P(x_i)\log_2 Q(x_i)$$

  - 분포 P에 대해서 계산한 분포 Q의 평균 정보량

$$H(P,Q) = \mathrm{E}_{x\sim P}[\log_2 Q(x_i)] \qquad H(P,Q) = \mathrm{E}_{x\sim P}[\log_2 Q(x_i)]$$

  - 교차 엔트로피는 참인 분포 P대신에 주어진 확률 분포 Q에 기반한 코딩 방법이 사용된다면 확률의 집합으로부터 하나의 사건을 알아 맞추는데 필요한 평균 비트 수
- 상대 엔트로피(Relative Entropy) = KL 다이버전스 D_{KL}
  - 상대 엔트로피는 같은 확률 변수에 대한 두 개의 확률 분포의 차이를 측정

$$D_{KL}(P||Q) = \sum_{i=1}^{N} P(x_i)\log_2 \frac{P(x_i)}{Q(x_i)} \qquad D_{KL}(P\,\|\,Q) = \sum_{i=1}^{N} P(x_i)\log_2 \frac{P(x_i)}{Q(x_i)}$$

  - 분포 P와 분포 Q간의 코드 길이 차이의(분포 P에 대한) 평균값

$$D_{KL}(P||Q) = \mathrm{E}_{\mathrm{x}\sim\mathrm{P}}\left[\log_2 \frac{P(x_i)}{Q(x_i)}\right] = \mathrm{E}_{\mathrm{x}\sim\mathrm{P}}[\log_2 P(x_i) - \log_2 Q(x_i)]$$

11.2 정보이론과 상대 엔트로피 (6/6)

- 상대 엔트로피는 다음의 깁스 부등식(Gibbs' inequality)을 만족
  - $P = Q$일 때만 0의 값을 가짐.

$$D_{KL}(P||Q) \geq 0$$

$$\begin{aligned} D_{KL}(P||Q) &= \sum_{i=1}^{N} P(x_i) \log_2 \frac{P(x_i)}{Q(x_i)} \\ &= \sum_{i=1}^{N} P(x_i) \log_2 P(x_i) - \sum_{i=1}^{N} P(x_i) \log_2 Q(x_i) \\ &= -\sum_{i=1}^{N} P(x_i) \log_2 Q(x_i) - \left[-\sum_{i=1}^{N} P(x_i) \log_2 P(x_i)\right] \\ &= H(P,Q) - H(P) \end{aligned}$$

- 상대 엔트로피의 최소화
  - P와 Q의 교차 엔트로피
  - 최소 교차 엔트로피의 원리
- 상호정보량(MI)과 KL 다이버전스의 관계
  - 결합 확률 $P(X,Y)$와 주변확률분포의 곱 $P(X)P(Y)$ 간의 상대 엔트로피.

$$\begin{aligned} I(Y;X) &= \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \\ &= D_{KL}(P(X,Y)||P(X)P(Y)) \end{aligned}$$

KL 다이버전스의 특성

KL Divergence (KLD) between $p_{\mathbf{X}}(\mathbf{x})$ and $g_{\mathbf{X}}(\mathbf{x})$

$$D_{p||g} = \int_{-\infty}^{\infty} p_{\mathbf{X}}(\mathbf{x}) \log\left(\frac{p_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})}\right) d\mathbf{x}$$

$$= \mathbf{E}\left[\log\left(\frac{p_{\mathbf{X}}(\mathbf{x})}{g_{\mathbf{X}}(\mathbf{x})}\right)\right]$$

A distance between two pr obability distributions, but no symmetricity, thus diver gence.

$$D_{p||g} \neq D_{g||p}$$

Property 1. Nonnegativity

$$D_{p||g} \geq 0$$

Property 2. Invariance

$$D_{p_{\mathbf{X}}||g_{\mathbf{X}}} = D_{p_{\mathbf{Y}}||g_{\mathbf{Y}}}$$

참고: 확률분포추정과 통계역학

p_i : probability of occurrence of state i of a stochastic system

$p_i \geq 0$ (for all i) and $\sum_i p_i = 1$

E_i : energy of the system when it is in state i

In thermal equilibrium, the probability of state i is

(Canonical distribution / Gibbs distribution)

$$p_i = \frac{1}{Z} \exp\left(- \frac{E_i}{k_B T} \right)$$

$$Z = \sum_i \exp\left(- \frac{E_i}{k_B T} \right)$$

$\exp\left(- E / k_B T \right)$: Boltzmann factor

Z : sum over states (partition function)

We set $k_B = 1$ and view $- \log p_i$ as "energy"

1. States of low energy have a higher probability of occurrence than the st ates of high energy.
2. As the temperature T is reduced, th e probability is concentrated on a s maller subset of low-energy states.

참고: 몬테칼로 시뮬레이션
Metropolis Algorithm

Metropolis Algorithm

A stochastic algorithm for simulating the evolution of a physical system to thermal equilibrium. A modified Monte Carlo method.

Markov Chain Monte Carlo (MCMC) method

Algorithm Metropolis

1. $X_n = x_i$. Randomly generate a new state x_j.

2. $\Delta E = E(x_j) - E(x_i)$

3. If $\Delta E < 0$, then $X_{n+1} = x_j$

   else if $\Delta E \geq 0$, then

   { Select a random number χ U[0,1].

   If $\chi < \exp(-\Delta E / T)$, then $X_{n+1} = x_j$, (*accept*)

   else $X_{n+1} = x_i$. (*reject*)

   }

참고: 몬테칼로 시뮬레이션
Gibbs Sampling

Gibbs sampling

An iterative adaptive scheme that generates a single value for the conditional distribution for each component of the random vector X, rather than all values of the variables at the same time.

$\mathbf{X} = X_1, X_2, ..., X_K$: a random vector of K components

Assume we know $P(X_k \mid \mathbf{X}_{-k})$, where $\mathbf{X}_{-k} = X_1, X_2, ..., X_{k-1} X_{k+1}, ..., X_K$

Gibbs sampling algorithm (Gibbs sampler)

1. Initialize $x_1(0), x_2(0), ..., x_K(0)$.
2. $i \leftarrow 1$

   $x_1(1) \sim P(X_1 \mid x_2(0), x_3(0), x_4(0), ..., x_K(0))$

   $x_2(1) \sim P(X_2 \mid x_1(1), x_3(0), x_4(0), ..., x_K(0))$

   $x_3(1) \sim P(X_3 \mid x_1(1), x_2(1), x_3(0), ..., x_K(0))$

   …

   $x_k(1) \sim P(X_k \mid x_1(1), x_2(1), ..., x_{k-1}(1), x_{k+1}(0), x_K(0))$

   …

   $x_K(1) \sim P(X_K \mid x_1(1), x_2(1), ..., x_{K-1}(1))$
3. If (termination condition not met), then $i \leftarrow i+1$ and go to step 2.

요약

- 변별모델과 생성모델
  - 변별모델은 주어진 x에 대한 y의 조건부 확률을 학습
  - 데이터의 확률 분포를 학습
  - GAN은 변별모델을 이용하여 생성모델을 학습하려는 시도
- 정보이론과 상대 엔트로피
  - 엔트로피는 물리 시스템의 혼돈(disorder)의 정도를 재는 척도
  - 상대 엔트로피(relative entropy)는 같은 확률 변수에 대한 두 개의 확률 분포의 차이를 측정

질문

- 변별 모델과 생성 모델의 차이점은 무엇인가? 각각의 장단점은 무엇인가? 그 활용 분야는 어떻게 다른가?
- 대립학습모델(Adversarial Learning) 은 생성 모델의 장점을 살리되 그 단점을 변별 모델로 극복하는 방법으로 볼 수 있다. 어떤 점에서 이 해석이 가능한가?
- 교차 엔트로피, 상대 엔트로피, 상호 정보를 정의하시오. 이들간의 관계를 설명하시오. 목적함수로 이 측도들을 사용하는 머신러닝/딥러닝 방법을 하나씩 예로 드시오.