심층학습 5장. 기계 학습의 기초 (5.5장~5.9장)

송선영

5.5 최대가능도 추정

- 최대가능도 원리 (maximum likelihood, ML)

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\text{모형}}(\mathbb{X};\boldsymbol{\theta}), \tag{5.56}$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{모형}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}). \tag{5.57}$$

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\text{모형}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}). \tag{5.58}$$

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim \hat{p}_{\text{자료}}} \log p_{\text{모형}}(\boldsymbol{x};\boldsymbol{\theta}). \tag{5.59}$$

$$D_{\mathrm{KL}}(\hat{p}_{\text{자료}} \| p_{\text{모형}}) = \mathbb{E}_{\mathbf{x}\sim \hat{p}_{\text{자료}}} [\log \hat{p}_{\text{자료}}(\boldsymbol{x}) - \log p_{\text{모형}}(\boldsymbol{x})]. \tag{5.60}$$

$$- \mathbb{E}_{\mathbf{x}\sim \hat{p}_{\text{자료}}} [\log p_{\text{모형}}(\boldsymbol{x})]. \tag{5.61}$$

5.51 조건부 로그가능도와 평균제곱오차

- 최대가능도 추정량 x 가 주어졌을 때 y 를 예측하기 위한 조건부 확률 $P(y|x;\theta)$를 추정하는 것으로 일반화하는 것은 자주 쓰임

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X};\boldsymbol{\theta}). \tag{5.62}$$

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)};\boldsymbol{\theta}). \tag{5.63}$$

- 예제: 최대가능도로서의 선형회귀
  - $p(y|x) = \mathcal{N}(y; \hat{y}(x;w), \sigma^2)$

$$\sum_{i=1}^{m} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)};\boldsymbol{\theta}) \tag{5.64}$$

$$= -m\log\sigma - \frac{m}{2}\log(2\pi) - \sum_{i=1}^{m} \frac{\|\hat{y}^{(i)} - y^{(i)}\|^2}{2\sigma^2}. \tag{5.65}$$

$$\mathrm{MSE}_{\text{훈련}} = \frac{1}{m}\sum_{i=1}^{m} \|\hat{y}^{(i)} - y^{(i)}\|^2 \tag{5.66}$$

5.52 최대가능도의 여러 성질

- 최대가능도 추정량의 특징
  - 견본 개수가 m이 무한대에 접근함에 따라 점근적으로 최상의 추정량임을 증명할 수 있음
    - 여기서 '최상 ' 이란?
      - **m의 증가에 따른 수렴률을** 기준으로 함
  - 즉, 어떤 한 매개변수의 최대가능도 추정값은 훈련 견본 개수가 무한대에 접근함에 따라 그 매개변수의 참값으로 수렴함

  - 몇 가지 조건들이 충족되면 최대가능도 추정량은 일치성을 가짐 **(일치 추정량)**
    - 조건1. 진 분포 $P_{\text{자료}}$ 가 반드시 모형족인 $P_{\text{모형}}$ 에 속해야 함 (= 그 어떤 추정량으로도 $P_{\text{자료}}$를 복원할 수 없어야 함)
    - 조건2. 진 분포 $P_{\text{자료}}$는 반드시 θ의 정확한 하나의 값에 대응되어야 함

  - 일치 추정량이라도 통계적 효율성은 서로 다를 수 있음
    - 즉, 고정된 개수의 표본들에 대한 일반적 오차가 서로 다를 수 있음
    - 즉, 같은 수준의 일반화 오차를 얻는데 필요한 견본 개수가 서로 다를 수 있음

  - 통계적 효율성
    - 매개변수적 사례들에서 고찰함
    - 추정값이 매개변수의 참값과 얼마나 가까운지는 기대 평균오차로 파악할 수 있음
    - 이런 매개변수 평균제곱오차는 m이 증가함에 따라 감소함
    - 그리고 m이 클 때, 크라메르-라오 하계(Cramer-Rao lower bound)에 따르면, 일치 추정량 중 최대가능도 추정량보다 MSE가 낮은 것은 없음

5.6 베이즈 통계학

- 최대가능도 추정과 베이즈 추정의 차이
  - **차이점1**
    - 최대가능도 추정: θ의 점 추정값 하나를 이용해서 예측
    - 베이즈 추정: θ에 관한 분포 전체를 이용해서 예측

$$p(x^{(m+1)} \mid x^{(1)}, \ldots, x^{(m)}) = \int p(x^{(m+1)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) d\boldsymbol{\theta}. \quad (5.68)$$

      - 빈도론자 접근 방식에서, θ의 점 추정값의 불확실성은 분산을 이용하여 측정함
      - 추정량의 분산: 관측된 자료에서 서로 다른 표본들을 추출했을 때 그 추정량이 얼마나 달라질 것인지를 나타내는 측도
      - 베이즈 통계학에서는 추정량의 불확실성을 적분을 통해서 해결 → 과대적합을 잘 방지하는 경향이 있음

  - **차이점2**
    - 베이즈 접근 방식에서는 사전분포가 예측에 기여함
      - 훈련 자료가 제한적일 경우, 최대가능도 접근 방식보다 베이즈 접근 방식이 훨씬 잘 일반화됨
      - 훈련 견본이 많을 때는 계산 비용이 커진다는 단점이 있음

5.6 베이즈 통계학

- **예제1: 베이즈 방식의 선형회귀**
  - 베이즈 추정 접근 방식에서는 선형회귀 매개변수를 어떻게 학습하는지?
  - 선형회귀에서 학습하고자 하는 것: 입력 벡터 $x \in \mathbb{R}^n$ 을 스칼라 $y \in \mathbb{R}$ 의 예측값으로 사상하는 일차함수
    (이 예측 함수의 매개변수: 벡터 w $\in \mathbb{R}^n$)

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}. \tag{5.69}$$

$$\hat{\boldsymbol{y}}^{(\text{훈련})} = \boldsymbol{X}^{(\text{훈련})} \boldsymbol{w}. \tag{5.70}$$

$$p(\boldsymbol{y}^{(\text{훈련})} \mid \boldsymbol{X}^{(\text{훈련})}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}^{(\text{훈련})}; \boldsymbol{X}^{(\text{훈련})}\boldsymbol{w}, \boldsymbol{I})]$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{y}^{(\text{훈련})} - \boldsymbol{X}^{(\text{훈련})}\boldsymbol{w})^\top (\boldsymbol{y}^{(\text{훈련})} - \boldsymbol{X}^{(\text{훈련})}\boldsymbol{w})\right). \tag{5.72}$$

- 모형 매개변수 벡터 w에 관한 사후분포를 구하려면 먼저 사전분포를 지정해야 함

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Lambda}_0^{-1}(\boldsymbol{w} - \boldsymbol{\mu}_0)\right). \tag{5.73}$$

5.6 베이즈 통계학

- 사전분포를 구한 후에는 사후분포를 구해야 함

$$p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w}) \tag{5.74}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^{\top}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})\right)\exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu}_0)^{\top}\boldsymbol{\Lambda}_0^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_0)\right) \tag{5.75}$$

$$\propto \exp\left(-\frac{1}{2}\left(-2\boldsymbol{y}^{\top}\boldsymbol{X}\boldsymbol{w}+\boldsymbol{w}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{w}+\boldsymbol{w}^{\top}\boldsymbol{\Lambda}_0^{-1}\boldsymbol{w}-2\boldsymbol{\mu}_0^{\top}\boldsymbol{\Lambda}_0^{-1}\boldsymbol{w}\right)\right). \tag{5.76}$$

- $A_m = (X^T X + A_0^{-1})^{-1}, \quad \mu_m = A_m(X^T y + A_0^{-1}\mu_0)$ 으로 정의할 수 있음.

$$p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu}_m)^{\top}\boldsymbol{\Lambda}_m^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_m)+\frac{1}{2}\boldsymbol{\mu}_m^{\top}\boldsymbol{\Lambda}_m^{-1}\boldsymbol{\mu}_m\right) \tag{5.77}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu}_m)^{\top}\boldsymbol{\Lambda}_m^{-1}(\boldsymbol{w}-\boldsymbol{\mu}_m)\right). \tag{5.78}$$

5.61 최대 사후확률(MAP) 추정

- 원리가 있는 접근 방식들은 대부분 매개변수 θ에 관한 베이즈 사후확률 전체를 이용해서 예측을 수행함.
- 하지만 점 추정값 하나만 구하는 것이 바람직할 때도 많음
  - 최대가능도 접근 방식으로 점 추정값을 구하는 대신, 사전분포가 점 추정값의 선택에 영향을 미치게 함으로써 베이즈 접근 방식의 일부 장점을 취할 수 있음
    → 이 때 합리적인 방법이 **최대 사후 확률(MAP)** 점 추정값을 사용하는 것
    - MAP 추정에서는 사후확률이 가장 큰 점을 선택함

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}). \qquad (5.79)$$

5.7 지도 학습 알고리즘

- 지도 학습 알고리즘
  - 특정 입력을 특정 출력에 연관시키는 방법을 인력 견본 모음 x와 출력 모음 y로 이루어진 하나의 훈련집합으로부터 배우는 알고리즘
  - y의 출력 사례들을 자동으로 수집하기가 어려워서 사람, 즉 '지도 교사(supervisor)'가 출력들을 제공하는 경우가 많음
  - 하지만 훈련 집합의 목표들을 사람의 개입 없이 자동으로 수집하는 경우에도 '지도'학습 이라는 용어를 사용함

5.71 확률적 지도 학습

- 분포 $p(y|x;\theta)$들의 한 매개변수적 모임에 대해 최상의 매개변수 벡터 θ를 최대 가능도 추정을 이용해서 확률분포를 추정할 수 있음
- 선형회귀가 매개변수적 확률분포 모임(분포족)에 대응됨

$$p(y \mid \boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(y;\boldsymbol{\theta}^\top \boldsymbol{x}, \boldsymbol{I}). \tag{5.80}$$

- 이진 변수에 관한 분포의 평균은 항상 0과 1 사이여야 함
- 이 문제를 해결하기 위해 로그 S자형 함수를 이용해서 일차함수의 출력을 구간 (0,1)로 압축하고 그 구간의 값을 확률로 사용
- 이런 방식을 **로지스틱 회귀**라 부름

$$p(y = 1 \mid \boldsymbol{x};\boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}). \tag{5.81}$$

5.72 서포트 벡터 머신 (SVM)

- 서포트 벡터 머신 (SVM)
  - 지도 학습에 대한 가장 영향력 있는 접근 방식 중 하나
  - 일차함수 $w^T x + b$ 가 학습을 주도한다는 점에서 로지스틱 회귀와 비슷하지만, 로지스틱 회귀와는 달리 확률을 제공하지 않음
  - 주어진 입력이 속한 부류만 알려줌
    - $w^T x + b > 0$: SVM은 양성 부류(positive class)에 있다는 예측 결과 제시
    - $w^T x + b < 0$: SVM은 음성 부류(negative class)에 있다는 예측 결과 제시

  - 서포트 벡터 머신은 **핵 요령**을 사용함
    - 핵 요령: 여러가지 기계 학습 알고리즘을 전적으로 견본들의 내적으로만 표현할 수 있다는 통찰에 기초

$$\boldsymbol{w}^\top \boldsymbol{x} + b = b + \sum_{i=1}^{m} \alpha_i \boldsymbol{x}^\top \boldsymbol{x}^{(i)}. \qquad (5.82)$$

    - x를 주어진 특징함수 $\phi(x)$의 출력으로 대체하고, 내적을 함수 $k\left(x, x^{(i)}\right) = \phi(x) \cdot \phi(x^{(i)})$ 로 대체할 수 있음
    - 이 때, $\phi(x) \cdot \phi(x^{(i)})$ 를 핵(kernel) 이라고 부름

$$f(\boldsymbol{x}) = b + \sum_i \alpha_i k(\boldsymbol{x}, \boldsymbol{x}^{(i)}). \qquad (5.83)$$

5.72 서포트 벡터 머신 (SVM)

- 가장 흔히 쓰이는 핵 함수인 **가우스 핵**

$$k(\boldsymbol{u},\boldsymbol{v}) = \mathcal{N}(\boldsymbol{u}-\boldsymbol{v};0,\sigma^2\boldsymbol{I}). \qquad (5.84)$$

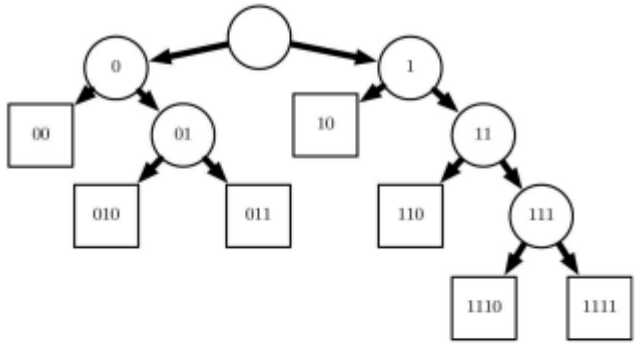
- $\mathcal{N}(x;\mu,\Sigma)$는 표준 정규 밀도임. 이 핵을 방사상 기저함수 핵이라고도 부름

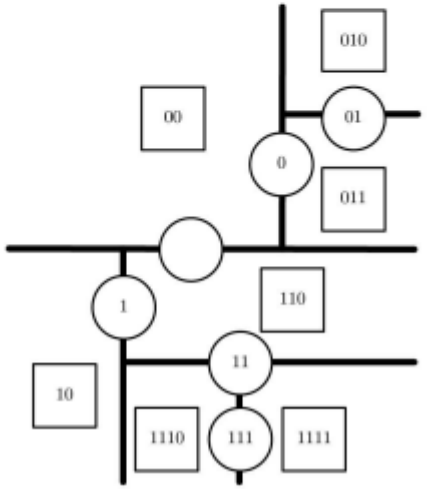
- 서포트 벡터 머신 외에 다른 여러 선형 모델들도 핵 요령 알고리즘으로 개선할 수 있음
  - 핵 요령을 사용하는 알고리즘을 통칭해 **핵 기계** 또는 **핵법** 알고리즘이라고 부름

- 핵 기계의 단점
  - 결정함수의 평가 비용이 훈련 견본 개수에 정비례해서 증가
  - 자료 집합이 클 때 훈련의 계산 비용이 높아짐

5.73 그 밖의 간단한 지도 학습 알고리즘

- K-최근접 이웃 회귀
- 결정 트리





5.8 비지도 학습 알고리즘

- 비지도 학습 알고리즘
  - 어떤 지도, 지시도 받지 않고 그냥 '특징 ' 들만 경험해서 뭔가를 배움
  - 사람이 견본들에 일일이 이름표를 부여해 주지 않아도 분포로부터 정보를 최대한 뽑아내려는 학습 알고리즘

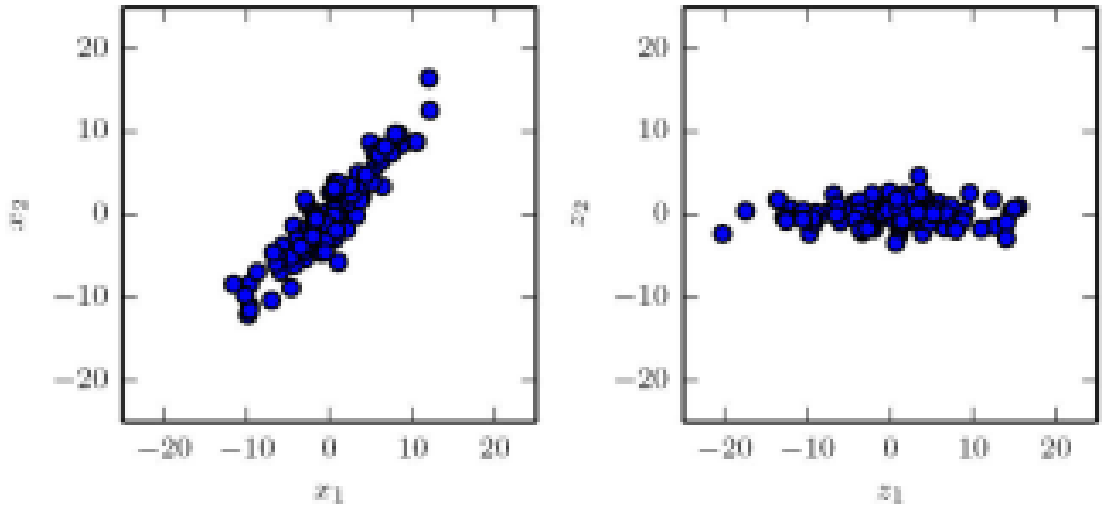
  - 밀도 추정, 분포에서 표본을 추출하는 방법의 학습, 분포에서 얻은 자료의 잡음 제거 방법 학습, 자료가 근처에 있는 다양체 찾기, 서로 연관된 견본들로 무리 짓기 와 같은 응용에 연관됨

  - 비지도 학습 알고리즘에서의 '최상 ' 의 의미는 x에 관한 정보를 최대한 유지하는 표현을 최상의 표현이라고 함.

5.81 주성분분석

- 주성분분석(PCA) 알고리즘
  - 자료를 배우는 비지도 학습 알고리즘으로 사용 가능

  - PCA는 원래의 입력보다 차원이 낮은 표현을 학습함
  - PCA는 성분들 사이에 선형 관계가 없는 표현을 학습함



5.81 주성분분석

- PCA 표현이 원래의 자료 표현 X 의 상관관계를 제거함
  - $m \times n$ 설계 행렬 X
  - 자료의 평균은 0이고, 즉 $\mathbb{E}[x] = 0$ 이라고 가정

$$\mathrm{Var}[\boldsymbol{x}] = \frac{1}{m-1}\boldsymbol{X}^\top \boldsymbol{X}. \tag{5.85}$$

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^\top. \tag{5.86}$$

$$\boldsymbol{X}^\top \boldsymbol{X} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top. \tag{5.87}$$

$$\mathrm{Var}[\boldsymbol{x}] = \frac{1}{m-1}\boldsymbol{X}^\top \boldsymbol{X} \tag{5.88}$$

$$= \frac{1}{m-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top \tag{5.89}$$

$$= \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top \tag{5.90}$$

$$= \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top. \tag{5.91}$$

$$\mathrm{Var}[\boldsymbol{z}] = \frac{1}{m-1}\boldsymbol{Z}^\top \boldsymbol{Z} \tag{5.92}$$

$$= \frac{1}{m-1}\boldsymbol{W}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{W} \tag{5.93}$$

$$= \frac{1}{m-1}\boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top \boldsymbol{W} \tag{5.94}$$

$$= \frac{1}{m-1}\boldsymbol{\Sigma}^2. \tag{5.95}$$

5.82 k-평균 군집화

- 단순 표현 학습 알고리즘의 또 다른 예로 k-평균 군집화 (k-means clustering)이 있음

- 평균 군집화 알고리즘
  - 훈련 집합의 견본들을 서로 가까이 있는 것들끼리 모아서 k개의 서로 다른 군집으로 분할
  - 입력 x를 나타내는 k차원 원핫 부호 벡터 h를 제공한다고 생각할 수 있음.
    - 표현 h에서 만일 x가 클러스터 i 에 속하면 h_i = 1이고 나머지 모든 성분은 0임

  - k-평균 군집화가 제공하는 원핫 부호 벡터는 모든 입력에 대해 성분들의 대다수가 0이라는 점에서 희소 표현에 해당함

  - 원핫 표현보다 더 좋은 것이 분산 표현

Thank You

감사합니다.