불확실한 세상에서 인공지능은 어떻게 판단할까?
: 인공지능의 확률적 접근

뇌공학 B조

July 25, 2025

1 서론: 인공지능의 본질과 직면 과제

Machine Learning(ML)은 입력값에 대한 함숫값 $f(x)$에 매핑하는 함수 f를 찾는 것을 목표로 한다. 예컨대, 시력을 예측하는 인공지능 모델은 컴퓨터 사용 시간이나 부모님의 시력 등의 데이터(즉, 입력값)에 대하여 시력을 예측하는 함수를 찾으며, LLM은 사용자가 작성한 프롬프트(이 역시 입력값)에 대하여 만족할만한 대답을 매핑하는 함수를 찾는다. 비단, 우리가 조사한 데이터에는 잡음(noise)이 존재하여 완전한 함수를 찾을 수 없다는 문제가 존재한다. 시력을 예측하는 인공지능 모델을 만들기 위해 피험자들의 컴퓨터 사용 시간과 시력을 조사한 예를 들어보자. 시력 측정 방식이 정확하지 아니하거나 우리가 고려하지 못한 외의 조건(e.g., 눈병 여부)으로 인해 관측값에 잡음이 생긴다. ML을 확률적으로 접근할 경우 이러한 잡음을 확률 분포로 모델링 할 수 있으며, 본 기사에서 소개하는 MLE 또한 확률 모델로 잡음을 확률 분포로 취급한다(Section 3, 참고: Figure 1).

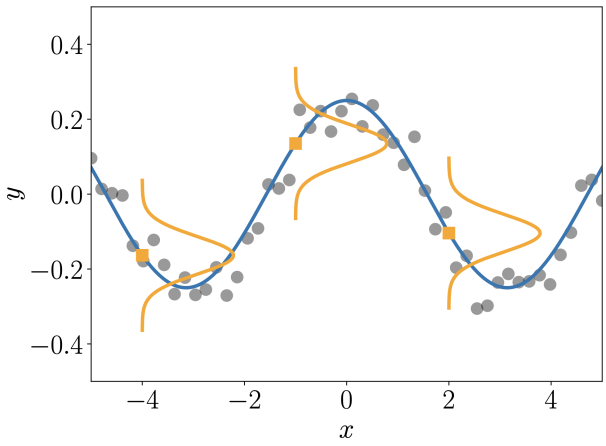

Figure 1: ML의 확률적 접근
©Mathematics for Machine Learning[3]

최적의 함수를 찾는 문제는 '함수가 찾은 값과 실제 값의 차이(또는 오차, 손실)를 최소화하는 문제'와 동일하다. 일반적으로 이러한 오차를 함수로 나타내며 이를 loss function이라 부른다. ML에서 최적의 함수를 찾아가는 것은 loss function을 줄여나가는 '학습'을 의미하는데, ML을 확률적으로 접근하게 되면 이는 우도를 최대화하거나 negative log-likelihood를 최소화하는 문제로 취급 가능하다(Section 2.2, Section 3).

잡음을 확률 분포로 모델링하는 관점과 동일하게 오차값 또한 확률 분포로 모델링 가능하다. 일반적으로 ML은 특정한 함수의 꼴을 정해 두고 모수를 바꾸며 최적의 함수를 찾아간다(Section 2.1). 위

예에서 컴퓨터 사용 시간과 시력은 서로 반비례 관계가 있다고 판단하여 특정한 함수의 꼴을 일차방정식과 같은 직선의 꼴로 정해두었다고 하자. 실제로 이 둘의 관계가 직선의 꼴로 표현되지 않는다면 필연적으로 오차가 발생하게 된다. 이러한 이유가 아니어도, ML을 실생활에 응용 시 잡음이 필연적으로 발생하여 현실을 완벽히 대변하는 함수는 찾을 수 없다. 이러한 이유로 오차는 확률 분포로 모델링하는 것이 바람직하다. 예컨대, 하루 평균 컴퓨터를 7시간 사용하는 사람에 대하여 '시력이 1.0일 것이다'라고 함숫값을 예측하는 것이 아닌 시력의 확률 분포를 예측한다.

본 기사에서는 ML을 확률적으로 접근하였을 때 함수를 찾는 대표적인 방법 MLE(Section 3)와 현재도 대표적으로 사용되는 loss function인 MSE와 MAE의 숨겨진 확률적 요소를 소개한다(Section 4).

2 확률 이론: 오차 모델링의 열쇠

2.1 모수(parameter)

ML은 최적의 함수를 찾으며, 그 함수는 모수를 포함한다. 예컨대 선형 모델을 이용하는 선형회귀의 경우 $f(x) = x^T\theta$를 함수로 사용하며(여기서 선형은 $x^T\theta$꼴로 표현 되었으며, 이의 요소인 x까지 선형적인 변환일 필요는 없다.) θ는 모델을 학습하며 구해야하는 모수이다. 즉, 모델을 학습하는 행위는 적절한 모수를 찾아 최적의 함수를 찾아가는 과정이다.

2.2 우도(likelihood)

Bayes' theorem은 다음과 같다:

$$\underbrace{p(\theta \mid y)}_{\text{posterior}} = \frac{\overbrace{p(y \mid \theta)}^{\text{likelihood}} \cdot \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(y)}_{\text{evidence}}} \tag{1}$$

좌변을 posterior, 우변 분자의 첫 p를 likelihood, 두번째 p를 prior이라고 부른다. 해석하면, prior은 y를 관측하기 전 지식, likelihood(우도)는 y를 관측하여 학습한 값($p(\theta \mid y)$와 $p(y \mid \theta)$의 혼동을 피하기 위해 likelihood(우도)는 $\mathcal{L}(\theta)$로 표기하기도 한다.), posterior은 y 관측 이후 업데이트된 지식이다. 즉, ML은 posterior을 최대화하여 y에 대한 최적의 함수의 모수 를 구하는 것을 목표로 한다.

3 최대 우도 추정 (MLE): 이론적 기반

3.1 MLE

MLE(Maximum Likelihood Estimation, 최대 우도 추정법)는 우도를 최대화하는 방법으로, 결과적으로 우도와 비례 관계를 가지는 posterior을 최대화할 수 있다. 즉, MLE는 수식으로 다음과 같이 표현 가능하다:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathcal{L}(\theta) \tag{2}$$

예컨대, ML의 모델을 $y = x^T\theta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$으로 정의하자($\epsilon$은 잡음). 관측한 데이터 — 각각은 서로 영향을 주지 않으며(독립적), 같은 특성을 가진다(동일하게 분포)고 가정하자. 이를 independent and identically distributed, 줄여서 i.i.d.라고 한다. — $x_1, x_2, \ldots, x_n$과 $y_1, y_2, \ldots, y_n$에 대한 집합을

각각 dataset $\mathcal{X}$, $\mathcal{Y}$로 표현하면, 우도는 다음과 같다:

$$\mathcal{L}(\theta \mid \mathcal{Y}, \mathcal{X}) = p(\mathcal{Y} \mid \mathcal{X}, \theta)[1] = \prod_{i=1}^{n} \mathcal{N}(y_i \mid x_i^T \theta, \sigma^2) \tag{3}$$

이전에 예로든 시력 예측 인공지능에 대입하면, 각 x_i는 피험자 개개인의 컴퓨터 사용시간이고 y_i는 시력이다. 인공지능을 학습시키면, 모수 θ가 결정되어 시력을 예측하는 함수 $y = x^T\theta$가 만들어진다.

MLE는 최댓값의 인덱스를 구하는 문제로, 마이너스를 취해 최솟값을 구하는 문제(최솟값이 되는 인덱스는 loss function을 최소화하게 한다.)로 변환 가능하다. 또한 최댓값(또는 최솟값)의 인덱스가 동일하다면 동일한 문제로 취급되어 상수를 더하거나 빼고, log를 취하는 등의 연산이 가능하다. 이러한 이유로 MLE는 우도에 log를 취해 미분하는 방법(미분을 통해 최솟값의 인덱스를 찾는다.)으로 구할 수 있다.

3.2 MAP

만약 표본이 적다면 어떤 일이 발생할까? 예컨대, 피험자가 4명이고 이 중 한명은 하루 평균 15시간 컴퓨터를 사용하지만 시력이 2.0이다. 이 경우 함수는 현실을 대변하지 못하며, 표본을 더 모집하여 추가 학습을 통해 개선 가능하다. 모두가 '컴퓨터를 오래하면 시력이 나쁘지 아니할까?'라고 생각할 것이다. 이러한 '믿음'은 학습전 지식, 즉 prior로 모델링 가능하다(컴퓨터 사용시간과 시력에 반비례한 함수에 기반한 분포로 prior를 정의한다.).

MAP(Maximum A Posteriori Estimation)는 다음과 같이 표현 가능하다:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta \mid y) = \arg\max_{\theta} \mathcal{L}(\theta) p(\theta) \tag{4}$$

MAP는 학습이 진행됨에 따라 초기에는 prior의 지식이 강하게, 후에는 새로운 학습된 우도의 지식이 강하게 작동하는 특성을 가진다.

4 MSE와 MAE의 확률적 해석

인공지능 모델이 내는 오차 ($y_i - \hat{y}_i$, 즉 실제 값과 예측 값의 차이) 가 어떤 특정한 확률 분포를 따른다고 가정해보자. 이렇게 오차 분포를 정해두면, 그 분포에 가장 잘 맞는 모델을 찾는 것 (통계에서는 '우도'를 가장 높인다고 표현한다) 이 특정 '오차 지표'를 가장 낮추는 것과 같은 의미가 된다. RMSE와 MAE가 바로 이런 방식으로 설명될 수 있는 대표적인 오차 지표이다.

- MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: mean-squared error
- RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$: root-mean-squared error[2]
- MAE $= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$: mean absolute error

4.1 정규 (Gaussian) 오차의 경우: RMSE/MSE가 최적인 이유

만약 모델의 예측 오차들이 정규 분포 (가우시안 분포) 를 따르고, i.i.d.를 만족한다고 가정해 보자. 이 경우, 실제 y_i는 모델의 예측 값 $\hat{y}_i = f(\theta, x_i)$ 를 중심으로 정규 분포 모양으로 퍼져 있을 것이라고 본다.

[1] $p(\mathcal{Y}, \mathcal{X} \mid \theta)$ 보다 $p(\mathcal{Y} \mid \mathcal{X}, \theta)$는 맥락을 고려한 표현임.
[2] RMSE는 MSE와 다르게 y_i와 동일한 척도(scale)와 단위(unit)를 가지는 정규화된 값임

이런 상황에서, 모든 데이터에 대한 우도와 로그 우도 (log-likelihood) 를 식으로 나타내면 다음과 같다.

$$\mathcal{L}(\theta, \sigma \mid \mathcal{Y}, \mathcal{X}) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid f(\theta, x_i), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{i=1}^{n} \exp\left[-\frac{(y_i - f(\theta, x_i))^2}{2\sigma^2}\right] \tag{5}$$

$$\log \mathcal{L} = -n \log \sigma - \frac{n}{2} \log 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(\theta, x_i))^2 \tag{6}$$

이 식에서 $\log \mathcal{L}$ 값을 가장 크게 (최대) 만들려면, 맨 마지막에 있는 $\sum_{i=1}^{n}(y_i - f(\theta, x_i))^2$ 부분을 가장 작게 (최소) 만들어야 한다. 이 값은 오차를 제곱해서 모두 더한 것과 같다. 여기에 데이터 개수 n으로 나누면 바로 평균 제곱 오차 (MSE) 가 된다.

결국, 오차가 정규 분포를 따른다고 가정하면, MSE(또는 RMSE) 를 가장 낮추는 모델이 통계적으로 가장 가능성이 높은 모델이 되는 것이다. 반대로, MSE를 사용하는 것은 오차가 정규 분포를 따른다는 가정이 암묵적으로 함유된 것이다. 이것이 바로 '최소 제곱법'이라는 통계 기법의 바탕이 된다.

4.2 라플라스 (Laplace) 오차의 경우: MAE가 최적인 이유

이번에는 모델의 예측 오차가 라플라스 분포를 따른다고 가정해보자. 라플라스 분포는 정규 분포와는 다른 모양을 가지는데, 특히 특이하게 큰 오차 (이상치, outlier) 가 나와도 그 영향이 덜하다.

오차가 라플라스 분포를 따른다고 했을 때의 우도와 로그 우도는 다음과 같다.

$$\mathcal{L}(\theta, \sigma \mid \mathcal{Y}, \mathcal{X}) = \frac{1}{2b} \exp\left[-\frac{|y - f(\theta, x_i)|}{b}\right] : \text{라폴라스 분포의 PDF} \tag{7}$$

$$\log \mathcal{L} = -n \log 2b - \frac{1}{b} \sum_{i=1}^{n} |y_i - f(\theta, x_i)| \tag{8}$$

마찬가지로, $\log \mathcal{L}$ 을 최대로 만들려면, 마지막 항인 $\sum_{i=1}^{n} |y_i - f(\theta, x_i)|$ 를 가장 작게 만들어야 한다. 이 값은 오차의 절댓값을 모두 더한 것과 같다. 여기에 데이터 개수 n으로 나누면 바로 평균 절대 오차 (MAE) 가 된다.

즉, 오차가 라플라스 분포를 따른다고 가정하면, MAE를 가장 낮추는 모델이 통계적으로 가장 가능성이 높은 모델이 된다. 반대로, MAE를 사용하는 것은 오차가 라플라스 분포를 따른다는 가정이 암묵적으로 함유된 것이다. 이러한 MAE의 특징 때문에, MAE를 최소화하는 방식은 데이터에 특이값이 있어도 잘 작동하는 '강건한(robust) 회귀' 방법으로 불린다.

4.3 그 외의 경우: 복잡한 오차 분포와 대안

실제 인공지능 모델의 오차는 항상 정규 분포나 라플라스 분포처럼 깔끔하지 않다. 때로는 오차의 크기가 입력값에 따라 달라지거나 (이분산성), 특정 값이 유난히 많거나 (예: 0이 빈번하게 나타나는 경우), 시간이 지남에 따라 오차가 서로 영향을 주는 (자기 상관) 등 복잡한 패턴을 보인다. 이럴 때는 단순히 RMSE나 MAE만으로는 모델 성능을 정확히 평가하기 어렵다.

이런 복잡한 상황에 대해 몇 가지 대안이 제시된다:

1. 데이터 변환 (Transformation): 데이터를 로그를 취하거나 다른 수학적 변환을 통해 오차 분포가 정규 분포와 비슷해지도록 만든 후 RMSE 등을 적용한다.

2. 강건한 추정 방법 (Robust inference): 이상치에 덜 민감한 다른 통계 기법(예: MAD, Median Absolute Deviation)을 사용한다.

3. 우도 기반 추론 (Likelihood-based inference): 가장 강력하고 유연한 방법이다. 여러 오차 분포를 조합하거나, 문제 특성에 맞는 복잡한 오차 분포를 직접 모델링하여 '우도 함수'를 새롭게 만들고 이를 통해 모델을 평가한다. 이는 RMSE나 MAE보다 오차의 복잡한 특성을 더 정확하게 반영할 수 있다.

5 결론: 인공지능과 확률적 사고의 미래

본 기사에서는 인공지능의 확률적 사고 방법을 다루며 함수를 찾는 대표적인 방법 MLE를 살펴보고 손실함수 RMSE와 MAE의 확률적 통찰을 하였다.

이 논의를 통해 우리는 인공지능 모델 평가에 쓰이는 RMSE와 MAE가 단순히 사용되는 지표가 아니라, 예측 오차가 어떤 확률 분포를 따를 것이라는 통계적 가정(정규 분포 또는 라플라스 분포)에 기반을 둔 것임을 알게 되었다. 즉, 어떤 오차 지표를 선택할지는 우리 모델의 오차가 어떤 특성을 가질 것이라고 '믿는지'에 따라 달라진다.

"어떤 오차 지표가 무조건 최고인가?"라는 질문에는 정답이 없다. 실제 데이터의 오차는 정규 분포나 라플라스 분포처럼 단순하지 않은 경우가 많기 때문에, 단순히 RMSE와 MAE 중 하나를 고르거나 둘 다 나열하는 것만으로는 충분하지 않다.

인공지능 연구자들에게 다음 세 가지 중요한 점들이 강조된다.

1. 오차 분포 가정의 중요성: 모델을 평가하고 개선하려면, 예측 오차가 어떤 확률 분포를 가지는지 정확히 이해하고 모델링해야 한다.

2. '우도' 기반 접근의 유연성: 오차가 복잡한 패턴을 보일 때는, 단순히 RMSE/MAE가 아니라 오차의 복잡성을 직접 반영한 '우도 함수'를 설계하는 것이 더 강력한 평가 방법이 될 수 있다.

3. 다양한 지표 사용 시 주의: 여러 모델 성능 지표를 함께 쓸 때는, 각 지표가 어떤 통계적 의미를 가지는지, 그리고 서로 어떻게 관련되어 있는지 충분히 이해해야 한다.

5.1 새로운 연구 질문 및 방향

인공지능이 단순히 답만 내놓는 것이 아니라, 그 답이 얼마나 '확실한지'(불확실성) 를 어떻게 이해하고 활용할 것인가 하는 중요한 질문으로 이어진다. 베이지안 딥러닝과 같은 확률적 인공지능 분야가 이러한 방향으로 나아가고 있지만, 여전히 다음 과제들이 남아있다. 이는 미래 인공지능이 신뢰받고 책임감 있는 의사 결정을 지원하는 데 필수적이다.

- AI가 스스로 오차 분포를 이해하고 적응하도록: 인공지능 모델이 학습 과정에서 데이터로부터 최적의 오차 분포 형태를 스스로 '발견'하고, 그에 맞춰 학습 방식을 '적응적으로' 변화시키는 능력을 가질 수 있을까?
  $\longrightarrow$ Mixture Density Networks, Heteroscedastic Regression

- 불확실성을 명확히 설명하고 소통하는 AI: 의료나 자율 주행처럼 중요한 분야에서, 인공지능이 예측 값과 함께 그 예측이 얼마나 '확실한지'를 정량적으로 제시하고, 비전문가도 이해할 수 있도록 '설명'하는(Explainable Uncertainty) 방법은 무엇일까?
  $\longrightarrow$ Bayesian Deep Learning, MC Dropout, Deep Ensembles

- 불확실성을 고려한 강건한 의사결정 AI: 예측 오차가 큰 영향을 미치는 복잡한 시스템에서, 인공지능이 단순히 평균 예측값을 넘어 불확실성까지 고려하여 리스크를 최소화하고 안전성을 극대화하는 '강건한 의사결정'을 내리도록 도울 수 있을까?
  $\longrightarrow$ Risk-sensitive RL, Distributional RL, CVaR Optimization

이러한 질문들에 대한 탐구는 인공지능이 인류 문명의 복잡한 문제 해결에 기여하는 진정한 지적 파트너로 진화하는 데 중요한 기반이 될 것이다.

References

[1] Timothy O. Hodson. Root-mean-square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development*, 15(14):5481–5487, Jul 2022.

[2] Y. Anzai. *Pattern recognition and machine learning*. Elsevier Science, 2012.

[3] Marc Peter Deisenroth, Cheng Soon Ong, and Aldo A. Faisal. *Mathematics for Machine Learning*. Cambridge University Press, 2021.

[4] Kevin P. Murphy. *Machine learning: A probabilistic perspective*. The MIT Press, 2012.

[5] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.