STAT404 - 베이즈통계입문
Mini Project

평균장 방법과 변분 베이즈

Presenter: 윤민서
데이터과학과 2021320322

Scope and Purpose

- Scope
  - 평균장 방법을 통한 이미지 잡음 제거
  - 변분 베이즈를 통한 선형 회귀
- Additional Method
  - 베이지안 최적화를 통한 위 방법들에 대한 초모수 결정

Scope and Purpose

- Purpose
  - 현재 AI 분야가 빠른 속도로 발전하면서 딥러닝에 대한 관심이 매우 상승
  - 그러나 딥러닝은 컴퓨팅 리소스가 매우 많이 드는 방식으로 간단한 문제를 해결할 때에는 과한 방법
  - 베이즈 통계학을 기반으로 한 머신러닝 알고리즘으로만 구성된 모델이 간단한 문제를 해결할 수 있음을 보여주는 데 의의

Literature Review

- 평균장 방법
  - A Mean-Field Variational Inference Approach to Deep Image Prior for Inverse Problems in Medical Imaging
  - https://openreview.net/forum?id=DvV_blKLiB4
- 변분 베이즈
  - Variational Bayesian inference for linear and logistic regression
  - https://arxiv.org/abs/1310.5438
- 수식 전개
  - Machine Learning: a Probabilistic Perspective.
  - https://probml.github.io/pml-book/book0.html

Literature Review

- 평균장 방법

$$\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} \mathrm{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{w}) \,\|\, p(\boldsymbol{w})] - \mathbb{E}_{\boldsymbol{w}\sim q_{\boldsymbol{\phi}}}[\log p(\mathcal{D} \,|\, \boldsymbol{w})] \tag{2}$$

using backpropagation without weight decay. This effectively doubles the number of trainable parameters and is known as Bayes by backprop (Blundell et al., 2015). The first term in Eq. (2) is usually approximated with MC integration by

$$\mathrm{KL}[q\|p] \approx \frac{1}{T}\sum_{i=1}^{T} \log q_{\boldsymbol{\phi}}(\boldsymbol{w}_i) - \log p(\boldsymbol{w}_i) \;, \tag{3}$$

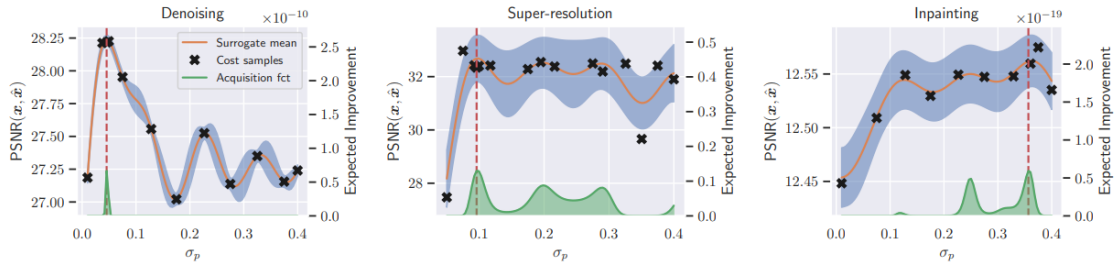


Figure 2: Results of Bayesian optimization. The acquisition function selects the next candidate for σ_p based on the maximum of the expected improvement.

Literature Review

- 변분 베이즈

2.2. **Variational Bayesian inference.** The variational posteriors are found by maximizing the variational bound

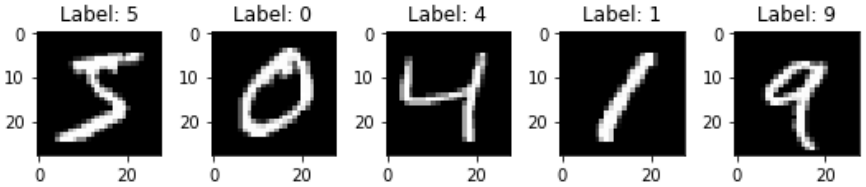
$$\mathcal{L}(\mathsf{Q}) = \iiint \mathsf{Q}(\boldsymbol{w}, \tau, \alpha) \ln \frac{\mathsf{P}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{w}, \tau)\mathsf{P}(\boldsymbol{w}, \tau|\alpha)\mathsf{P}(\alpha)}{\mathsf{Q}(\boldsymbol{w}, \tau, \alpha)} \mathrm{d}\boldsymbol{w}\mathrm{d}\tau\mathrm{d}\alpha \leq \ln \mathsf{P}(\mathcal{D}), \quad (5)$$

where $\mathsf{P}(\mathcal{D})$ is the model evidence. To maximize this bound, we assume that the variational distribution $\mathsf{Q}(\boldsymbol{w}, \tau, \alpha)$, which approximates the posterior $\mathsf{P}(\boldsymbol{w}, \tau, \alpha|\mathcal{D})$, factors into $\mathsf{Q}(\boldsymbol{w}, \tau)\mathsf{Q}(\alpha)$.

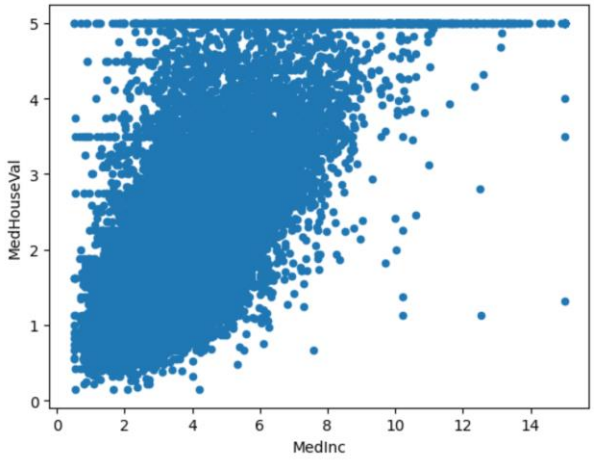
- 수식 전개

Description

- Data
  - 평균장 방법: MNIST 손글씨 데이터셋 (torchvision)



  - 변분 베이즈: California Housing 데이터셋 (sklearn)



Description

- Method - 평균장 방법

평균장 근사를 위한 접근에서는 posterior를 fully factorized approximation 형태로 가정한다

$$q(\mathbf{x}) = \prod_i q_i(\mathbf{x}_i)$$

우리의 목표는 아래의 최적화 문제를 푸는 것이다.

$$\min_{q_1,\ldots,q_D} \mathbb{KL}(q||p)$$

즉 각 주변확률분포인 q_i의 모수에 대하여 최적화한다.

변분 추론의 근본적인 목표는 다음의 하계를 최대화하는 것이다.

$$L(q) := -J(q) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})} \leq \log p(\mathcal{D})$$

이때 q_j를 제외한 나머지 변수를 상수로 취급하면

$$\begin{aligned}
L(q_j) &= \sum_{\mathbf{x}} \prod_i q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k)\right] \\
&= \sum_{\mathbf{x}_j} \sum_{\mathbf{x}_{_j}} q_j(\mathbf{x}_j) \prod_{i\neq j} q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k)\right] \\
&= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{_j}} \prod_{i\neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x}) \\
&\quad - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \sum_{\mathbf{x}_{_j}} \prod_{i\neq j} q_i(\mathbf{x}_i) \left[\sum_{k\neq j} \log q_k(\mathbf{x}_k) + q_j(\mathbf{x}_j)\right] \\
&= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) + \text{const}
\end{aligned}$$

이때

$$\log f_j(\mathbf{x}_j) := \sum_{\mathbf{x}_{\_j}} \prod_{i\neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x}) = \mathbb{E}_{-q_j}[\log \tilde{p}(\mathbf{x})]$$

위에서 유도된 식은 $\mathbf{x}_j$를 제외한 모든 잠재변수에 대하여 평균을 계산한 것과 같다. 그러므로 $L(q_j)$를 다음과 같이 쓸 수 있다.

$$L(q_j) = -\mathbb{KL}(q_j||f_j)$$

Description

- Method - 평균장 방법

앞에서 유도된 식은 $\mathbf{x}_j$를 제외한 모든 잠재변수에 대하여 평균을 계산한 것과 같다. 그러므로 $L(q_j)$를 다음과 같이 쓸 수 있다.

$$L(q_j) = -\mathbb{KL}(q_j||f_j)$$

여기에서 L을 최대화하는 것은 $q_j = f_j$인 상황에서 KL 분산을 최소화하는 것과 같다.

$$q_j(\mathbf{x}_j) = \frac{1}{Z_j}\exp\left(\mathbb{E}_{-q_j}[\log \tilde{p}(\mathbf{x})]\right)$$

이때 정규화 상수 Z_j은 무시할 수 있다. 또한 q_j이 정규분포를 따른다고 가정하면 위의 식을 다음과 같이 쓸 수 있다.

$$\log q_j(\mathbf{x}_j) = \mathbb{E}_{-q_j}[\log \tilde{p}(\mathbf{x})] + \text{const}$$

즉, q_j의 분포 함수는 변수 $\mathbf{x}_j$에 의하여 잘 정의된다. 즉 $\mathbf{x}_j$이 이산형 변수이면 q_j는 이산형 확률분포가 될 것이고, $\mathbf{x}_j$이 연속형 변수이면 q_j는 연속형 확률밀도함수가 된다.

$$p(\{\mathbf{x}_v\},\{\mathbf{z}_v\}) \propto \prod_{v\in\mathcal{V}} \Phi(\mathbf{x}_v,\mathbf{z}_v) \prod_{(u,v)\in\mathcal{E}} \Psi(\mathbf{z}_u,\mathbf{z}_v)$$

- Φ
  - Latent node embedding이 주어졌을 때, node feature vector에 대한 likelihood
- Ψ
  - 연결된 node 사이의 dependency

Description

- Method - 변분 베이즈

Prior: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}I)$

Likelihood: $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{X}\mathbf{w}, \lambda^{-1})$

이때 α와 λ에 대한 추정을 한다고 하면 아래와 같은 prior를 고려할 수 있다.

$$p(\mathbf{w}, \lambda, \alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, (\lambda\alpha)^{-1}\mathbf{I})\mathrm{Ga}(\lambda|a_0^\lambda, b_0^\lambda)\mathrm{Ga}(\alpha|a_0^\alpha, b_0^\alpha)$$

또한 posterior q를 계산하기 편한 형태로 인수분해하여 근사하는 방법을 사용할 수 있다.

$$q(\mathbf{w}, \alpha, \lambda) = q(\mathbf{w}, \lambda)q(\alpha)$$

이러한 가정 하에서 Design and Analysis of Learning Classifier Systems의 7장의 내용에 따르면 posterior 아래와 같이 계산할 수 있다.

$$q(\mathbf{w}, \alpha, \lambda) = \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \lambda^{-1}\mathbf{V}_N)\mathrm{Ga}(\lambda|a_N^\lambda, b_N^\lambda)\mathrm{Ga}(\alpha|a_N^\alpha, b_N^\alpha)$$

이때

$$\begin{aligned}
\mathbf{V}_N^{-1} &= \bar{\mathbf{A}} + \mathbf{X}^T\mathbf{X} \\
\mathbf{w}_N &= \mathbf{V}_N\mathbf{X}^T\mathbf{y} \\
a_N^\lambda &= a_0^\lambda + \frac{N}{2} \\
b_N^\lambda &= b_0^\lambda + \frac{1}{2}(||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \mathbf{w}_N^T\bar{\mathbf{A}}\mathbf{w}_N) \\
a_N^\alpha &= a_0^\alpha + \frac{D}{2} \\
b_N^\alpha &= b_0^\alpha + \frac{1}{2}\left(\frac{a_N^\lambda}{b_N^\lambda}\mathbf{w}_N^T\mathbf{w}_N + \mathrm{tr}(\mathbf{V}_N)\right) \\
\bar{\mathbf{A}} &= \frac{\alpha_N^\alpha}{b_N^\alpha}\mathbf{I}
\end{aligned}$$

여기에서 $q(\mathbf{w}, \lambda)$와 $q(\boldsymbol{\alpha})$를 번갈아가며 업데이트하는 방식으로 알고리즘이 작동한다. 또한 $\mathbf{w}$와 λ가 추정되고 나면 posterior predictive distribution은 Student distribution을 따른다. 간편함을 위해서 아래와 같이 데이터가 1개인 경우를 생각하자.

$$p(y|\mathbf{x}, \mathcal{D}) = \mathcal{T}\left(\mathbf{w}_N^T\mathbf{x}, \frac{b_N^\lambda}{a_N^\lambda}(1 + \mathbf{x}^T\mathbf{V}_N\mathbf{x}), 2a_N^\lambda\right)$$

Description

- Method - 변분 베이즈

marginal likelihood의 정확한 값을 계산하기 위해서는 아래와 같은 적분을 계산해야 한다.

$$p(\mathcal{D}) = \iiint p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \lambda) p(\mathbf{w}|\alpha) p(\lambda) d\mathbf{w} d\alpha d\lambda$$

그러나 이러한 적분은 intractable하므로 최종적인 계산은 $\log p(\mathcal{D})$의 하계 $L(q)$를 구하는 것으로 마무리한다.

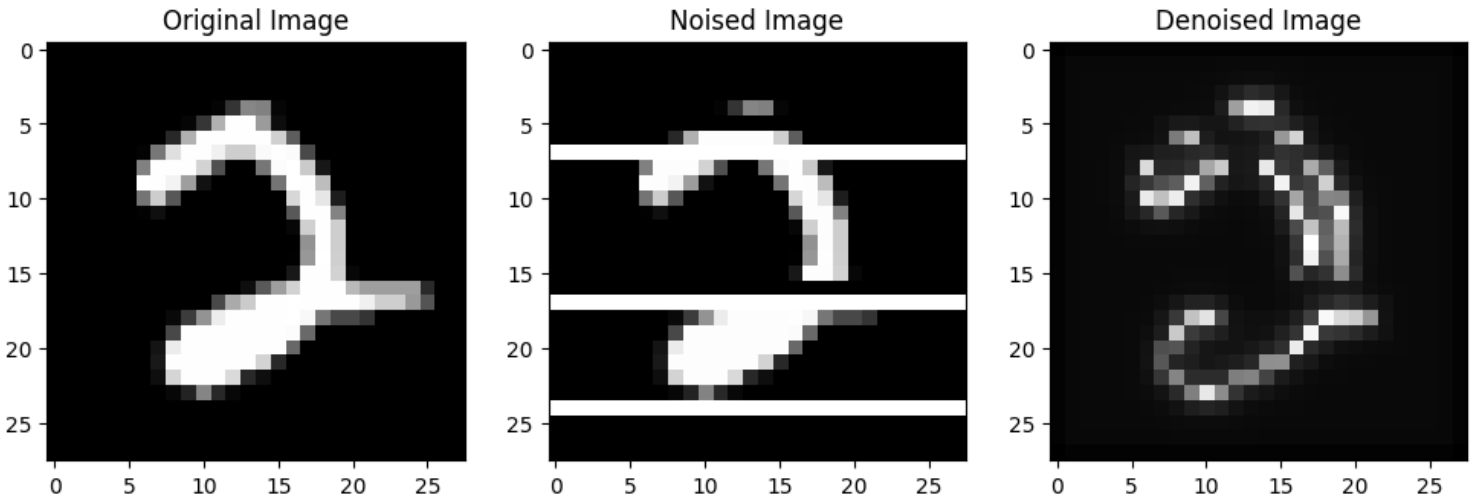
$$\begin{aligned} L(q) = & -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{a_N^\lambda}{b_N^\lambda}(y_i - \mathbf{w}_N^T\mathbf{x}_i)^2 + \mathbf{x}_i^T\mathbf{V}_N\mathbf{x}_i\right) \\ & + \frac{1}{2}\log|\mathbf{V}_N| + \frac{D}{2} \\ & - \log\Gamma(a_0^\lambda) + a_0^\lambda \log b_0^\lambda - b_0^\lambda \frac{a_N^\lambda}{b_N^\lambda} + \log\Gamma(a_N^\lambda) - a_N^\lambda \log b_N^\lambda + a_N^\lambda \\ & - \log\Gamma(a_0^\alpha) + \alpha_0^\alpha \log b_0^\alpha + \log\Gamma(a_N^\alpha) - a_N^\alpha \log b_N^\alpha \end{aligned}$$

추가로 prior를 uninformative prior로 가정하는 경우에는, 즉 $a_0^\alpha = b_0^\alpha = 0$이라고 설정하면 아래를 얻는다.

$$\bar{\alpha} = \frac{a_N^\alpha}{b_N^\alpha} = \frac{D/2}{\frac{1}{2}(\frac{a_N^\lambda}{b_N^\lambda}\mathbf{w}_N^T\mathbf{w}_N + \mathrm{tr}(\mathbf{V}_N))}$$

Discussion of Major Findings

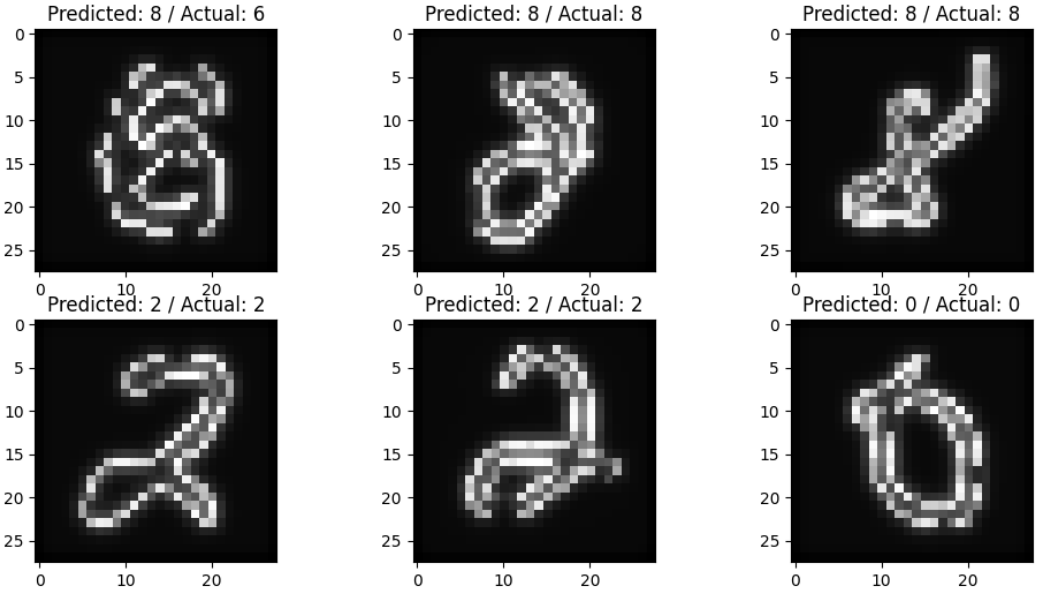
- 평균장 방법
  - Salt and Pepper 잡음을 추가한 뒤 잡음 제거



Discussion of Major Findings

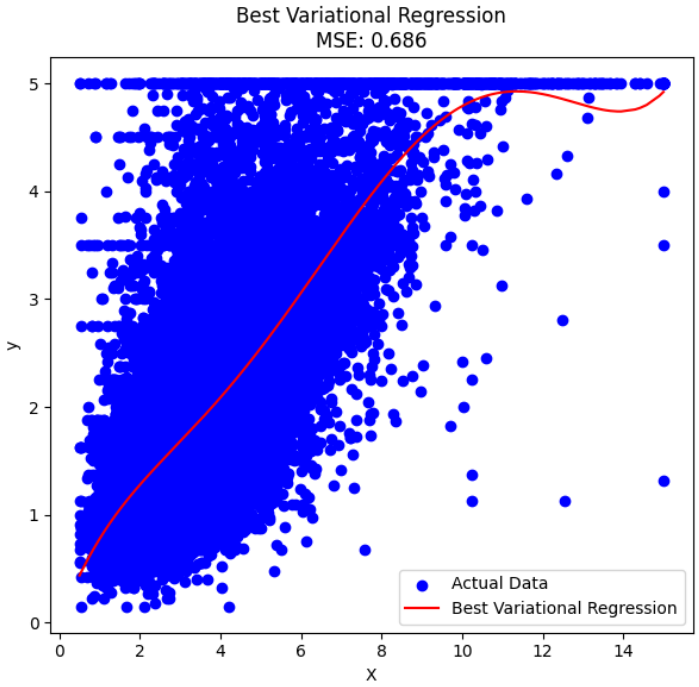
- 평균장 방법
  - 잡음이 제거된 이미지가 유효한지 알아보기 위하여 가벼운 CNN을 통해 분류 문제를 해결하는 방법으로 검증

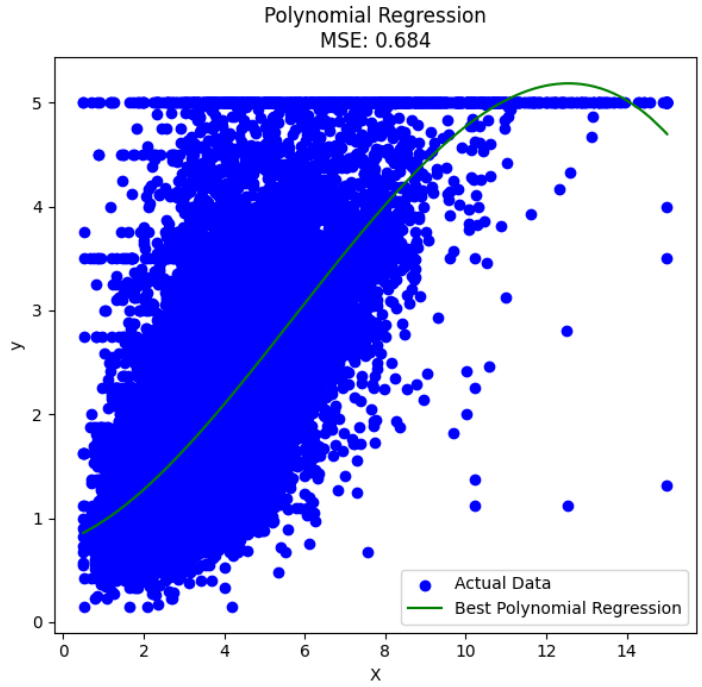
```
End of Epoch [2/2], Loss: 0.1097, Total Accuracy: 96.27%
```



Discussion of Major Findings

- 변분 베이즈
  - MSE에서 강점을 보이지는 않지만 데이터 이해도가 높음





Discussion of Major Findings

- 베이지안 최적화
  - 초모수의 개수에 따라 반복 횟수가 지수적으로 증가하는 Grid Search보다 효율적으로 탐색하는 것을 확인

53	-0.7464	64.69	58.92	1.779
54	-0.6937	86.55	19.68	5.0
55	-0.7049	85.65	18.75	3.073
56	-0.6937	67.11	61.02	5.0
57	-0.6977	71.14	60.58	4.873
58	-0.6977	71.32	64.09	4.751
59	-0.7464	72.99	62.93	1.869
60	-0.6937	68.95	62.97	5.0
61	-0.6937	66.0	64.63	5.0
62	-0.6977	68.0	66.51	4.017

Implications of The Study

- 베이즈 통계학을 기반으로 한 머신러닝 알고리즘으로만 구성된 모델이 간단한 문제를 해결할 수 있다고 결론
- Grid Search, Randomized Search 등의 방법론이 아닌 베이즈 통계학을 기반으로 한 베이지안 최적화 방법이 초모수를 결정하는 데 도움이 될 수 있음

Concluding Remark

- Conclusion
  - 베이지안 최적화를 통한 평균장 방법, 변분 베이즈 방법은 간단한 문제를 푸는 데에 효과적
  - 딥러닝 만능주의 경향에서 벗어나는 데 도움이 될 수 있는 연구

Concluding Remark

- Future work
  - 더욱 다양한 알고리즘
  - 더욱 다양한 Task에 적용