creative commons

COMMONS DEED

저작자표시-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사 학위논문

몬테카를로 탈락 방법을 이용한 신경망 모형의 불확실성 측정

Uncertainty Quantification of Neural Network Using Monte Carlo Dropout

2023년 2월

서울대학교 대학원

통계학과

박 수 원

몬테카를로 탈락 방법을 이용한 신경망 모형의 불확실성 측정

지도교수 이재용

이 논문을 이학석사 학위논문으로 제출함

2022년 10월

서울대학교 대학원
통계학과
박 수 원

박수원의 이학석사 학위논문을 인준함

2023년 1월

위 원 장	이 상 열	(인)
부 위 원 장	이 재 용	(인)
위 원	이 권 상	(인)

국문초록

신경망 모형의 급격한 발전으로, 이미지 인식 분야에서부터 신약 개발에 이르기까지 다양한 산업 현장에 신경망 모형이 도입되었다. 그에 발맞춰 신경망 모형의 불확실성을 예측할 수 있는 방법론에 대한 많은 연구가 진행되고 있다. 몬테카를로 탈락 방법은 기존의 신경망 모형을 크게 조정하지 않으면서도 불확실성을 측정할 수 있다는 점으로 인해 불확실성에 대한 연구가 필요한 산업 현장과 학계에서 자주 쓰이는 방법들 중 하나가 되었다. 다른 한편으로는 몬테카를로 탈락 방법의 한계가 이론적, 실증적으로 연구되고 있다. 본 논문은 몬테카를로 탈락 방법의 이론적 근거를 소개하고 모의 데이터에서 실제로 불확실성 측정이 어떤 양상을 띄는지 분석한다.

주요어: 불확실성 측정, 베이즈 신경망 모형, 심층 가우시안 과정
학번: 2020-28703

목차

국문초록 **i**

1 서론 **1**

2 몬테카를로 탈락 방법 **3**

2.1 사전분포가 가우시안 과정을 따르는 모형 3

2.2 사후 예측분포의 도출 5

2.3 변분 추론 및 최적화 6

2.4 심층 가우시안 과정을 도입한 모형 11

3 모의 데이터 분석 **13**

3.1 데이터 생성 13

3.2 분석 결과 15

4 결론 **20**

Abstract in English **24**

Chapter 1

서론

최근들어 실제 산업 현장에 적극적으로 적용되기 시작한 고도화된 신경망 모형은 사회적으로 큰 영향력을 갖게 되었다.[1] 예를들어 이미지 인식 분야에서 신경망 모형이 인간과 유사하거나 우월한 수준의 분류 성능을 보이게 되면서, 시각적 판단이 필요한 생산 과정 및 서비스에서 고성능 이미지 센서가 사람을 대체하게 되었다.

신경망 모형이 사회에 미치는 파급력이 커진 만큼, 신경망 모형의 불확실성을 이론적으로 규명하고 예측하기 위한 연구 또한 활발히 이루어지고 있다.[2] Gal and Ghahramani [3]는 신경망 모형의 학습 과정에서 과적합 방지를 위해 도입한 탈락 방법[4]을 심층 가우시안 과정(Deep Gaussian Process) 모형의 관점에서 해석하여 불확실성을 예측하고자 하였다. 몬테카를로 탈락(Monte Carlo Dropout)으로 불리는 이 방식은 기존의 탈락 방법과 같은 방식으로 각 학습 세대(epoch)마다 은닉층의 유닛들 중 일부를 일정한 확률로 활성 또는 비활성화하여 학습을 진행한다. 그 후에는 기존의 방식과 다르게 예측 단계에서도 은닉층의 유닛들 중 일부를 일정한 확률로 활성 또는 비활성화하여 예측을 여러번 수행하게 된다. 이렇게 뽑은 몬테카를로 표본에 대해 예측평균(predictive

mean), 예측분산(predictive variance)등을 구해서 불확실성을 추정하고자하는 것이 해당 방법론의 골자이다.

몬테카를로 탈락 방법의 핵심은 심층 신경망 모형에서 각 유닛에 탈락 방법을 적용해 L_2 정규화된 제곱 손실 함수에 대해 모형을 훈련시키는 것과 심층 가우시안 과정에 대한 근사 사후분포를 변분추론으로 구하는 것이 근사적으로 같은 최적화라는 점이다. 따라서 몬테카를로 탈락 방법을 적용하는 신경망 모형의 기존 훈련 방식을 크게 바꾸지 않고 적용할 수 있기 때문에 2D 영상분석[56], 자율주행 제어[78], 신약개발[910]등 다양한 분야에서 핵심 방법론 및 벤치마크 방법론으로 활용되고 있는 실정이다. 그러나, Osband [11]는 몬테카를로 탈락 방법으로 훈련한 신경망 모형에서 데이터의 양이 인식론적 불확실성(epistemic uncertainty)에 영향을 미치지 않음을 보였다. 즉, 데이터를 많이 확보하는 것을 통해 줄일 수 있는 불확실성을 줄이지 못한다는 것이다. 또한 Liu et al. [12]는 실증 연구를 통해 몬테카를로 탈락 방법이 분포 밖(Out of Distribution) 데이터의 불확실성을 낮게 측정함을 보였다.

본 논문에서는 몬테카를로 탈락 방법의 이론적 근거를 소개하고 모의 데이터에 대해 몬테카를로 탈락 방법을 적용한 신경망 모형과 가우시안 과정을 비교하는 것을 통해 실제로 어떤 현상이 일어나는지 알아보고자 한다.

Chapter 2

몬테카를로 탈락 방법

몬테카를로 탈락 방법의 이론적 바탕은 심층 가우시안 과정을 가정하는 것으로부터 시작한다. 심층 가우시안 과정은 공분산 함수의 변수가 가우시안 과정인 구조를 갖는다.[13] 따라서 함수에 대한 사전분포가 가우시안 과정을 따르는 기본적인 모형을 먼저 살펴보도록 한다.

2.1 사전분포가 가우시안 과정을 따르는 모형

새로운 데이터에 대한 확률변수인 $\mathbf{X}^*$(차원 $N \times Q$), $\mathbf{Y}^*(N \times D)$, 기존 데이터에 대한 확률변수인 $\mathbf{X}, \mathbf{Y}$, 가중치 행렬과 편향 벡터 $\mathbf{W}_1(Q \times K_1), \mathbf{b}_1(1 \times K_1)$, 그리고 가우시안 과정을 따르는 함수 $\mathbf{F}_1(N \times D)$가 다음과 같은 모형으로 표현된다고 하자.

$$\mathbf{F}_1 | \mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1 \sim \mathcal{MN}_{N \times D}(\mathbf{0}_{N \times D}, \widehat{\mathbf{K}}_1(\mathbf{X}^*, \mathbf{X}^*), I_D) \tag{2.1}$$

$$\mathbf{Y}^* | \mathbf{F}_1 \sim \mathcal{MN}_{N \times D}(\mathbf{F}_1, \tau^{-1}\mathbf{I}_N, I_D) \tag{2.2}$$

$$\text{where } \widehat{\mathbf{K}}_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{K_1} \sum_{k=1}^{K_1} \sigma_1(\mathbf{x}_1^\top [\mathbf{W}_1]_{\cdot k} + b_k)\sigma_1(\mathbf{x}_2^\top [\mathbf{W}_1]_{\cdot k} + b_k)$$

여기서 $\widehat{\mathbf{K}}_1$은 공분산 함수

$$\mathbf{K}_1(\mathbf{x}_1, \mathbf{x}_2) = \int p(\mathbf{w})p(\mathbf{b})\sigma_1(\mathbf{x}_1^\top \mathbf{w} + b)\sigma_1(\mathbf{x}_2^\top \mathbf{w} + b)\, d\mathbf{w}db$$

에 대한 몬테카를로 적분이다.

$\mathbf{K}_1(\mathbf{x}_1, \mathbf{x}_2)$가 잘 정의되는, 즉 양의 준정부호 공분산 함수임은 다음과 같이 보일 수 있다. $\mathbf{K}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^\top \mathbf{x}_2$가 양의 준정부호 공분산 함수인 것과 같은 논리로 $\sigma_1(\mathbf{x}_1^\top \mathbf{w} + b)\sigma_1(\mathbf{x}_2^\top \mathbf{w} + b)$ 또한 양의 준정부호 공분산 함수로 생각할 수 있다. 따라서 양의 준정부호 공분산 함수의 정의에 따라

$$\int p(\mathbf{w})p(\mathbf{b})\sigma_1(\mathbf{x}_1^\top \mathbf{w} + b)\sigma_1(\mathbf{x}_2^\top \mathbf{w} + b)\, d\mathbf{w}db \geq 0$$

이다. 따라서 $\mathbf{K}_1$으로 구성한 그람 행렬 G와 임의의 벡터 v에 대해 $v^\top G v \geq 0$이 항상 성립함을 알 수 있다. 그람 행렬 G가 양의 준정부호 공분산 행렬이므로 $\mathbf{K}_1$은 양의 준정부호 공분산 함수이다.[14]

본 논문에서 σ는 Rectified Linear Unit(ReLU) 비선형함수로, 가중치 행렬 $\mathbf{W}_1$의 각 열과 $\mathbf{b}_1$은 표준 다변량 정규분포로 놓는다. K_1은 후술될 신경망 모형 은닉층의 유닛 개수에 상응하는 값이다. 또한 조건부 $\mathbf{F}_1$와 조건부 $\mathbf{Y}^{**}$는 각 열이 서로 독립인 다변량 정규분포라 가정하기 때문에 행렬 정규분포 $\mathcal{MN}_{N\times D}(\cdot, \cdot, I_D)$의 형태를 갖는다.

2.2 사후 예측분포의 도출

위의 모형에 대한 사후 예측분포를 다음과 같이 표현할 수 있다.

$$
\begin{aligned}
&p(\mathbf{Y}^*|\mathbf{X}^*, \mathbf{X}, \mathbf{Y}) \\
&= \iiint p(\mathbf{Y}^*|\mathbf{F}_1)p(\mathbf{F}_1|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1)p(\mathbf{W}_1, \mathbf{b}_1|\mathbf{X}, \mathbf{Y})d\mathbf{F}_1 d\mathbf{W}_1 d\mathbf{b}_1 \\
&= \iint p(\mathbf{Y}^{**}|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1)p(\mathbf{W}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_1|\mathbf{X}, \mathbf{Y})d\mathbf{W}_1 d\mathbf{b}_1 \qquad (2.3) \\
&\quad \text{s.t } \mathbf{Y}^{**}|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1 \sim \mathcal{MN}_{N\times D}(\mathbf{0}_{N\times D}, \widehat{\mathbf{K}}(\mathbf{X}^*, \mathbf{X}^*) + \tau^{-1}\mathbf{I}_N, I_D) \\
&= \iint p(\mathbf{Y}^{**}|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1)p(\mathbf{W}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_1|\mathbf{X}, \mathbf{Y})d\mathbf{W}_1 d\mathbf{b}_1 \qquad (2.4) \\
&\quad \text{s.t } \mathbf{Y}^{**}|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1 \sim \mathcal{MN}_{N\times D}(\mathbf{0}_{N\times D}, \Phi_1\Phi_1^\top + \tau^{-1}\mathbf{I}_N, I_D) \\
&= \iiint p(\mathbf{Y}'|\mathbf{X}^*, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1)p(\mathbf{W}_2|\mathbf{X}, \mathbf{Y})p(\mathbf{W}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_1|\mathbf{X}, \mathbf{Y})d\mathbf{W}_2 d\mathbf{W}_1 d\mathbf{b}_1 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2.5) \\
&\quad \text{s.t } \left[\mathbf{Y}'\right]_{\cdot j}|\mathbf{X}^*, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1 \sim \mathcal{N}(\Phi_1 \left[\mathbf{W}_2\right]_{\cdot j}, \tau^{-1}\mathbf{I}_N)
\end{aligned}
$$

식 (2.3)는 [Bishop and Nasrabadi [15] p.93]에 수록된 주변 정규분포와 조건부 정규분포와의 관계를 조건 (2.1), (2.2)에 적용하여, $p(\mathbf{Y}^*|\mathbf{F}_1)p(\mathbf{F}_1|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1)$를 $\mathbf{F}_1$에 대해 적분하는 것으로 부터 도출된다. 다음으로, 식 (2.4)은 차원 $N \times K_1$의 행렬

$$
\Phi_1 = \begin{bmatrix} \sqrt{\frac{1}{K_1}}\boldsymbol{\sigma}_1(\mathbf{x}_1^{*T}\mathbf{W}_1 + \mathbf{b}_1^\top) \\ \sqrt{\frac{1}{K_1}}\boldsymbol{\sigma}_1(\mathbf{x}_2^{*T}\mathbf{W}_1 + \mathbf{b}_1^\top) \\ \vdots \\ \sqrt{\frac{1}{K_1}}\boldsymbol{\sigma}_1(\mathbf{x}_N^{*T}\mathbf{W}_1 + \mathbf{b}_1^\top) \end{bmatrix} \qquad (2.6)
$$

를 도입하여 $\widehat{\mathbf{K}}_1(\mathbf{X}^*, \mathbf{X}^*) = \Phi_1\Phi_1^\top$로 나타낸 것이다. 최종적으로, 식 (2.5)은 [Bishop and Nasrabadi [15] p.93]에 수록된 주변 정규분포와 조건부 정규분포와의 관계를 식 (2.3)과는 반대로 적용하여,

$$p([\mathbf{Y}^{**}]_{\cdot j}\,|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1) = \int p([\mathbf{Y}']_{\cdot j}\,|\mathbf{X}^*, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) p([\mathbf{W}_2]_{\cdot j}\,|\mathbf{X}, \mathbf{Y})\, d\,[\mathbf{W}_2]_{\cdot j}$$
$$\text{s.t}\ \ [\mathbf{W}_2]_{\cdot j}\,|\mathbf{X}, \mathbf{Y} \sim \mathcal{N}(0, \mathbf{I}_{K_1}), \quad j \in \{1, \ldots, D\}$$

을 만족하는 차원 $N \times D$의 가중치 행렬 W_2를 도입한 결과이다.

2.3 변분 추론 및 최적화

위의 사후 예측분포 $\mathbf{Y}^*|\mathbf{X}^*, \mathbf{X}, \mathbf{Y}$를 구하기 위해 필요한 $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1$에 대한 사후분포

$$p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1|\mathbf{X}, \mathbf{Y}) = p(\mathbf{W}_2|\mathbf{X}, \mathbf{Y})p(\mathbf{W}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_1|\mathbf{X}, \mathbf{Y})$$

는 분석적으로 구하기 어렵다(analytically intractable). 따라서, 변분 추론(variational inference)을 통해 일정한 조건을 만족하는 분포를 근사 사후분포 $q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) = q(\mathbf{W}_1)q(\mathbf{W}_2)q(\mathbf{b}_1)$로 두고 이를 $p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1|\mathbf{X}, \mathbf{Y})$와 최대한 비슷한 분포를 갖도록 최적화한다. 몬테카를로 탈락 방법의 핵심은 혼합 베르누이분포에 가까운 혼합 정규분포를 가중치 행렬 $\mathbf{W}_1, \mathbf{W}_2$의 근사 사후분포로 뒀을 때, 위의 변분 추론 최적화 과정이 심층 신경망 모형에서 각 유닛에 탈락 방법을 적용해 L_2 정규화된 제곱 손실 함수에 대해 모형을 훈련시키는 것과 근사적으로 같다는 점이다.

위의 최적화는 구체적으로 $q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1)$와 $p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1|\mathbf{X}, \mathbf{Y})$ 사이의 쿨백-라이블러 발산(Kullback-Leibler divergence, 이하 KL)을 최소화 하는 것이다.

이는 다음의 로그 증거 하한(Log Evidence Lower Bound)을 최대화 하는 것[15]과 같다.

$$
\begin{aligned}
&\mathcal{L}_{ELBO} \\
&= \int q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) \log p(\mathbf{Y}|\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{X}) d\mathbf{W}_1 d\mathbf{W}_2 d\mathbf{b}_1 && (2.7) \\
&\quad - \mathrm{KL}(q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1)||p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1)) && (2.8)
\end{aligned}
$$

$\mathcal{L}_{ELBO}$의 몬테카를로 적분 근사를 용이하게 하기 위해, 새로운 모수 $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{m}$를 도입하여 $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$의 근사 사후분포 q를 다음과 같이 재매개변수화(reparameterization)할 수 있다.

$$
\begin{aligned}
\boldsymbol{\epsilon}_1 &\sim \mathcal{MN}_{Q\times K_1}(\mathbf{0}_{Q\times K_1}, \mathbf{I}_Q, \mathbf{I}_{K_1}) && (2.9) \\
\boldsymbol{\epsilon}_2 &\sim \mathcal{MN}_{K_1\times D}(\mathbf{0}_{K_1\times D}, \mathbf{I}_{K_1}, \mathbf{I}_D) && (2.10) \\
\boldsymbol{\epsilon}^\top &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K_1}) \\
[\mathbf{Z}_h]_{ij} &\sim \mathrm{Bernoulli}(p_h) \cdot \mathbf{I}(i=j);\ h \in \{1,2\} \\
\mathbf{W}_1 &= \mathbf{Z}_1(\mathbf{M}_1 + \delta\boldsymbol{\epsilon}_1) + (\mathbf{I}_Q - \mathbf{Z}_1)\delta\boldsymbol{\epsilon}_1 && (2.11) \\
\mathbf{W}_2 &= \mathbf{Z}_2(\mathbf{M}_2 + \delta\boldsymbol{\epsilon}_2) + (\mathbf{I}_{K_1} - \mathbf{Z}_2)\delta\boldsymbol{\epsilon}_2 && (2.12) \\
\mathbf{b}_1 &= \mathbf{m} + \delta\boldsymbol{\epsilon} && (2.13)
\end{aligned}
$$

여기서 스칼라 값 $\delta > 0$는 $\mathbf{W}_1, \mathbf{W}_2$의 혼합 정규분포를 혼합 베르누이분포에 가까운 분포로 만들기 위해 충분히 작은 값을 지니는 초모수이다. $\mathbf{Z}_1$은 대각 성분들이 각각 초모수 p_1인 베르누이분포들을 따르는 행렬이다. $\mathbf{Z}_2$도 $\mathbf{Z}_1$과 같은 방식으로 구성된 행렬이다. 목표는 $\mathcal{L}_{ELBO}$를 최대화 하게끔 $\mathbf{M}_1$(차원 $Q \times K_1$), $\mathbf{M}_2(K_1 \times D)$, $\mathbf{m}(1 \times K_1)$를 학습하는 것이다.

위의 가정하에서 (2.7)의 피적분 함수 부분을 살펴보면 다음과 같다.

$$
\begin{aligned}
& q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) \log p(\mathbf{Y}' | \mathbf{X}^*, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) \\
&= q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) \sum_{j=1}^{D} \log p([\mathbf{Y}']_{\cdot j} | \mathbf{X}^*, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) \\
&= q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) \sum_{n=1}^{N} \log p([\mathbf{Y}']_{n\cdot} \,|\, [\mathbf{X}^*]_{n\cdot}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) \qquad (2.14) \\
&= q(\mathbf{Z}_1, \boldsymbol{\epsilon}_1, \mathbf{Z}_2, \boldsymbol{\epsilon}_2, \boldsymbol{\epsilon}) \sum_{n=1}^{N} \log p([\mathbf{Y}']_{n\cdot} \,|\, [\mathbf{X}^*]_{n\cdot}, \mathbf{W}_1(\mathbf{Z}_1, \boldsymbol{\epsilon}_1), \mathbf{W}_2(\mathbf{Z}_2, \boldsymbol{\epsilon}_2), \mathbf{b}_1(\boldsymbol{\epsilon})) \\
&\approx \sum_{n=1}^{N} \log p([\mathbf{Y}']_{n\cdot} \,|\, [\mathbf{X}^*]_{n\cdot}, \widehat{\mathbf{W}}_1^n(\widehat{\mathbf{Z}}_1^n, \widehat{\boldsymbol{\epsilon}}_1^n), \widehat{\mathbf{W}}_2^n(\widehat{\mathbf{Z}}_2^n, \widehat{\boldsymbol{\epsilon}}_2^n), \widehat{\mathbf{b}}_1^n(\widehat{\boldsymbol{\epsilon}}^n)) \qquad (2.15) \\
&\approx \left(\sum_{n=1}^{N} \frac{\tau}{2} || \, [\mathbf{Y}']_{n\cdot} - \sqrt{\frac{1}{K_1}} \sigma_1(\mathbf{x}_n^{*T} \widehat{\mathbf{Z}}_1^n \mathbf{M}_1 + \mathbf{m}^\top) \widehat{\mathbf{Z}}_2^n \mathbf{M}_2 ||_2^2 \right) - N \log(\frac{\tau}{2\pi})^{\frac{D}{2}} \qquad (2.16)
\end{aligned}
$$

식 (2.14)은 다음의 관계로부터 유도된다.

$$
\begin{aligned}
& \sum_{j=1}^{D} \log p([\mathbf{Y}']_{\cdot j} | \mathbf{X}^*, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1) \\
&= \left(\sum_{j=1}^{D} \frac{\tau}{2} || \, [\mathbf{Y}']_{\cdot j} - \Phi_1 \, [\mathbf{W}_2]_{\cdot j} \, ||_2^2 \right) - D \log(\frac{\tau}{2\pi})^{\frac{N}{2}} \\
&= \left(\sum_{n=1}^{N} \frac{\tau}{2} || \, [\mathbf{Y}']_{n\cdot} - \sqrt{\frac{1}{K_1}} \sigma_1(\mathbf{x}_n^{*T} \mathbf{W}_1 + \mathbf{b}_1^\top) \mathbf{W}_2 ||_2^2 \right) - N \log(\frac{\tau}{2\pi})^{\frac{D}{2}} \qquad (2.17) \\
&= \sum_{n=1}^{N} \log p([\mathbf{Y}']_{n\cdot} \,|\, [\mathbf{X}^*]_{n\cdot}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1)
\end{aligned}
$$

식 (2.15)는 (2.11),(2.12),(2.13)의 근사 사후분포 q로부터 크기 1의 몬테카를로 표본 $\widehat{\mathbf{Z}}_1, \widehat{\boldsymbol{\epsilon}}_1, \widehat{\mathbf{Z}}_2, \widehat{\boldsymbol{\epsilon}}_2, \widehat{\boldsymbol{\epsilon}}$을 $\mathbf{Y}'$의 열 차원 N 묶음만큼 추출한 것이다. 식 (2.16)는 충분히 작은 $\delta > 0$에 대해 $\widehat{\mathbf{W}}_1^n \approx \widehat{\mathbf{Z}}_1^n \mathbf{M}_1, \widehat{\mathbf{W}}_2^n \approx \widehat{\mathbf{Z}}_2^n \mathbf{M}_2, \widehat{\mathbf{b}}_1^n \approx \mathbf{m}$로 근사할 수

있으므로, 이를 식 (2.17)에 대입한 결과이다.

다음으로, (2.8)은 $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1$가 서로 독립이므로 다음과 같이 분해된다[16].

$$
\begin{aligned}
&\mathrm{KL}(q(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1)||p(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1)) \\
&= \mathrm{KL}(q(\mathbf{W}_1, \mathbf{W}_2)||p(\mathbf{W}_1, \mathbf{W}_2))\mathrm{KL}(q(\mathbf{b}_1|\mathbf{W}_1, \mathbf{W}_2)||p(\mathbf{b}_1|\mathbf{W}_1, \mathbf{W}_2)) \\
&= \mathrm{KL}(q(\mathbf{W}_1)||p(\mathbf{W}_1))\mathrm{KL}(q(\mathbf{W}_2|\mathbf{W}_1)||p(\mathbf{W}_2|\mathbf{W}_1))\mathrm{KL}(q(\mathbf{b}_1|\mathbf{W}_1, \mathbf{W}_2)||p(\mathbf{b}_1|\mathbf{W}_1, \mathbf{W}_2)) \\
&= \mathrm{KL}(q(\mathbf{W}_1)||p(\mathbf{W}_1))\mathrm{KL}(q(\mathbf{W}_2)||p(\mathbf{W}_2))\mathrm{KL}(q(\mathbf{b}_1)||p(\mathbf{b}_1))
\end{aligned}
$$

[Gal and Ghahramani [3] Appendix. Proposition 1]에 따라, 일반화 된 χ^2 분포의 성질을 이용하면 충분히 큰 K_1와 Q에 대해 다음과 같이 근사할 수 있다. 또한, $\mathbf{b}_1$의 근사 사후분포에 대해서는 다음과 같이 정확한 식을 구할 수 있다.

$$
\mathrm{KL}(q(\mathbf{W}_1)||p(\mathbf{W}_1)) \approx \frac{p_1}{2}\sum_{q=1}^{Q}\left([\mathbf{M}_1]_{q\cdot}^{\top}[\mathbf{M}_1]_{q\cdot}\right) + QK_1(\delta^2 - \log(\delta^2) - 1) + C_1 \quad (2.18)
$$

$$
\mathrm{KL}(q(\mathbf{W}_2)||p(\mathbf{W}_2)) \approx \frac{p_2}{2}\sum_{k=1}^{K_1}\left([\mathbf{M}_2]_{k\cdot}^{\top}[\mathbf{M}_2]_{k\cdot}\right) + K_1D(\delta^2 - \log(\delta^2) - 1) + C_2 \quad (2.19)
$$

$$
\mathrm{KL}(q(\mathbf{b}_1)||p(\mathbf{b}_1)) = \frac{1}{2}\left(\mathbf{m}^{\top}\mathbf{m} + K_1(\delta^2 - \log(\delta^2) - 1)\right) + C_3 \quad (2.20)
$$

여기서, $C_i, i = 1, 2, 3$는 모수 $\mathbf{M}_1, \mathbf{M}_2, \mathbf{m}$와 상관없는 상수값이다. 또한, δ는 식 (2.16)을 구하는 과정에서 충분히 작지만 0보다는 큰 값으로 설정하기 때문에 모수 $\mathbf{M}_1, \mathbf{M}_2, \mathbf{m}$와 상관없는 상수로 취급할 수 있다.

최종적으로, $\mathbf{y}_n = [\mathbf{Y}']_{n\cdot}$, $\widehat{\mathbf{y}}_n = \sqrt{\frac{1}{K_1}}\sigma_1(\mathbf{x}_n^{*T}\widehat{\mathbf{Z}}_1^n\mathbf{M}_1 + \mathbf{m}^{\top})\widehat{\mathbf{Z}}_2^n\mathbf{M}_2$라 할 때 (2.16),

(2.18), (2.19), (2.20)를 이용해 $\mathcal{L}_{ELBO}$를 근사하면 다음과 같다.

$$
\begin{aligned}
&\mathcal{L}_{ELBO} \\
&\approx -\frac{\tau}{2}\left(\sum_{n=1}^{N}||\left[\mathbf{Y}'\right]_{n\cdot} - \sqrt{\frac{1}{K_1}}\sigma_1(\mathbf{x}_n^{*T}\widehat{\mathbf{Z}}_1^n\mathbf{M}_1 + \mathbf{m}^\top)\widehat{\mathbf{Z}}_2^n\mathbf{M}_2||_2^2\right) \\
&\quad - \frac{p_1}{2}||\mathbf{M}_1||_2^2 - \frac{p_2}{2}||\mathbf{M}_2||_2^2 - \frac{1}{2}||\mathbf{m}||_2^2 + C \\
&\propto -\frac{1}{2N}\left(\sum_{n=1}^{N}||\mathbf{y}_n - \widehat{\mathbf{y}}_n||_2^2\right) - \sum_{h=1}^{2}\frac{p_h}{2\tau N}||\mathbf{M}_h||_2^2 - \frac{1}{2\tau N}||\mathbf{m}||_2^2 \qquad (2.21)
\end{aligned}
$$

Paisley, Blei, and Jordan [17]에 따르면 확률적 목적함수 (2.21)를 최적화 하는 것은 $\mathcal{L}_{ELBO}$와 같은 극한에 수렴한다.

식 (2.21)과의 비교를 위해 얕은 신경망 모형에서 모수 $\mathbf{M}_1, \mathbf{M}_2, \mathbf{m}$의 L_2 정규화 및 각 유닛에 대한 몬테카를로 탈락 방법을 적용한 평균제곱오차 손실함수를 생각하자. $\mathbf{M}_1, \mathbf{M}_2, \mathbf{m}$의 가중치 감쇠 초모수를 각각 $\lambda_1 = \frac{p_1}{2\tau N}, \lambda_2 = \frac{p_2}{2\tau N}, \lambda_3 = \frac{1}{2\tau N}$라 하면 얕은 신경망 모형에 대해 위의 근사 식과 동일한 최대화 목적 함수를 다음과 같이 도출할 수 있다.

$$
-\mathcal{L}_{SNN} = -\frac{1}{2N}\left(\sum_{n=1}^{N}||\mathbf{y}_n - \widehat{\mathbf{y}}_n||_2^2\right) - \sum_{h=1}^{2}\lambda_h||\mathbf{M}_h||_2^2 - \lambda_3||\mathbf{m}||_2^2 \qquad (2.22)
$$

따라서, (2.22)에서 크기가 1인 배치에 대한 경사하강법을 수행하듯이 $\widehat{\mathbf{y}}_n$을 한 번 뽑은 후에 최적화하고 다시 $\widehat{\mathbf{y}}_n$을 한 번 뽑은 후에 최적화하는 방식을 반복하는 것을 통해 $\mathcal{L}_{ELBO}$를 근사할 수 있다.

여기서 $\lambda_1 = \frac{p_1}{2\tau N}, \lambda_2 = \frac{p_2}{2\tau N}, \lambda_3 = \frac{1}{2\tau N}$임을 통해 가중치 감쇠 초모수인 $\lambda_1, \lambda_2, \lambda_3$와 가우시안 과정의 초모수인 τ, 근사 사후분포의 초모수인 p_1, p_2 사이의 관계를 살펴볼 수 있다.

2.4 심층 가우시안 과정을 도입한 모형

지금까지의 도출 과정을 심층 신경망 모형에 적용하기 위해 심층 가우시안 과정을 다음과 같이 도입한다.

$$\mathbf{F}_1|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1 \sim \mathcal{MN}_{N\times K_2}(\mathbf{0}_{N\times K_2}, \widehat{\mathbf{K}}_1(\mathbf{X}^*, \mathbf{X}^*), I_{K_2}) \tag{2.23}$$

$$\mathbf{F}_2|\mathbf{X}^*, \mathbf{F}_1, \mathbf{b}_2 \sim \mathcal{MN}_{N\times D}(\mathbf{0}_{N\times D}, \widehat{\mathbf{K}}_2(\mathbf{F}_1, \mathbf{F}_1), I_D) \tag{2.24}$$

$$\mathbf{Y}^*|\mathbf{F}_2 \sim \mathcal{MN}_{N\times D}(\mathbf{F}_2, \tau^{-1}\mathbf{I}_N, I_D) \tag{2.25}$$

$$\text{where } \widehat{\mathbf{K}}_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{K_1}\boldsymbol{\sigma}_1(\mathbf{x}_1^\top \mathbf{W}_1 + \mathbf{b}_1^\top)\boldsymbol{\sigma}_1(\mathbf{x}_1^\top \mathbf{W}_1 + \mathbf{b}_1^\top)^\top$$

$$\widehat{\mathbf{K}}_2(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{K_2}\boldsymbol{\sigma}_2(\mathbf{x}_1^\top + \mathbf{b}_2^\top)\boldsymbol{\sigma}_2(\mathbf{x}_1^\top + \mathbf{b}_2^\top)^\top$$

위의 모형에 대한 사후 예측분포를 다음과 같이 표현할 수 있다.

$$\begin{aligned}
&p(\mathbf{Y}^*|\mathbf{X}^*, \mathbf{X}, \mathbf{Y}) \\
&= \int \cdots \int p(\mathbf{Y}^*|\mathbf{F}_2)p(\mathbf{F}_2|\mathbf{X}^*, \mathbf{F}_1, \mathbf{b}_2)p(\mathbf{F}_1|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1) \\
&\qquad p(\mathbf{W}_1, \mathbf{b}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_2|\mathbf{X}, \mathbf{Y})\, d\mathbf{F}_2 d\mathbf{F}_1 d\mathbf{W}_1 d\mathbf{b}_1 d\mathbf{b}_2 \\
&= \int \cdots \int p(\mathbf{Y}'|\mathbf{X}^*, \mathbf{F}_1, \mathbf{W}_3, \mathbf{b}_2)p(\mathbf{F}_1|\mathbf{X}^*, \mathbf{W}_1, \mathbf{b}_1)p(\mathbf{W}_3|\mathbf{X}, \mathbf{Y}) && (2.26) \\
&\qquad p(\mathbf{W}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_2|\mathbf{X}, \mathbf{Y})\, d\mathbf{F}_1 d\mathbf{W}_3 d\mathbf{W}_1 d\mathbf{b}_1 d\mathbf{b}_2 && (2.27) \\
&\quad \text{s.t } [\mathbf{Y}']_{\cdot j}\, |\mathbf{X}^*, \mathbf{F}_1, \mathbf{W}_3, \mathbf{b}_2 \sim \mathcal{N}(\Phi_2(\mathbf{F}_1)\, [\mathbf{W}_3]_{\cdot j}\,, \tau^{-1}\mathbf{I}_N) \\
&= \int \cdots \int p(\mathbf{Y}''|\mathbf{X}^*, \mathbf{W}_3, \mathbf{W}_2, \mathbf{b}_2, \mathbf{W}_1, \mathbf{b}_1)\, p(\mathbf{W}_3|\mathbf{X}, \mathbf{Y})p(\mathbf{W}_2|\mathbf{X}, \mathbf{Y}) \\
&\qquad p(\mathbf{W}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_1|\mathbf{X}, \mathbf{Y})p(\mathbf{b}_2|\mathbf{X}, \mathbf{Y})\, \mathbf{W}_3 d\mathbf{W}_2 d\mathbf{W}_1 d\mathbf{b}_1 d\mathbf{b}_2 && (2.28) \\
&\quad \text{s.t } [\mathbf{Y}'']_{\cdot j}\, |\mathbf{X}^*, \mathbf{W}_3, \mathbf{W}_2, \mathbf{b}_2, \mathbf{W}_1, \mathbf{b}_1 \sim \mathcal{N}(\Phi_2(\Phi_1\mathbf{W}_2)\, [\mathbf{W}_3]_{\cdot j}\,, \tau^{-1}\mathbf{I}_N)
\end{aligned}$$

식 (2.27)는 (2.3), (2.5), (2.4)의 과정과 마찬가지로 $\mathbf{F}_2$에 대해 적분하고 $\mathbf{W}_3$을 도입하는 것을 통해 얻어진다. 그 과정에서 다음의 행렬을 도입한다.

$$\Phi_2(\mathbf{F}_1) = \begin{bmatrix} \sqrt{\frac{1}{K_2}}\boldsymbol{\sigma}_2(\mathbf{F}_1 + \mathbf{b}_1^\top) \\ \sqrt{\frac{1}{K_2}}\boldsymbol{\sigma}_2(\mathbf{F}_1 + \mathbf{b}_1^\top) \\ \vdots \\ \sqrt{\frac{1}{K_2}}\boldsymbol{\sigma}_2(\mathbf{F}_1 + \mathbf{b}_1^\top) \end{bmatrix}$$

식 (2.27) 또한 $\mathbf{F}_1$에 대해 적분하고 $\mathbf{W}_2$를 도입하는 것을 통해 얻어진다. 이 때 Φ_1은 (2.6)과 같고 Φ_2에는 $\mathbf{F}$ 대신 $\Phi_1\mathbf{W}_2$를 대입한다. 그 후의 과정은 2.3의 논리를 따라가면 된다. 위의 모형은 2층의 은닉층에 해당되는 심층 신경망 모형에 대응된다. 더 깊은 은닉층을 갖는 심층 신경망 모형에 대응되는 심층 가우시안 과정을 구축하는 것은 조건 (2.23)과 (2.25) 사이에 (2.24)을 추가한 것과 같은 방식으로 (2.24)와 (2.25) 사이에 가우시안 과정은 따르는 조건부 함수를 추가하면 된다.

Chapter 3

모의 데이터 분석

3.1 데이터 생성

현실에서는 가우시안 과정을 데이터 생성 과정(data generating process)으로 갖는 데이터를 찾기 어렵다. 따라서 모의 데이터를 생성한 후 해당 데이터에 대해 몬테카를로 탈락 방법을 적용한 신경망 모형과 가우시안 과정이 적합되는 양상을 살펴본다.[1] 구체적으로는, 독립변수의 값이 표본의 최솟값과 최댓값 범위 사이를 어느정도 빠짐없이 채우고 있는 경우의 양상과 범위 가운데가 일정 부분 없어서 해당 부분을 분포 밖(out of distribution) 데이터로 볼 수 있는 경우의 양상을 살펴본다.

먼저, GPy[18] 파이썬 패키지를 이용하여 다음의 모형으로부터 데이터를 생성한다.

$$\mathbf{F}_1 \sim \mathcal{N}(\mathbf{0}, \widehat{\mathbf{K}}(\mathbf{X}, \mathbf{X})) \tag{3.1}$$

$$\mathbf{Y}|\mathbf{F}_1 \sim \mathcal{N}(\mathbf{F}_1, \mathbf{I}_N) \tag{3.2}$$

[1] https://github.com/sepiabrown/MCDropout

$$\text{where } \widehat{\mathbf{K}}(\mathbf{x}_1, \mathbf{x}_2) = \frac{2}{\pi}\sin^{-1}\left(\frac{\mathbf{x}_1^\top \mathbf{x}_2 + 1}{\sqrt{\mathbf{x}_1^\top \mathbf{x}_1 + 2}\sqrt{\mathbf{x}_2^\top \mathbf{x}_2 + 2}}\right)$$

여기서 $\widehat{\mathbf{K}}$은 은닉층의 비선형함수가 ReLU인 얕은 신경망 모형을 가우시안 과정으로 표현할 때의 공분산 함수이다.[14] 주어진 정보가 구간마다 다르게 하기 위해 $\mathbf{X}$는 $[-45, 35], [-25, -15], [-5, 5], [15, 25], [35, 45]$ 구간에서 등간격의 관측값들을 각각 10개씩 추출하고 $[35, -25], [-15, -5], [5, 15], [25, 35]$ 사이의 구간에서는 관측값을 설정하지 않는다. 그 후 $\mathbf{X}$값들과 $\mathbf{Y}$값들을 표준화한다. 그림 3.1은 추출된 $\mathbf{X}$,$\mathbf{Y}$에 대한 산포도이다.

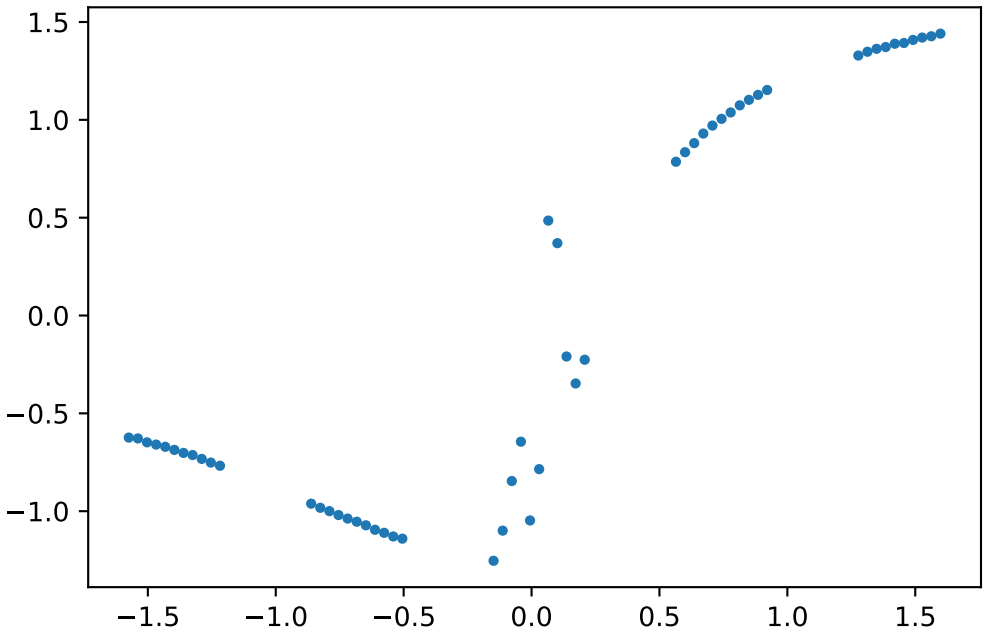

그림 3.1: 추출된 $\mathbf{X}$,$\mathbf{Y}$값들에 대한 산포도

3.2 분석 결과

추출된 값들에 대해 그림 3.2로 표현된 신경망 모형을 적합시킨다. 신경망 모형의 은닉층은 2000개의 유닛으로 구성한다. 탈락층의 각 유닛은 Bernoulli(p), $p = 0.8$을 따르고 L_2 정규화를 위한 가중치 감쇠 초모수 λ의 값은 0.01이다. 또한 확률적 경사 하강법(stochastic gradient descent) 최적화 방식을 사용하고 경사하강의 학습 세대(epoch) 횟수는 10000번으로 설정한다. 이렇게 훈련된 모형으로부터 몬테카를로 표본을 1000번 추출하여 예측평균과 예측분산을 구한 결과는 그림 3.3과 같다. 가운데의 파란색 선은 예측평균, 그 밖의 음영된 구간은 예측평균으로부터 예측표준편차 한 단위만큼의 범위를 나타낸 것이다. 신경망 모형에서 식 (2.21)과는 다르게 입력값에 완전연결층을 배치한 후 몬테카를로 탈

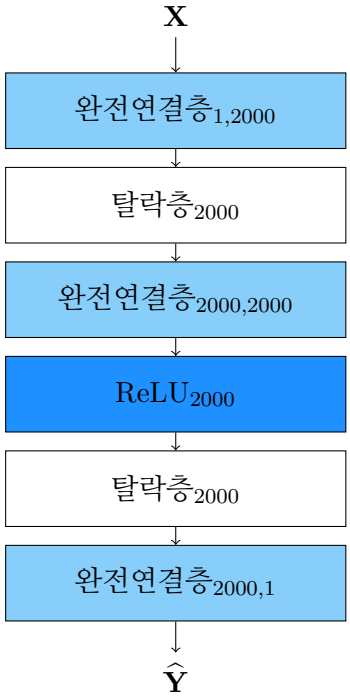

그림 3.2: 몬테카를로 탈락 방법이 적용된 신경망 모형의 구조도

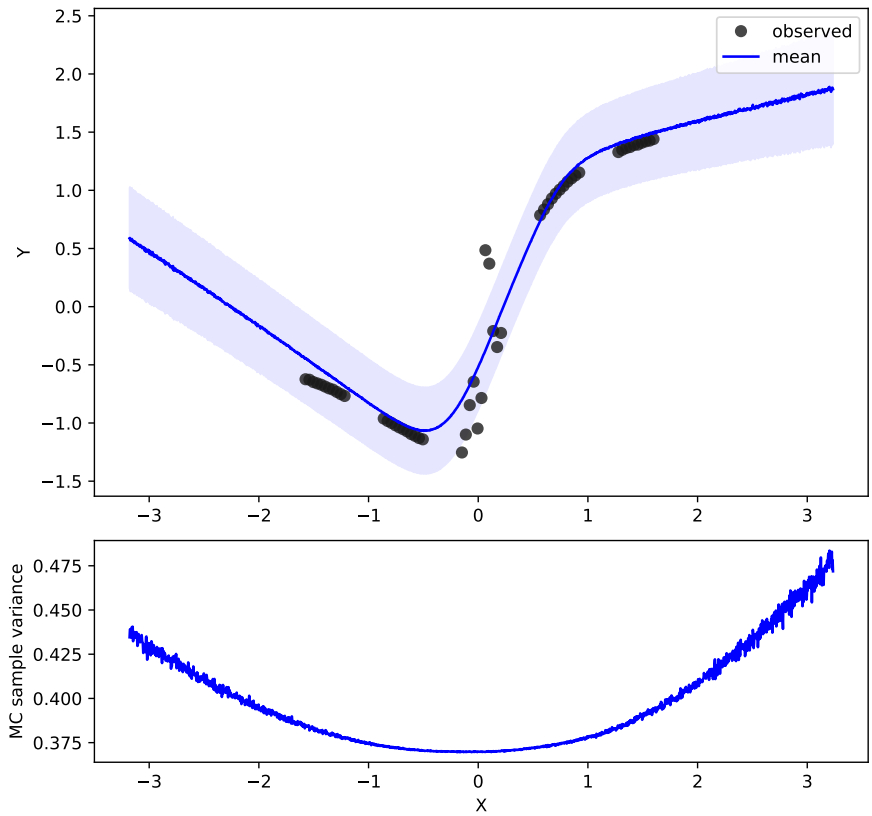

그림 3.3: 신경망 모형 적합 결과

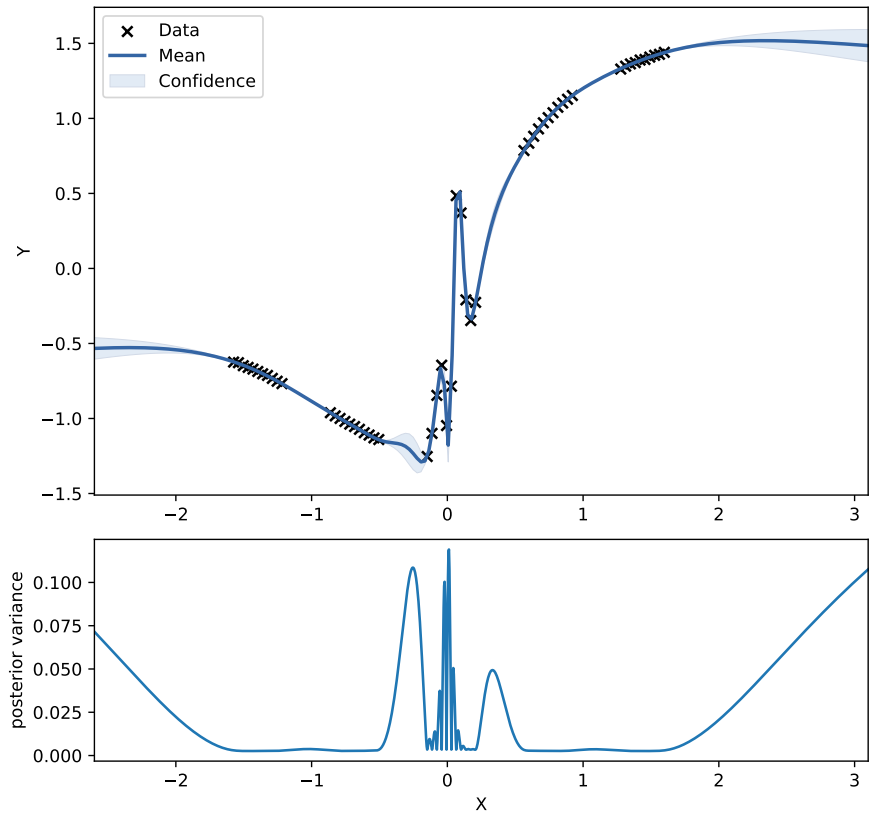

그림 3.4: 가우시안 과정 모형 적합 결과

락 방법을 적용한 이유는 낮은 차원을 갖는 입력값에 곧바로 탈락방식을 적용할 경우 정의역의 경계값 근처에서의 몬테카를로 표본의 예측분산이 급격하게 증가하는 현상이 발생하기 때문이다. 해당 완전연결층을 적용하지 않아도 정의역 안의 몬테카를로 표본의 예측분산은 비슷한 경향을 띈다.

그림 3.4는 위의 신경망 모형에 대응되는 가우시안 과정을 (3.1), (3.2)과 같은 가정으로 구성하여 적합한 결과이다. 그림 3.3와 비슷하게 가운데의 파란색 선은 예측평균, 그 밖의 음영된 구간은 95% 예측구간의 범위를 나타낸 것이다. 그림 3.3과 3.4를 비교하면 알 수 있듯이, 탈락 방법의 신경망 모형은 추출된 독립변수 값 범위의 가장자리에서 예측분산이 증가하는 패턴을 어느정도 근사하고 있다. 하지만 0 근방에서와 같이 주변과 질적으로 다른 불규칙한 패턴이 갑자기 등장하는 부분에 대한 예측분산에 대해서는 낮은 값을 갖는 것으로 파악하여 실제 가우시안 과정 모형에서의 예측분산을 잘 근사하지 못하는 경향을 보인다.

다음으로 추출된 독립변수 값 범위의 가운데가 비어있을 경우를 살펴보기 위해 $\mathbf{X}$의 값이 $[0, 1]$ 사이인 표본을 삭제한 후에 각각의 모형을 적합시킨다. 그림 3.5를 보면 알 수 있듯이 추출된 독립변수 값 범위의 가장자리 부분에서는 예측분산이 증가하는 패턴을 어느정도 표현하고 있다. 하지만, 그림 3.5와 3.6을 비교해보면 삭제된 $[0, 1]$ 부분에 대해 가우시안 과정 모형은 예측분산이 큰 값을 갖는 반면, 탈락 방법의 신경망 모형은 낮은 값을 갖는다. 신경망 모형에 대한 위의 경향성은 더 깊은 은닉층, 더 많은 유닛수에 대해서도 일관되게 나타난다. 또한 다양한 값의 가중치 감쇠 초모수를 적용해도 전반적인 패턴은 크게 달라지지 않는다.

위의 분석은 데이터의 양이 인식론적 불확실성에 영향을 미치지 않는다는 Osband [11]의 결론과 유사하다. 왜냐하면 $\mathbf{X}$의 값이 $[0, 1]$ 사이인 표본의 삭제 유무와 상관없이 $[0, 1]$ 사이의 불확실성은 일관되게 낮게 측정되었다. 이러한 경향성은 모의 데이터의 표본수를 증가시켜도 일관되게 관찰된다.

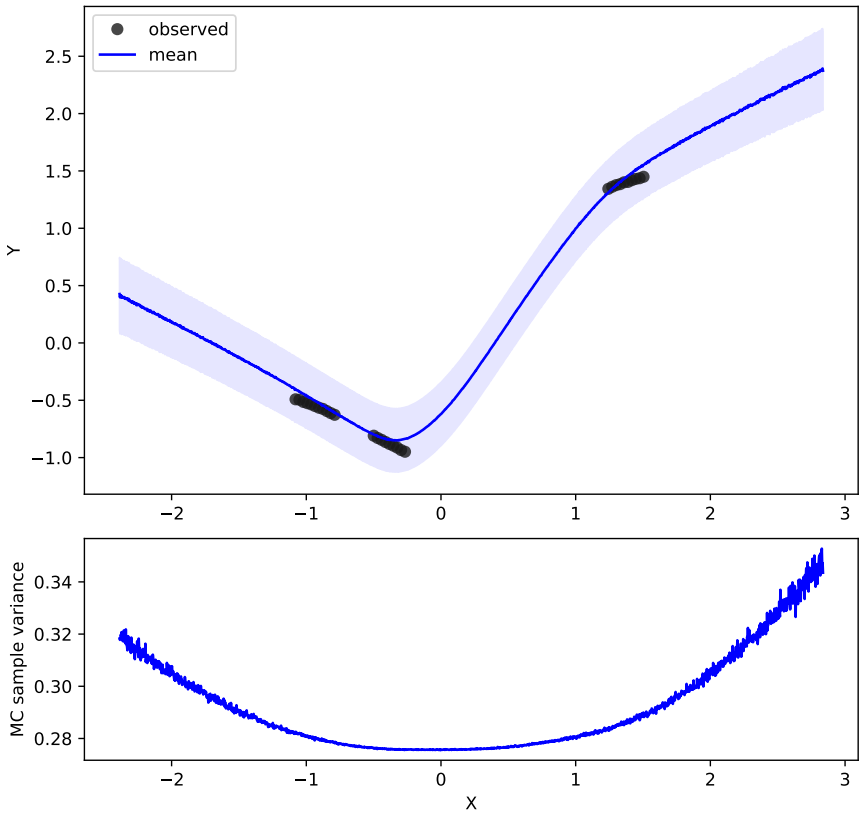

그림 3.5: 삭제된 표본에 대한 신경망 모형 적합 결과

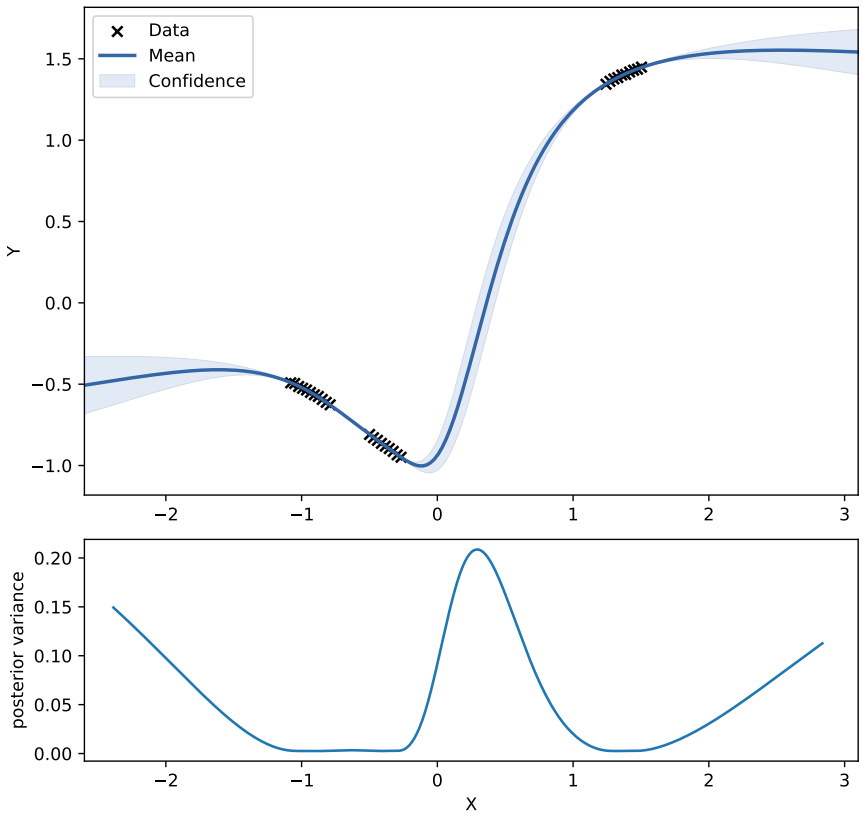

그림 3.6: 삭제된 표본에 대한 가우시안 과정 모형 적합 결과

또한 위의 분석은 분포 밖의 데이터에 대한 불확실성이 낮게 측정된다는 Liu et al. [12]의 실증연구 결과와도 유사하다. 그림 3.5에서 $\mathbf{X}$의 값이 $[0, 1]$인 구간은 분포 밖의 데이터라 할 수 있는데, 이에 대한 불확실성이 반영되지 않은 것을 볼 수 있다. 이렇게 $\mathbf{X}$ 범위 안쪽의 분포 밖 데이터에서 불확실성이 낮게 측정되는 현상은 $\mathbf{X}$에서 삭제하지 않은 $[-2, 0]$ 구간의 값들과 $[1, 2]$ 구간의 값들 사이를 더 크게 벌렸을 경우에도 일관되게 관찰된다.

Chapter 4

결론

본 논문에서는 몬테카를로 탈락 방법의 이론적 근거를 살펴보고 이에 대한 모의 데이터 분석을 진행하였다. 먼저 심층 가우시안 과정을 가정한 후 이에 대한 사후분포를 근사하기 위하여 변분추론을 수행하였다. 이 때, 특정 가정을 만족하는 혼합 정규분포를 근사 사후분포로 두고 로그 증거 하한을 최적화하는 것이 심층 신경망 모형에서 각 유닛에 탈락 방법을 적용하여 모형을 훈련시키는 것과 근사적으로 같음을 소개하였다. 다음으로 가우시안 과정으로부터 모의 데이터를 생성하였다. 이에 대해 몬테카를로 탈락 방법을 적용한 신경망 모형과 가우시안 과정을 훈련시키고 훈련된 신경망 모형으로부터 몬테카를로 표본을 표집하였다. 표집된 몬테카를로 표본으로부터 예측분산을 도출한 후 이를 가우시안 과정의 예측분산과 비교하였다. 그 결과 몬테카를로 탈락 방법은 데이터의 수에 영향을 받지 않으며 표본의 범위 안에 있는 분포 밖 데이터에 대해 예측분산을 낮게 평가함을 알 수 있다.

참고문헌

[1] Kenji Doya et al. "Social impact and governance of AI and neurotechnologies". In: *Neural Networks* (2022).

[2] Moloud Abdar et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". In: *Information Fusion* 76 (2021), pp. 243–297.

[3] Yarin Gal and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059.

[4] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[5] Alex Kendall and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems* 30 (2017).

[6] David Hall et al. "Probabilistic object detection: Definition and evaluation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 1031–1040.

[7] Rhiannon Michelmore et al. "Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 7344–7350.

[8] Jiaxin Liu et al. "PNNUAD: Perception Neural Networks Uncertainty Aware Decision-Making for Autonomous Vehicle". In: *IEEE Transactions on Intelligent Transportation Systems* (2022).

[9] Seongok Ryu, Yongchan Kwon, and Woo Youn Kim. "A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification". In: *Chemical Science* 10.36 (2019), pp. 8438–8446.

[10] Lewis H Mervin et al. "Uncertainty quantification in drug design". In: *Drug discovery today* 26.2 (2021), pp. 474–489.

[11] Ian Osband. "Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout". In: *NIPS workshop on bayesian deep learning*. Vol. 192. 2016.

[12] Yehao Liu et al. "The Peril of Popular Deep Learning Uncertainty Estimation Methods". In: *Bayesian Deep Learning (BDL) Workshop at NeurIPS 2021*. 2021.

[13] Andreas Damianou and Neil D Lawrence. "Deep gaussian processes". In: *Artificial intelligence and statistics*. PMLR. 2013, pp. 207–215.

[14] Matthias Seeger. "Gaussian processes for machine learning". In: *International journal of neural systems* 14.02 (2004), pp. 69–106.

[15] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[16] MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.

[17] John Paisley, David M. Blei, and Michael I. Jordan. "Variational Bayesian Inference with Stochastic Search". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, 2012, pp. 1363–1370. ISBN: 9781450312851.

[18] The GPy authors. *GPy: A Gaussian process framework in python*. http://github.com/SheffieldML/GPy. 2012–2014.

Abstract in English

With the rapid development of neural network models, from the field of image recognition to drug discovery, neural network models have been introduced in a variety of industrial settings at an unprecedented rate. In line with the advancement, many studies are being actively conducted on methodologies that can predict the uncertainty of neural network models. Due to the fact that it can measure the uncertainty without significantly adjusting the existing model, Monte Carlo dropout has become one of the methods frequently used in the industry and academia where research on uncertainty is needed. On the other hand, the limitations of the Monte Carlo dropout method are being studied theoretically and empirically. This paper introduces the theoretical basis of the Monte Carlo dropout method and analyzes the uncertainty measurement in simulated data.

Keywords: Uncertainty Quantification, Bayesian Neural Network, Deep Gaussian Process

Student Number: 2020-28703