The Korean Journal of Applied Statistics
2023, Vol. 36, No. 2, 115–127

DOI: https://doi.org/10.5351/KJAS.2023.36.2.115
Print ISSN1225-066X / Online ISSN 2383-5818

Variational Bayesian multinomial probit model with Gaussian process classification on mice protein expression level data

Donghyun Son[a], Beom Seuk Hwang[1,a]

[a]Department of Applied Statistics, Chung-Ang University

Abstract

Multinomial probit model is a popular model for multiclass classification and choice model. Markov chain Monte Carlo (MCMC) method is widely used for estimating multinomial probit model, but its computational cost is high. However, it is well known that variational Bayesian approximation is more computationally efficient than MCMC, because it uses subsets of samples. In this study, we describe multinomial probit model with Gaussian process classification and how to employ variational Bayesian approximation on the model. This study also compares the results of variational Bayesian multinomial probit model to the results of naive Bayes, K-nearest neighbors and support vector machine for the UCI mice protein expression level data.

Keywords: variational Bayesian approximation, Gaussian process, multinomial probit model, latent variable

1. 서론

다중 분류(multi-class classification) 문제는 두 개 이상의 이산형 클래스를 종속 변수로 가지는 분류 문제를 말한다. 다중 분류 문제를 해결하기 위해 흔히 나이브 베이즈(naive Bayes) 분류, K-최근접 이웃(K-nearest neighbors), 서포트 벡터 머신(support vector machine)과 같은 분류 방법이 사용되기도 하고, 다항 로지스틱 회귀 모형(multinomial logistic regression model), 다항 프로빗 회귀 모형(multinomial probit regression model)과 같은 통계적 모형을 사용할 수도 있다. 다항 로지스틱 회귀 모형은 통계 모형 중에서는 가장 널리 사용되는 모형이지만 독립성 공리 가정(IIA)을 만족해야만 한다는 점 때문에 모형의 사용이 어려운 경우가 존재한다. 반면, 다항 프로빗 모형의 경우 독립성 공리 가정의 제약을 받지 않는다는 장점이 있어 다항 로짓 모형의 대안으로 사용될 수 있다 (Hausman과 Wise, 1978). Albert와 Chib (1993)은 다항 프로빗 모형을 사용한 다중 분류 모형의 추정을 위해 잠재 변수를 추가하여 깁스 샘플링(Gibbs sampling)으로 사후 분포의 도출이 가능한 방법을 제안하였다. 잠재 변수를 추가하여 사후 분포의 추정이 가능하도록 하는 방법은 가우시안 과정(Gaussian process) 분류 모형에서도 적용이 가능하다 (Neal, 1998). 가우시안 과정 분류 모형은 가우시안 과정 회귀 모형처럼 사전 분포로 가우시안 과정 사전 분포를 사용한다는 점에서 유사하나, 잠재 변수를 도입하고 이 잠재 변수가 가우시안 과정 사전 분포를 따른다는 점에서 차이가 있다 (Williams와 Rasmussen, 2006). 또한,

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2021, and supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2019R1C1C1011710). This paper was prepared by extracting parts of Donghyun Son's master's thesis.

[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail:bshwang@cau.ac.kr

Published 30 April 2023 / journal homepage: http://kjas.or.kr

가우시안 과정 회귀 모형은 정규 분포를 따르는 가능도(likelihood)를 사용하여 사후 분포의 유도 과정에서 적분의 계산이 비교적 쉬운 편이지만, 가우시안 과정 분류 모형은 정규 분포를 따르지 않는 가능도를 사용하여 잠재 변수에 대한 사후 예측 분포를 유도하는 과정에서 적분을 계산하기 어렵다는 특징을 갖는다 (Williams와 Rasmussen, 2006). 가우시안 과정 분류 모형에서 사후 분포를 근사적으로 유도하기 위해 William과 Barber (1998)은 라플라스 근사법을 고안하였고, Neal (1998)은 Markov chain Monte Carlo (MCMC)를 사용하여 잠재 변수를 샘플링하는 방식의 사후 분포 추론 방법을 제안하였으며, Minka (2001)는 expectation propagation (EP) 방법을 제안하였다. 이 중 MCMC 방법은 가우시안 과정 분류 모형뿐만 아니라 적분의 계산이 어려운 다양한 모형에 널리 사용되고 좋은 성능을 보이는 베이지안 추론 방법이지만, 어떠한 값으로 수렴되기까지 시간이 오래 걸린다는 단점이 있다 (Jordan 등, 1999). 이러한 MCMC 방법을 대체하기 위해서 Beal (2003)은 변분 베이지안 근사법(variational Bayes approximation)을 사용한 잠재 변수의 사후 분포 추론 방법을 제안하였다. 가우시안 과정 분류에 대한 변분 베이지안 다항 프로빗 모형은 Girolami와 Rogers (2006)에 의해 소개되었고, Lama와 Girolami (2008)는 해당 방법을 Kote-Jarai 등 (2006)이 연구한 유전자 발현 마이크로어레이(microarray) 데이터에 적용하여 실제로 좋은 분류 성능을 갖는다는 것을 보이기도 하였다.

본 논문에서는 가우시안 과정 다중 분류 모형에 적용할 수 있는 변분 베이지안 근사법을 소개하고 이를 실제 데이터에 적용할 수 있음을 보이고자 한다. 2장에서는 다항 프로빗 회귀 모형을 적용한 가우시안 과정 분류를 소개한다. 3장에서는 변분 베이지안 방법과 가우시안 과정 다중 프로빗 모형에 적용한 변분 근사법을 설명한다. 4장에서는 쥐 단백질 발현 데이터에 가우시안 과정 분류에 대한 변분 베이지안 다항 프로빗 모형을 적용하여 분류 결과와 성능을 확인한다. 5장에서는 본 논문에서 얻은 결론을 요약하여 정리하고 한계점과 후속 연구에 대해 논의한다.

2. 다항 프로빗을 적용한 가우시안 과정 분류

2.1. 가우시안 과정

가우시안 과정(Gaussian process; GP)은 함수에 대한 분포를 나타내기 위해서 유한한(finite) 개수의 확률 변수의 집합을 사용한 결합 정규 분포를 말한다 (Williams와 Rasmussen, 2006). 가우시안 과정은 다변량 정규 분포의 확장이므로 다변량 정규 분포와 같은 성질을 가진다. 예를 들어, 다변량 정규 분포의 주변 분포는 정규 분포를 따르는데, 가우시안 과정으로 나타낸 함수의 주변 분포도 마찬가지로 가우시안 과정을 따른다 (Williams와 Rasmussen, 2006). 또한, 다변량 정규 분포는 평균 벡터와 공분산 행렬로 표현할 수 있는데, 가우시안 과정도 마찬가지로 평균 함수와 공분산 함수를 사용하여 다음과 같이 정의할 수 있다.

$$f(\mathbf{x}) \sim \mathrm{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right).$$

이때, 일반적으로 평균 함수 $m(\mathbf{x})$는 영함수(zero function)로 두고, 공분산 함수 $k(\mathbf{x}, \mathbf{x}')$로 함수의 모양을 결정한다 (Neal, 1998). 공분산 함수로는 흔히 squared exponential (SE) 공분산 함수를 많이 사용하고, 이외에도 Matérn 공분산 함수, ornstein-uhlenbeck (OU) process 공분산 함수, rational quadratic (RQ) 공분산 함수 등 다양한 유형의 공분산 함수를 사용하여 함수에 대한 분포를 나타낼 수 있다 (Williams와 Rasmussen, 2006).

2.2. 가우시안 과정 분류

가우시안 과정은 크게 가우시안 과정 회귀 문제와 가우시안 과정 분류 문제로 구분된다. 먼저, 가우시안 과정 회귀는 연속형 종속 변수 y에 대하여 $y = f(\mathbf{x}) + \varepsilon$이고, 함수 f는 $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w}$로 입력값의 벡터인 $\mathbf{x}$와 가중치 벡터 $\mathbf{w}$의 내적 형태로 주어졌을 때, 함수 $f(\mathbf{x})$의 사전 분포로 가우시안 과정 사전 분포를 설정한 모형을 말한다 (Williams와 Rasmussen, 2006). 다음으로 가우시안 과정 분류 모형을 설명하기에 앞서, 일반적인 이진 분류 모형은 종속 변수 y가 이산형 변수(클래스)이며, 클래스 집합이 $\{-1, +1\}$이라고 할 때 가능도 함수로

$P(y = +1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^T\mathbf{w})$를 갖는 모형을 말한다. 여기서 $\sigma(z)$는 시그모이드 함수를 의미하는데, 만약 시그모이드 함수로 로지스틱 함수를 사용한다면 $\sigma(z) = 1/(1+e^{-z})$가 되고 이를 로지스틱 회귀 모형이라 한다 (Williams와 Rasmussen, 2006). 가우시안 과정 분류 모형은 어떤 잠재 함수 $f(\mathbf{x})$가 사전 분포로 GP를 따른다고 하고, 이 잠재 함수를 이진 분류 모형의 시그모이드 함수 안에 대입한 모형을 말한다. 이를 식으로 나타내면 식 (2.1)과 같다.

$$\pi(\mathbf{x}) \triangleq p(y = +1 \mid \mathbf{x}) = \sigma(f(\mathbf{x})). \tag{2.1}$$

가우시안 과정 분류 모형으로 새로운 데이터 $\mathbf{x}^*$에 대한 잠재 함수의 분포를 계산하는 식은 다음과 같다 (Williams와 Rasmussen, 2006).

$$p(f^* \mid \boldsymbol{X}, \mathbf{y}, \mathbf{x}^*) = \int p(f^* \mid \boldsymbol{X}, \mathbf{x}^*, \mathbf{f})\, p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})\, d\mathbf{f}. \tag{2.2}$$

가우시안 과정 회귀 모형은 가능도 함수와 사후 분포 모두 다변량 정규 분포를 따르고, 식 (2.2)와 같이 적분을 계산할 때 다변량 정규 분포의 성질에 따라 적분의 계산도 비교적 쉽게 가능하게 해준다. 하지만 가우시안 과정 분류 모형은 가능도 함수가 정규 분포를 따르지 않으므로 적분의 계산식을 유도하기 어려운 경우가 발생하기도 한다 (Williams와 Rasmussen, 2006). 이 문제를 해결하기 위해 적분을 근사적으로 계산하는 다양한 방법들이 개발되어 왔다. Williams와 Barber (1998)는 라플라스 근사법, Minka (2001)는 EP 방법, Neal (1998)은 MCMC 근사법을 제안했다. 또한, Beal (2003)은 변분 베이지안 근사법을 제안하였고, 이 방법은 다른 방법들에 비해 시간 복잡도가 낮다는 장점을 가지고 있다 (Girolami와 Rogers, 2006).

2.2.1. 프로빗 함수를 적용한 가우시안 과정 분류

시그모이드 함수로 표준 정규 분포의 누적 분포 함수인 프로빗 함수 $\Phi(z)$를 사용한 프로빗 회귀 모형은 식 (2.1)을 활용하여 다음과 같이 표현될 수 있다.

$$\pi(\mathbf{x}) \triangleq p(y = +1 \mid \mathbf{x}) = \Phi(f(\mathbf{x})),$$

여기서 잠재 함수 $f(\mathbf{x})$가 GP 사전 분포를 따른다고 하면 프로빗 함수를 적용한 가우시안 과정 분류 모형이 된다. 이때 Girolami와 Rogers (2006)는 프로빗 함수를 적용한 가우시안 과정 분류 모형을 다음과 같이 정의하였다. 먼저, 종속 변수 t_n은 $t_n \in \{0, 1\}$인 클래스를 가진다고 정의한다. 다음으로, 잠재 변수를 y_k라는 변수로 두고 다음과 같이 정의하였다.

$$y_k = \sum_j \beta_{kj} x_j + \epsilon = m_k + \epsilon, \quad \epsilon \sim N(0, 1),$$

이때 $(t_n = 1)$인 n 번째 데이터 표본의 가능도 함수는 아래와 같이 계산된다.

$$\begin{aligned} P(t_n = 1 \mid \mathbf{x}_n, \boldsymbol{\beta}) &= \int P(t_n = 1, y_n \mid \mathbf{x}_n, \boldsymbol{\beta})\, dy_n \\ &= \int P(t_n = 1 \mid y_n)\, P(y_n \mid \mathbf{x}_n, \boldsymbol{\beta})\, dy_n \\ &= \int \delta(y_n > 0)\, P(y_n \mid \mathbf{x}_n, \boldsymbol{\beta})\, dy_n, \ \text{where} \ y_n \mid \mathbf{x}_n^T, \ \boldsymbol{\beta} \sim N\left(\mathbf{x}_n^T\boldsymbol{\beta}, 1\right) \\ &= \Phi\left(\mathbf{x}_n^T\boldsymbol{\beta}\right). \end{aligned}$$

위의 이진 분류 프로빗 모형에서 프로빗 함수 안의 $\mathbf{x}_n^T\boldsymbol{\beta}$는 앞서 정의한 바에 의하여 m_k로 나타낼 수 있다. 그 다음 m_k를 GP 확률 변수라고 정의하여 사전 분포로 가우시안 과정을 따르도록 하면 $m_k \sim \text{GP}(\mathbf{0}, \boldsymbol{C}_{\varphi_k})$이

되고, t_n의 클래스를 두 개 이상으로 확장함으로써 다항 프로빗을 적용한 가우시안 과정 분류 모형을 정의할 수 있게 된다. 다항 프로빗 모형에서 사용되는 n 번째 데이터의 k 번째 클래스에 대한 잠재 변수는 y_{nk}이다. 다항 프로빗 모형에서는 잠재 변수로 계산된 값을 사용하여 다중 분류를 진행할 수 있게 된다. 이 잠재 변수를 프로빗 함수에 맞는 형태로 적절히 나타내기 위하여 y_{nk}의 사전 분포를 m_{nk}를 사용하여 나타낼 수 있다. 즉, $y_{nk}|m_{nk} \sim N(m_{nk}, 1)$. 이때 n 번째 관측 데이터에 대하여 각 클래스에 대한 잠재 변수의 값을 모두 계산한 후 그 중에서 가장 큰 값을 갖는 y_{nj}를 찾으면, 그때의 j가 바로 n 번째 데이터가 속하는 클래스가 된다 (Albert와 Chib, 1993).

2.2.2. 모형의 사전 분포와 변수 사이의 위계적 관계

이 장에서는 2.2.1장에서 정의한 가우시안 과정 분류 이진 프로빗 모형을 확장하여 가우시안 과정 사전 분포를 사용한 다항 프로빗 모형을 설명하기 위해 변수들의 사전 분포와 그 관계를 식으로 정의한다 (Girolami와 Rogers, 2006). 먼저, 데이터 행렬 $\boldsymbol{X}_{N\times D} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$, $t_n \in \{1, \ldots, K\}$을 원소로 하는 종속 변수 벡터 $\mathbf{t}_{N\times 1}$, 잠재 변수 y_{nk}로 이루어진 행렬 $\boldsymbol{Y}_{N\times K}$와 m_{nk}로 이루어진 GP 확률 변수 행렬 $\boldsymbol{M}_{N\times K}$, $M \times 1$의 크기를 갖는 공분산 함수의 하이퍼파라미터(hyperparameter) 벡터 $\boldsymbol{\varphi}_k$, 마지막으로 $\boldsymbol{\varphi}_k$의 위계적 사전 분포를 나타내는 벡터 $\boldsymbol{\psi}_k$와 $\boldsymbol{\alpha}_k$를 정의한다. 다음으로, GP 확률 변수 m_k에 대한 사전 분포를 $m_k|\boldsymbol{X}, \boldsymbol{\varphi}_k \sim \mathrm{GP}(\boldsymbol{\varphi}_k) = N_{\mathbf{m}_k}(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\varphi}_k})$으로 정의한다. 여기서 $\boldsymbol{C}_{\boldsymbol{\varphi}_k}$는 클래스별 공분산 행렬이고, 공분산 함수로 어떤 함수를 사용하는가에 따라 $\boldsymbol{C}_{\boldsymbol{\varphi}_k}$가 다르게 정의될 수 있다. 예를 들어, 공분산 함수로 radial basis function (RBF) kernel을 사용한다면, $\boldsymbol{C}_{\boldsymbol{\varphi}_k}$는 $\exp\{-\sum_{d=1}^{M} \varphi_{kd}(x_{id} - x_{jd})^2\}$로 정의된다. 그 다음, 공분산 행렬의 하이퍼파라미터 φ_{kd}에 대한 사전 분포는 $\varphi_{kd} \sim \mathrm{Exp}(\psi_{kd})$으로 설정한다. 그리고 켤레 관계를 활용하여 ψ_{kd}의 사전 분포를 $\psi_{kd} \sim \Gamma(\sigma_k, \tau_k)$로 설정하고, 사전 분포의 하이퍼파라미터인 $\boldsymbol{\alpha} = \{\sigma_{k=1,\ldots,K}, \tau_{k=1,\ldots,K}\}$의 사전 분포로 무정보적 사전 분포를 사용하거나 사전 지식을 활용한 사전 분포를 부여한다.

예를 들어 어떤 데이터가 주어졌을 때, 데이터에서 주어진 클래스의 개수 K가 3이고 전체 데이터의 개수 N은 10개, 주어진 독립 변수의 개수 D는 2개라고 하고, 공분산 함수의 커널로 앞서 설명한 RBF 커널을 사용한다고 하자. 데이터 행렬 $\boldsymbol{X}_{10\times 2} = [\mathbf{x}_1, \ldots, \mathbf{x}_{10}]^T$이고, 종속 변수 벡터는 $\boldsymbol{t}_{10\times 1} = \{t_1, \ldots, t_{10}\}$, $t_n \in \{1, 2, 3\}$이 된다. 종속 변수의 클래스를 결정하는 데에 사용되는 잠재 변수 행렬 $\boldsymbol{Y}_{10\times 3}$와 GP 확률 변수 행렬 $\boldsymbol{M}_{10\times 3}$은 각각 $\boldsymbol{Y}_{10\times 3} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3]$와 $\boldsymbol{M}_{10\times 3} = [\boldsymbol{m}_1, \boldsymbol{m}_2, \boldsymbol{m}_3]$ 행렬로 나타난다. 그리고 여기서 행렬 $\boldsymbol{Y}$와 행렬 $\boldsymbol{M}$의 원소는 y_{nk}와 m_{nk}이다. GP 확률 변수의 공분산 행렬의 하이퍼파라미터로 사용되는 $\boldsymbol{\varphi}_k$는 3개의 클래스에 대해서 각각 $\boldsymbol{\varphi}_1$, $\boldsymbol{\varphi}_2$, $\boldsymbol{\varphi}_3$ 벡터를 가지며, 그 차원은 사용한 공분산 함수 커널의 하이퍼파라미터 개수 M에 따라 달라진다. 앞서 설명한 RBF 커널의 경우에는 하이퍼파라미터의 개수 $M = D$로 독립 변수의 개수와 같고 이에 따라 $\boldsymbol{\varphi}_k$의 차원은 2×1이 된다. 사용하는 공분산 함수 커널의 종류에 관계없이 GP 확률 변수의 공분산 행렬의 차원은 10×10이 된다. $\boldsymbol{\varphi}_k$의 하이퍼파라미터인 $\boldsymbol{\psi}_k$는 3개의 클래스에 대해서 각각 $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$, $\boldsymbol{\psi}_3$ 벡터를 가지며 그 차원은 앞서 결정된 M에 의해서 $M \times 1$로 결정되는데, 예시에서 RBF 커널을 사용하였으므로 $\boldsymbol{\psi}_k$의 차원은 2×1이 된다. $\boldsymbol{\psi}_k$의 하이퍼파라미터인 $\boldsymbol{\alpha}_k$는 3개의 클래스에 대해서 각각 $\boldsymbol{\alpha}_1 = [\sigma_1, \tau_1]$, $\boldsymbol{\alpha}_2 = [\sigma_2, \tau_2]$, $\boldsymbol{\alpha}_3 = [\sigma_3, \tau_3]$ 벡터를 가진다.

위에서 설명한 예시를 통해 모수가 업데이트 되는 순서를 살펴보면, 먼저 클래스 1에 대한 하이퍼파라미터 $\boldsymbol{\alpha}_1$의 원소 σ_1와 τ_1을 사용하여 $\psi_{11} \sim \Gamma(\sigma_1, \tau_1)$과 $\psi_{12} \sim \Gamma(\sigma_1, \tau_1)$을 업데이트 한다. 그리고 ψ_{kd}의 사전 분포로 $\mathrm{Exp}(\psi_{kd})$라는 사전 분포를 설정한 바에 따라 ψ_{11}로 φ_{11}을, ψ_{12}로 φ_{12}를 업데이트 한다. 업데이트 된 φ_{11}와 φ_{12}를 하나의 벡터로 합치면 클래스 1에 대한 GP 확률 변수의 공분산 함수 하이퍼파라미터 벡터 $\boldsymbol{\varphi}_1$을 결정할 수 있다. 결정된 $\boldsymbol{\varphi}_1$는 공분산 함수 커널의 하이퍼파라미터 벡터이므로 이를 통해 GP 확률 변수의 공분산 함수를 업데이트 한다. 클래스 1에 대한 GP 확률 변수 $\boldsymbol{m}_1$은 $\boldsymbol{m}_1 \sim \mathrm{GP}(\boldsymbol{\varphi}_1) = N(\mathbf{0}_{10\times 1}, \mathbf{C}_{\boldsymbol{\varphi}_1})$으로 업데이트 한다. 이제 GP 확률 변수 $\boldsymbol{m}_1$으로 잠재 변수 $\boldsymbol{y}_1$을 업데이트 할 때 사전 분포 $y_{n1} \sim N(m_{n1}, 1)$을 사용한다. 위와 같은 방식으로

클래스 2, 클래스 3에 대해서도 모수의 업데이트를 동일하게 진행한 후 최종적으로 업데이트 된 y_{11}, y_{12}, y_{13}를 비교하여 가장 큰 값을 갖는 값이 예를 들어 y_{13}이라면, 첫 번째 데이터의 클래스는 3으로 분류되게 된다.

다항 프로빗 회귀 모형에 대한 종속 변수 t_n과 잠재 변수 $\mathbf{y}_n$ 사이의 관계는 $t_n = j$ if $y_{nj} = \max_{1\le k\le K}\{y_{nk}\}$로 정의된다. 이를 다른 방식으로 표현하자면, $P(t_n = i|\mathbf{y}_n)$는 $\delta(y_{ni} > y_{nk}, \forall k \neq i)$로 나타낼 수 있고, 따라서 종속 변수에 대한 사후 분포를 다음과 같이 표현할 수 있게 된다.

$$\begin{aligned} P(t_n = i \mid \mathbf{m}_n) &= \int \delta(y_{ni} > y_{nk} \;\forall\; k \neq i) \prod_{j=1}^{K} p\left(y_{nj} \mid m_{nj}\right) d\mathbf{y} \\ &= \int_{\boldsymbol{C}_i} \prod_{j=1}^{K} p\left(y_{nj} \mid m_{nj}\right) d\mathbf{y} \\ &= E_{p(u)}\left\{\prod_{j\neq i} \Phi\left(u + m_{ni} - m_{nj}\right)\right\}. \end{aligned}$$

이때 확률 변수 u는 표준 정규 분포를 따른다. 잠재 변수와 관련된 집합을 $\boldsymbol{\Theta} = \{\boldsymbol{Y}, \boldsymbol{M}\}$라 하고, 하이퍼파라미터의 집합을 $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_{k=1,\ldots,K}, \boldsymbol{\psi}_{k=1,\ldots,K}\}$로 정의하면 결합 가능도는 다음과 같이 나타낼 수 있다.

$$\begin{aligned} p(\mathbf{t}, \boldsymbol{\Theta}, \boldsymbol{\Phi} \mid \boldsymbol{X}, \boldsymbol{\alpha}) = &\prod_{n=1}^{N} \left\{\sum_{i=1}^{K} \delta(y_{ni} > y_{nk},\; \forall\; k \neq i)\, \delta(t_n = i)\right\} \\ &\times \prod_{k=1}^{K} p(y_{nk} \mid m_{nk})\, p(\mathbf{m}_k \mid \boldsymbol{X}, \boldsymbol{\varphi}_k)\, p(\boldsymbol{\varphi}_k \mid \boldsymbol{\psi}_k)\, p(\boldsymbol{\psi}_k \mid \boldsymbol{\alpha}_k). \end{aligned}$$

이 결합 가능도 식을 기반으로 하여 추후에 업데이트하게 될 구체적인 식을 유도한다. 하지만 그 과정에서 적분의 계산을 필요로 하고 그 계산이 어렵기 때문에 적분을 근사적으로 계산할 방법이 필요하게 된다. 따라서 Beal (2003)이 제안한 변분 베이지안 근사법을 다항 프로빗 모형에 적용하기 위한 과정을 3장에서 설명한다.

3. 변분 베이지안 근사법

변분 베이지안 근사법은 근사적인 분포를 구하는 베이지안 방법이다 (Jordan 등, 1999). 대표적인 베이지안 근사법인 마르코프 연쇄 몬테카를로(MCMC) 방법은 여러 분야에서 널리 사용되고 있지만 어떠한 값으로 수렴되기까지 시간이 오래 걸린다는 단점이 있는 반면, 변분 근사법은 계산 속도가 빨라서 MCMC 방법의 대체 방법으로 사용되고 있다 (Jordan 등, 1999). 변분 근사법을 사용한 베이지안 학습의 목적은 크게 두 가지가 있는데 첫째로, 주변 가능도를 근사적으로 구하여 모형 간 비교를 하기 위함이 있고, 둘째로 모형의 모수에 대한 사후 분포를 근사적으로 구하여 예측에 사용하기 위함이 있다 (Beal, 2003).

변분 근사법에서 근사적인 분포를 구하기 위한 핵심 아이디어는 최적화(optimization)이다. 잠재 변수의 근사분포족(a family of approximate densities; Q)으로부터 분포의 후보군 $q(\boldsymbol{\Theta}) \in Q$을 선정하고, 최적화 과정을 통해 후보군 중에서 쿨백-라이블러 발산(Kullback–Leibler divergence)을 최소화하는 분포를 찾아내어 조건부 분포를 대신하는 근사 분포로 사용한다 (Blei 등, 2017). 쿨백-라이블러 발산을 계산하는 과정에는 종속 변수의 주변 확률 밀도 함수를 구하기 위한 결합 확률 밀도 함수의 적분이 포함되어 있는데, 이를 계산하기가 어렵기 때문에 쿨백-라이블러 발산을 다음과 같이 ELBO (evidence lower bound)라는 식으로 대체하여 표현한다.

$$\mathrm{ELBO}(q) = E\left[\log p(\boldsymbol{\Theta}, \mathbf{t} \mid \mathbf{X}, \boldsymbol{\Phi}, \boldsymbol{\alpha})\right] - E\left[\log q(\boldsymbol{\Theta})\right].$$

ELBO(q)가 커지면 쿨백-라이블러 발산이 작아지는 반비례 관계를 활용하여 ELBO(q)를 최대화하는 분포를 찾는다면, 해당 분포를 쿨백-라이블러 발산을 최소화하는 분포로 사용할 수 있게 된다 (Blei 등, 2017).

Girolami와 Rogers (2006)는 변분 근사법을 사용하여 다항 프로빗 가우시안 과정 분류 모형을 추정하였는데, 평균장 근사법(mean-field approximation) (Beal, 2003)을 활용하여 모형의 사후 분포를 인수들의 곱 형태로 나타내었다. 그 식은 $p(\boldsymbol{\Theta}|\mathbf{t}, \boldsymbol{X}, \boldsymbol{\Phi}, \boldsymbol{\alpha}) \approx \prod_i Q(\boldsymbol{\Theta}_i) = Q(\boldsymbol{Y})Q(\boldsymbol{M})$로 표현된다. 해당 식을 활용한 변분 베이지안 다항 프로빗 모형의 시간 복잡도는 $O(\text{KNS}^2)$이고, 깁스 샘플링과 라플라스 근사법의 시간 복잡도는 $O(\text{KN}^3)$인데, 여기서 S는 표본의 부분 집합으로 $S \ll N$ 이기 때문에 변분 근사법이 더 효율적인 방법으로 알려져 있다 (Girolami와 Rogers, 2006).

3.1. 변분 근사법 적용

이번 장에서는 Girolami와 Rogers (2006)의 다항 프로빗 가우시안 과정 분류 모형에 대한 변분 근사법을 소개한다. 앞서, 모형의 사후 분포가 평균장 근사법에 의하여 $p(\boldsymbol{\Theta}|\mathbf{t}, \boldsymbol{X}, \boldsymbol{\Phi}, \boldsymbol{\alpha}) \approx \prod_i Q(\boldsymbol{\Theta}_i) = Q(\boldsymbol{Y})Q(\boldsymbol{M})$로 근사적인 표현이 된다고 하였다. 여기에서, GP 확률 변수($\mathbf{M}$)와 잠재 변수($\mathbf{Y}$)에 대한 분포족 $Q(\mathbf{M}), Q(\mathbf{Y})$은 다음의 식으로 표현된다.

$$Q(\mathbf{M}) = \prod_{k=1}^{K} Q(\mathbf{m}_k) = \prod_{k=1}^{K} N_{\mathbf{m}_k}(\widetilde{\mathbf{m}}_k, \Sigma_k), \tag{3.1}$$

$$Q(\mathbf{Y}) = \prod_{n=1}^{N} Q(\mathbf{y}_n) = \prod_{n=1}^{N} N^{t_n}_{\mathbf{y}_n}(\widetilde{\mathbf{m}}_n, \boldsymbol{I}), \tag{3.2}$$

여기서 틸드 연산자(~)는 사후 기댓값(posterior expectation)을 말하며 $\widetilde{f(a)} = E_{Q(a)}\{f(a)\}$를 뜻한다. 식 (3.1)의 사후 기댓값과 공분산 함수는 k에 따라 $\widetilde{\mathbf{m}}_k = \boldsymbol{\Sigma}_k \widetilde{\mathbf{y}}_k$, $\boldsymbol{\Sigma}_k = \boldsymbol{C}_{\boldsymbol{\varphi}_k}(\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{\varphi}_k})^{-1}$로 계산된다. 그리고 식 (3.2)의 $N^{t_n}_{\mathbf{y}_n}(\widetilde{\mathbf{m}}_n, \boldsymbol{I})$은 잘린 다변량 정규 분포(truncated multivariate Gaussian)를 말하며, 종속 변수 t_n으로 표시되는 클래스의 차원이 가장 큰 값을 갖는다. 여기서, 사후 기댓값 $\widetilde{\mathbf{m}}_k$의 계산에 필요한 잠재 변수의 사후 기댓값 $\widetilde{y}_{nk}$와 $\widetilde{y}_{ni}$는 $k \neq i$에 대해 다음 식을 갖는다.

$$\widetilde{y}_{nk} = \widetilde{m}_{nk} - \frac{E_{p(u)}\left\{N_u(\widetilde{m}_{nk} - \widetilde{m}_{ni}, 1)\,\Phi_u^{n,i,k}\right\}}{E_{p(u)}\left\{\Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nk})\,\Phi_u^{n,i,k}\right\}}, \tag{3.3}$$

$$\widetilde{y}_{ni} = \widetilde{m}_{ni} - \left(\sum_{j \neq i} \widetilde{y}_{nj} - \widetilde{m}_{nj}\right), \tag{3.4}$$

이때 $\Phi_u^{n,i,k} = \prod_{j \neq i,\, k} \Phi(u + \widetilde{m}_{ni} - \widetilde{m}_{nj})$이고, $p(u) = N_u(0, 1)$이다. 다음으로, 하이퍼파라미터 집합인 $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_{k=1,\ldots,K}, \boldsymbol{\psi}_{k=1,\ldots,K}\}$에 속하는 공분산 함수에 대한 하이퍼파라미터 $\boldsymbol{\varphi}_k$의 분포족은 평균장 근사법에 의해 식 (3.5)를 따르고, 사후 기댓값의 추정은 $\varphi^s_{kd} \sim \text{Exp}(\widetilde{\psi}_{kd})$로부터 S개의 표본을 추출하여 중요도 샘플링(importance sampling)을 통해 계산된 근사적인 값으로 진행하여 식 (3.6)을 따르게 된다 (Lawrence 등, 2004). 그리고 ψ_{kd}의 분포는 $Q(\psi_{kd}) = \Gamma_{\psi_{kd}}(\sigma_k + 1, \tau_k + \widetilde{\varphi}_{kd})$이고, ψ_{kd}의 사후 평균은 $\widetilde{\psi}_{kd} = (\sigma_k + 1)/(\tau_k + \widetilde{\varphi}_{kd})$로 계산된다.

$$Q(\boldsymbol{\varphi}_k) \propto N_{\widetilde{\mathbf{m}}_k}\left(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\varphi}_k}\right) \prod_{d=1}^{M} \text{Exp}\left(\varphi_{kd} \mid \widetilde{\psi}_{kd}\right), \tag{3.5}$$

$$\widetilde{f(\boldsymbol{\varphi}_k)} \approx \sum_{s=1}^{S} f(\boldsymbol{\varphi}_k^s)\, w(\boldsymbol{\varphi}_k^s) \quad \text{where} \quad w(\boldsymbol{\varphi}_k^s) = \frac{N_{\widetilde{\mathbf{m}}_k}\left(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\varphi}_k^s}\right)}{\sum_{s'=1}^{S} N_{\widetilde{\mathbf{m}}_k}\left(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\varphi}_k^{s'}}\right)}. \tag{3.6}$$

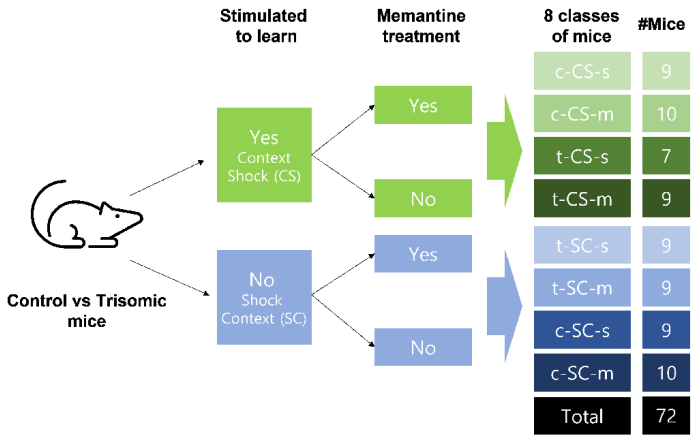


Figure 1: *Eight classes of mice based on genotype (control, c, and trisomy, t), stimulation to learn (context-shock, CS, and shock-context, SC) and treatment (saline, s, and memantine, m).*

마지막으로, 반복을 통해 업데이트 된 $\boldsymbol{\varphi}_k$, $\boldsymbol{\psi}_k$가 고정된 값이라고 가정하고, $\widetilde{\mathbf{m}}_k$와 $\widetilde{\mathbf{y}}_n$을 사용하여 계산한 잠재 변수 집합 $\boldsymbol{\Theta}$에 대한 주변 가능도가 어떤 하한값으로 수렴하게 되고, 그때 계산된 확률로 다중 분류가 이루어지게 된다. 식 (3.8)에서 사용된 $\varphi_{kd}^s \sim \text{Exp}(\widetilde{\psi}_{kd})$와 $w(\boldsymbol{\varphi}_k^s)$는 식 (3.6)에서 설명한 것과 같다. 식 (3.7)에서 사용된 $\mathbf{p}_k$는 $N \times K$ 행렬 $\mathbf{P}$의 k 번째 열벡터를 말한다. 여기서 행렬 $\mathbf{P}$는 p_{nk}를 원소로 가지며 각각의 원소 p_{nk}와 p_{ni}는 $t_n = i$일 때 k (for all $k \neq i$)에 대해서 식 (3.10)과 같이 정의된다. 이 p_{nk}와 p_{ni}는 잠재 변수의 사후 기댓값을 나타내는 식 (3.3)과 식 (3.4)의 가장 오른쪽에 있는 항에 의해서 정의된다.

$$\widetilde{\mathbf{m}}_k \leftarrow \boldsymbol{C}_{\widetilde{\boldsymbol{\varphi}}_k} \left(\mathbf{I} + \boldsymbol{C}_{\widetilde{\boldsymbol{\varphi}}_k}\right)^{-1} \left(\widetilde{\mathbf{m}}_k + \mathbf{p}_k\right), \tag{3.7}$$

$$\widetilde{\boldsymbol{\varphi}}_k \leftarrow \sum_s \boldsymbol{\varphi}_k^s w\left(\boldsymbol{\varphi}_k^s\right), \tag{3.8}$$

$$\widetilde{\psi}_{kd} \leftarrow \frac{\sigma_k + 1}{\tau_k + \widetilde{\varphi}_{kd}}, \tag{3.9}$$

$$p_{nk} = -\frac{E_{p(u)}\left\{N_u\left(\widetilde{m}_{nk} - \widetilde{m}_{ni}, 1\right) \Phi_u^{n,i,k}\right\}}{E_{p(u)}\left\{\Phi\left(u + \widetilde{m}_{ni} - \widetilde{m}_{nk}\right) \Phi_u^{n,i,k}\right\}} \quad \text{and} \quad p_{ni} = -\sum_{j \neq i} p_{nj}. \tag{3.10}$$

3.2. 변분 근사법을 적용한 사후 예측 분포

새로운 표본에 대한 GP 사후 예측 분포 $p(\mathbf{m}^{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t})$는 아래와 같은 평균과 분산을 갖는 K개 정규 분포의 곱 형태로 나타난다 (Girolami와 Rogers, 2006).

$$\widetilde{m}_k^{\text{new}} = \widetilde{\mathbf{y}}_k^T \left(\mathbf{I} + \boldsymbol{C}_{\widetilde{\boldsymbol{\varphi}}_k}\right)^{-1} \boldsymbol{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{\text{new}},$$

$$\widetilde{\sigma}_{k,\text{new}}^2 = c_{\widetilde{\boldsymbol{\varphi}}_k}^{\text{new}} - \left(\boldsymbol{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{\text{new}}\right)^T \left(\mathbf{I} + \boldsymbol{C}_{\widetilde{\boldsymbol{\varphi}}_k}\right)^{-1} \boldsymbol{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{\text{new}}.$$

이때, $\boldsymbol{C}_{\widetilde{\boldsymbol{\varphi}}_k}^{\text{new}}$는 $N \times 1$ 크기의 벡터로 새로 주어진 데이터 포인트와 기존 데이터 $\mathbf{X}$ 사이에서 계산된 공분산 함수의 값을 말하고, $c_{\widetilde{\boldsymbol{\varphi}}_k}^{\text{new}}$는 새로 주어진 데이터 포인트로만 계산된 공분산 함수의 값을 말한다. 그리고 종속 변수로

가능한 k값에 대한 예측 분포는 다음 식으로 나타나게 된다.

$$P\left(t_{new} = k \mid \mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}\right) = E_{p(u)}\left\{\prod_{j\neq k}\Phi\left(\frac{1}{\widetilde{v}_j^{\text{new}}}\left[u\widetilde{v}_k^{\text{new}} + \widetilde{m}_k^{\text{new}} - \widetilde{m}_j^{\text{new}}\right]\right)\right\},$$

여기에서 $u \sim N_u(0,1)$이고, $\widetilde{v}_k^{\text{new}} = \sqrt{1+\widetilde{\sigma}_{k,\text{new}}^2}$ 이다.

4. 실제 데이터 분석

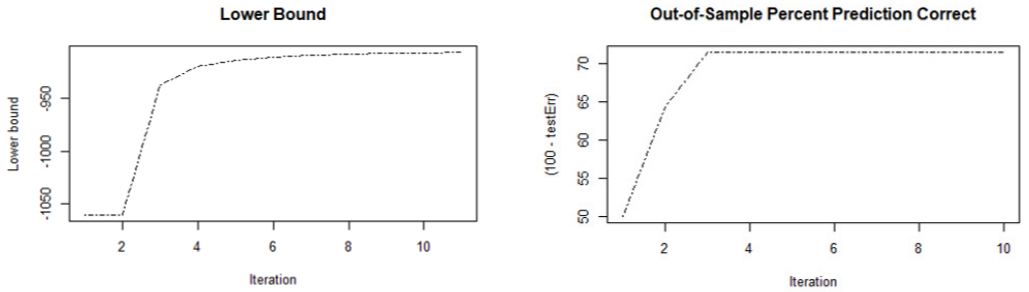
4.1. 데이터 소개

UCI (https://archive.ics.uci.edu/ml/datasets.php)에서 제공하는 쥐의 단백질 발현 데이터는 8가지 클래스에 속하는 쥐들을 어떤 상황에 노출시키고 피질 핵으로부터 77가지 단백질의 발현 수준을 측정한 것으로 자극과 반응에 대한 연합 학습을 평가하는 데 사용된다 (Higuera 등, 2015). 77개의 독립 변수는 모두 연속형 변수로 각 단백질의 발현 수준을 나타내고, 종속 변수는 8가지 클래스로 구성된 범주형 변수이며, 총 1,080개 행으로 구성되어 있다. 종속 변수는 c-CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m으로 c는 대조군을, t는 삼염색체군을 의미하고, CS는 학습을 위한 자극을 부여한 집단, SC는 그렇지 않은 집단을 말하며, s는 식염수(saline)를 투여한 집단, m은 메만틴(memantine)을 투여한 집단을 의미한다. Figure 1를 보면, 쥐 단백질 발현 데이터는 각 클래스에 7마리부터 10마리의 쥐가 속해있어 비교적 균형 잡힌 데이터라고 할 수 있다. 본 분석에서는 개별 측정된 쥐의 77가지 독립 변수를 이용해 쥐가 어떤 클래스에 속하는지 확인하는 다중 분류를 진행하였다. 가우시안 과정 분류에 대한 변분 다항 프로빗 모형을 적용하기 위해 R 소프트웨어 (R Core Team, 2022; ver. 4.2.1)와 R의 vbmp 패키지 (Lama와 Girolami, 2022; ver. 1.64.0)를 사용하였다. vbmp 패키지는 해당 모형을 적용하기 위해 만들어진 R 패키지로, 훈련 데이터와 테스트 데이터가 주어졌을 때 변분 다항 프로빗 모형을 적용한 다중 분류를 수행할 수 있고, 커널 함수로 가우시안, 코시, 라플라스, 내적 커널 등과 같은 다양한 커널을 지원하고 있다 (Lama와 Girolami, 2008).

분석에 앞서 수행한 전처리 과정은 다음과 같다. 먼저, 원 데이터는 종속 변수(target) 포함 총 81개 열로 구성되어 있는데, 이 중 3개의 범주형 변수 칼럼은 각각 종속 변수의 클래스를 설명하는 대조군, 삼염색체군, 학습-자극 집단, 자극-학습 집단, 식염수 투여 집단과 메만틴 투여 집단을 나타내어 분류에 큰 영향을 줄 수 있다. 또한 본 모형에는 범주형 변수의 사용이 불가능하기 때문에 3개의 범주형 변수 열을 제거하였다. 그 다음, 77개 독립 변수 중에서 결측으로 주어진 행은 각 독립 변수의 평균값으로 대체하였다. 마지막으로 본 데이터에서 1,080개의 행은 총 8개 클래스에 속하는 72마리 쥐의 단백질 발현 수준을 각각 15번씩 측정한 것이므로 개별 쥐의 15번 측정에 대한 평균을 구하여 새로운 72개 행의 데이터로 만들었다. 위의 전처리 과정으로 새롭게 생성된 데이터는 72행, 77열의 크기를 갖게 되었으며 해당 데이터를 사용하여 72마리의 쥐를 8가지 클래스로 분류하는 것을 목표로 하였다.

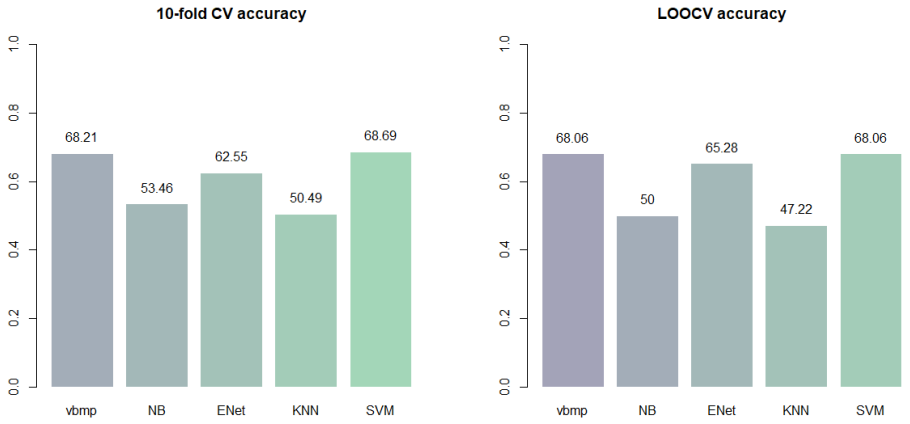
4.2. 분석 결과

가우시안 과정 분류에 대한 변분 베이지안 다항 프로빗 모형을 쥐 단백질 발현 데이터에 적용하여 홀드아웃 교차 검증을 진행한 결과는 Figure 2에서 확인할 수 있다. 훈련 데이터와 검증 데이터의 비율은 8 : 2로 설정하였는데, 그 이유는 종속 변수의 클래스는 8가지로 많은 반면 주어진 관측 데이터는 72개로 적은 편이어서 이보다 더 적은 비율로 훈련데이터를 구성할 경우 학습이 효과적으로 진행되지 않았기 때문이다. 모형의 학습에 사용한 파라미터로 최대 반복수는 10을 지정하였고, 커널로는 내적 커널(inner-product kernel)을 사용하였다. 그리고 주어진 데이터가 변분 모형의 반복을 진행함에 따라 주변 가능도 함수의 하한으로 근사가 되는지 확인하였다. Figure 2(a)를 보면, 약 10번의 반복을 진행했을 때 주변 가능도 함수의 하한값이 −900



(a) Marginal likelihood lower bound (b) Out-of-Sample prediction accuracy

Figure 2: *Hold-Out CV: UCI mice protein expression data.*



(a) 10-fold CV (b) LOOCV

Figure 3: *Cross-Validation: UCI mice protein expression data.*

근처로 근사적으로 접근함을 확인할 수 있고, Figure 2(b)에서 반복에 따른 예측 정확도가 약 72%로 수렴함을 알 수 있다.

본 데이터는 사용 가능한 검증 데이터 개수가 적고 이로 인해 모형의 분산이 커질 수도 있기 때문에 홀드아웃 교차 검증만으로 성능 평가를 진행하기에는 과적합이 될 가능성이 크다. 따라서, 모형의 분산을 줄이면서 조금 더 균형 잡힌 방식으로 성능을 평가하기 위해서 10-fold 교차 검증과 함께, 분산은 다소 크게 나타나서 과적합의 우려가 있지만 모형의 편향을 줄일 수 있는 LOOCV (leave-one-out cross-validation)도 진행하였다. 그리고 그 결과를 나이브-베이즈, K-최근접 이웃법, elastic-net, 서포트 벡터 머신의 10-fold 교차 검증, LOOCV의 성능과 비교해 보았다.

Figure 3(a)에서 변분 다항 프로빗 모형의 10-fold 교차 검증 정확도가 약 68.21%로 나타남을 알 수 있었다. 이는 다른 분류기 중 가장 성능이 뛰어난 서포트 벡터 머신의 정확도 68.69%보다는 다소 낮았지만, 두 모형의 성능이 거의 비슷하게 나타남을 확인할 수 있다. 그리고 서포트 벡터 머신을 제외한 나이브-베이즈, K-최근접 이웃법 등의 분류기와는 성능 차이가 두드러지게 나타났다. Figure 3(b)에서는 변분 다항 프로빗 모형의 LOOCV 성능이 약 68.06%로 나타남을 확인할 수 있었다. 이는 10-fold 교차 검증에서 가장 성능이 좋

Table 1: LOOCV results from VBMP

		Predicted class								
		c-CS-m	c-CS-s	c-SC-m	c-SC-s	t-CS-m	t-CS-s	t-SC-m	t-SC-s	Total
True class	c-CS-m	**5**	4	0	0	0	0	0	1	10
	c-CS-s	5	**1**	0	0	1	1	0	1	9
	c-SC-m	0	0	**10**	0	0	0	0	0	10
	c-SC-s	1	0	0	**7**	0	0	1	0	9
	t-CS-m	0	1	0	0	**8**	0	0	0	9
	t-CS-s	0	1	0	0	1	**5**	0	0	7
	t-SC-m	0	0	2	1	0	0	**5**	1	9
	t-SC-s	0	0	0	0	0	0	1	**8**	9
	Total	11	7	12	8	10	6	7	11	72

은 분류기인 서포트 벡터 머신의 LOOCV 정확도 68.06%와 동일한 수치로 두 모형의 성능이 크게 차이 나지 않고 거의 유사하게 나타난다는 점을 다시 한번 확인할 수 있었다. 또한, 나이브-베이즈의 정확도는 50.0%, K-최근접 이웃법은 47.22%로 나타나 변분 다중 프로빗 모형의 성능에 비해 크게 떨어짐을 알 수 있었다.

Table 1과 Table 2는 LOOCV를 진행하면서 모든 쥐를 한 번씩 테스트 데이터로 사용한 결과를 종합하여 각 클래스에 대한 정분류 개수와 오분류 개수를 보여준다. Table 1은 변분 다항 프로빗 모형의 분류 결과이며, Table 2는 서포트 벡터 머신의 분류 결과를 나타낸 것이다. 먼저, c-CS-s 클래스에 속하는 쥐의 분류가 변분 다항 프로빗 모형과 서포트 벡터 머신에서 모두 공통적으로 잘 이루어지지 않아 분류 정확도를 떨어뜨린 점을 확인할 수 있었다. Table 1의 변분 다항 프로빗 모형의 분류 결과의 경우, c-CS-m 클래스 중 40%에 달하는 4마리의 쥐를 c-CS-s 클래스로 분류하고, 반대로 c-CS-s의 56%에 달하는 5마리의 쥐를 c-CS-m으로 분류하는 등 c-CS-m 클래스와 c-CS-s 클래스를 분류하는 데에 어려움을 겪은 것으로 보인다. 하지만 나머지 클래스 중 c-SC-m, c-SC-s, t-CS-m 클래스의 분류에 있어서는 서포트 벡터 머신보다 더 정확하게 분류를 진행하였다. 특히 변분 다항 프로빗 모형은 c-SC-m 클래스에 속하는 10마리의 쥐를 모두 정확하게 분류함을 확인할 수 있고, 총 9마리의 t-CS-m 클래스에 속하는 쥐 중에서 8마리를 정확하게 분류함을 확인할 수 있었다. 반면, Table 2에서는 c-CS-m 클래스와 c-CS-s 클래스를 분류하는 데에 있어 서포트 벡터 머신이 변분 다항 프로빗 모형보다 더 정확한 결과를 도출함을 확인하였다. 이때 서포트 벡터 머신은 변분 다항 프로빗 모형과 달리 c-CS-m, c-CS-s 두 클래스 간의 분류에 어려움을 겪은 것이 아니라, t-CS-m 클래스까지 포함하여 세 가지 클래스를 분류하는 데에 어려움을 겪은 것으로 보인다. 그리고 서포트 벡터 머신은 c-SC-m, c-SC-s, t-CS-m 클래스에 대한 분류에서 변분 다항 프로빗 모형에 비해 분류를 정확하게 진행하지 못했는데, 특히 t-CS-m 클래스의 분류를 정확하게 진행하지 못하였다. 하지만 나머지 클래스들인 t-SC-m과 t-SC-s의 분류는 서포트 벡터 머신이 변분 다항 프로빗 모형보다 더 정확하게 분류를 진행하였으며, 총 9마리의 쥐가 속해있는 t-SC-s 클래스의 분류를 모두 정확하게 분류하였음을 확인할 수 있다.

5. 결론

본 연구에서는 다중 분류 모형에 프로빗 함수를 사용하고 사전 분포로 가우시안 과정을 따르도록 하는 가우시안 과정 분류 다항 프로빗 모형에 변분 베이지안 근사법을 적용하는 방법에 주목하였다. 변분 베이지안 근사법은 잠재 변수의 조건부 사후 분포를 근사적으로 구하는 것을 목적으로 하여, 근사적인 분포를 구하기 위해서 근사분포족을 설정하고 이로부터 분포의 후보군을 선정하여 쿨백-라이블러 발산을 최소화하는 분포를 최적화 과정을 통해 찾아낸다.

본 연구에서는 UCI의 쥐 단백질 발현 데이터에 변분 근사법을 적용하여 그 정확도를 다른 다중 분류기

Table 2: LOOCV results from SVM

		Predicted class								
		c-CS-m	c-CS-s	c-SC-m	c-SC-s	t-CS-m	t-CS-s	t-SC-m	t-SC-s	Total
True class	c-CS-m	**6**	2	0	0	2	0	0	0	10
	c-CS-s	3	**2**	0	0	3	1	0	0	9
	c-SC-m	0	0	**9**	0	0	0	1	0	10
	c-SC-s	0	0	1	**6**	0	0	1	1	9
	t-CS-m	1	1	0	0	**6**	1	0	0	9
	t-CS-s	0	1	0	0	1	**5**	0	0	7
	t-SC-m	0	0	3	0	0	0	**6**	0	9
	t-SC-s	0	0	0	0	0	0	0	**9**	9
	Total	10	6	13	6	12	7	8	10	72

와 비교하였다. 예측 정확도 면에서 변분 근사법은 68.21%의 정확도를 보였다. 이는 나이브 베이즈 방법의 53.46%, elastic-net의 62.55%, K-최근접 이웃법의 50.49%보다 더 높았고, 서포트 벡터 머신의 68.69%보다는 다소 낮지만 비슷한 성능을 보였다. 이러한 점에서 변분 근사법이 다중 분류 문제에 대한 하나의 대안으로써 사용될 수 있다는 결론을 얻을 수 있었다. 또한, 해당 분석은 표본의 개수가 적은 상황에서 진행한 분석으로 변분 베이지안 근사법이 적은 표본 수 하에서도 좋은 성능을 보일 수 있다는 관점을 제시할 수 있다. 본 연구를 통해 변분 베이지안 근사법의 예측 성능을 확인할 수 있었지만, 다항 프로빗 모형에서 잠재 변수의 사전 분포를 가우시안 과정으로 두는 제약 조건에 의하여 독립 변수를 연속형 변수로 제한한다는 점은 한계점으로 지적될 수 있다. 향후 연구에서는 사전 분포를 가우시안 과정으로 제한하지 않고 더 다양한 분포에 대해서 적용하기 위한 방법을 연구하고자 한다.

References

Albert JH and Chib S (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**, 669–679.

Beal MJ (2003). *Variational Algorithms for Approximate Bayesian Inference*. University of London, University College London (United Kingdom).

Blei DM, Kucukelbir A, and McAuliffe JD (2017). Variational inference: A review for statisticians, *Journal of the American Statistical Association*, **112**, 859–877.

Girolami M, and Rogers S (2006). Variational Bayesian multinomial probit regression with Gaussian process priors, *Neural Computation*, **18**, 1790–1817.

Hausman JA and Wise DA (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences, *Econometrica: Journal of the Econometric Society*, **46**, 403–426.

Higuera C, Gardiner KJ, and Cios KJ (2015). Self-Organizing feature maps identify proteins critical to learning in a mouse model of down syndrome, *PloS One*, **10**, e0129126.

Jordan MI, Ghahramani Z, Jaakkola TS, and Saul LK (1999). An introduction to variational methods for graphical models, *Machine Learning*, **37**, 183–233.

Kote-Jarai Z, Matthews L, Osorio A *et al.* (2006). Accurate prediction of BRCA1 and BRCA2 heterozygous genotype using expression profiling after induced DNA damage, *Clinical Cancer Research*, **12**, 3896–3901.

Lama N and Girolami M (2008). Vbmp: Variational Bayesian multinomial probit regression for multi-class clas-

sification in R, *Bioinformatics*, **24**, 135–136.

Lama N and Girolami M (2022). _vbmp: Variational Bayesian Multinomial Probit Regression_. R package version 1.64.0, Available from: http://bioinformatics.oxfordjournals.org/cgi/content/short/btm535v1

Lawrence ND, Milo M, Niranjan M, Rashbass P, and Soullier S (2004). Reducing the variability in cDNA microarray image processing by Bayesian inference, *Bioinformatics*, **20**, 518–526.

Minka TP (2001). *A family of algorithms for approximate Bayesian inference* (Doctoral dissertation), Massachusetts Institute of Technology, Cambridge, MA, USA.

Neal RM (1998). Regression and classification using gaussian process priors. In AP Dawid, M Bernardo, JO Berger, and AFM Smith (Eds), *Bayesian Statistics 6* (pp. 475–501), Oxford University Press, New York.

R Core Team (2022). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Williams CK and Barber D (1998). Bayesian classification with Gaussian processes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1342–1351.

Williams CK and Rasmussen CE (2006). *Gaussian Processes for Machine Learning*, MIT press, Cambridge, MA.

Received November 23, 2022; Revised December 13, 2022; Accepted December 16, 2022

가우시안 과정 분류에 대한 변분 베이지안 다항 프로빗 모형: 쥐 단백질 발현 데이터에의 적용

손동현[a], 황범석[1,a]

[a]중앙대학교 응용통계학과

요 약

다항 프로빗 모형은 다중 분류와 선택 모형에서 흔히 사용하는 모형이다. 다항 프로빗 모형을 추정하기 위해 일반적으로 널리 사용하는 베이지안 접근법인 마르코프 연쇄 몬테카를로(MCMC) 방법은 계산 복잡도가 매우 높다는 문제점을 가지고 있다. 반면, 변분 베이즈 방법은 MCMC 방법보다 계산 복잡도는 낮으면서도 분류 성능적인 면에서 큰 차이가 나지 않아 더 효율적인 방법으로 알려져 있다. 본 연구에서는 가우시안 과정에 기반한 다항 프로빗 모형을 설명하고 해당 모형에 적용할 수 있는 변분 베이지안 근사법을 알아보고자 한다. 그리고 UCI에서 제공되는 쥐 단백질 발현 데이터에 가우시안 과정 분류에 대한 변분 베이지안 다항 프로빗 모형을 적용하여 그 성능을 확인하고 나이브 베이즈, K-최근접 이웃법, 서포트 벡터 머신 분류기의 성능과 비교한다.

주요용어: 가우시안 과정, 다항 프로빗 모형, 변분 베이즈 방법, 잠재 변수

이 논문은 2021년도 중앙대학교 CAU GRS 지원에 의하여 작성되었고, 2019년도 정부 (과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2019R1C1C1011710).

이 논문은 손동현의 석사논문의 일부를 발췌하여 작성하였음.

[1]교신저자: (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과. E-mail: bshwang@cau.ac.kr