Physica A 513 (2019) 112–125

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Parameter inference in a probabilistic model using clustered data

Hirohito Kiwata

Division of Natural Science, Osaka Kyoiku University, Kashiwara, Osaka 582-8582, Japan

ARTICLE INFO

Article history:
Received 14 July 2018
Available online xxxx

Keywords:
Inverse ising problem
Inference of parameters in a probabilistic model
Clustered data samples
Mean-field theory
Pseudo-likelihood method

ABSTRACT

We propose a method to infer the parameters of a probabilistic model from given data samples. Our method is based on the pseudolikelihood and composite likelihood methods. We cluster the given data samples and apply the clustered data samples to the pseudo-likelihood and composite likelihood methods. From an expansion of the pseudolikelihood method around the mean of a cluster, the mean-field and Thouless–Anderson–Palmer equations are derived. Likewise, from an expansion of the composite likelihood method around the mean of a cluster, a method that is similar to the Bethe approximation is derived. We then perform numerical simulations using our method. We find that our method gives an accurate estimate in the range of weak coupling parameters but has an inferior accuracy compared to the pseudolikelihood and composite likelihood methods in the range of strong coupling parameters. In the range of strong coupling parameters, as the number of clusters increases, the inference accuracy of our method improves. Compared to the pseudolikelihood and composite likelihood methods, our method reduces the number of computational tasks for the estimation, therefore, sacrificing the inference accuracy.

1. Introduction

With the increase in computer performances and the prevalence of computers, it is easy to obtain large amounts of data. Feature extraction methods have attracted much attention in applications to investigate the relationships contained within such data. In particular, the interpretation of the statistical properties of data has been a major focus [1–3]. We assume that a probabilistic model generates the given data and investigate methods to determine the parameters of this model. As the amount of data increases, we can more accurately predict the properties behind the data by using a suitable method. However, the number of computational tasks for the probabilistic inference increases as the amount of data increases. Moreover, an increase in the number of parameters of a probabilistic model increases the difficulty of the probabilistic inference. When the number of parameters of a probabilistic model or the amount of data increases, it becomes difficult to determine the properties of the data. Therefore, even if computer performances continue to progress, probabilistic inference will remain a difficult task. There is a dilemma regarding the treatment of the data. In this paper, we propose a method to reduce the dilemma.

The maximum likelihood method is a standard method for inferring the parameters of a probabilistic model from given data samples. The parameters of a probabilistic model are determined to maximize the likelihood function. Because it is impossible to simultaneously obtain the optimum parameters, an iterative method is adopted to evaluate the parameters. In the iteration process, the parameters are gradually tuned to match the expectations of the empirical distribution with those of a probabilistic model. However, it is difficult to evaluate the expectations of a probabilistic model barring exceptions.

E-mail address: kiwata@cc.osaka-kyoiku.ac.jp.

https://doi.org/10.1016/j.physa.2018.08.158
0378-4371/

To settle this problem, one alternative to the maximum likelihood method is the pseudolikelihood method [4–8]. In the pseudolikelihood method, the likelihood function is replaced with the product of a full conditional. The pseudolikelihood method has two outstanding properties: consistency and convexity [9,10]. The pseudolikelihood method surpasses the maximum likelihood method for a low number of computational tasks. However, the pseudolikelihood method requires many computational tasks to accurately estimate the parameters.

The mean-field method from statistical physics significantly reduces the number of computational tasks for estimation [11–14]. However, the mean-field method is restricted in its applicability. When the correlation between random variables is strong, the mean-field method fails to infer the parameters of a probabilistic model. The reason for this failure originates in the origin of the mean-field method. To improve the estimation accuracy, there is a simple approximate inference method known as the Bethe approximation or belief propagation [15–21]. The Bethe approximation outperforms the mean-field method for the inference of parameters. The mean-field method and the Bethe approximation are interpreted in an expansion with respect to the parameters and a partial summation of their higher-order terms [3]. Even though both methods have the advantage of a small number of computational tasks, they fail in the range of strong coupling parameters and have an inferior estimation accuracy compared to the pseudolikelihood method. To improve the inference accuracy of the mean-field method, the application of clustered data samples has been proposed [22,23]. The idea for such an application originates from a physical consideration, which we consider to be heuristic. The mean-field method with clustered data samples gives excellent results for restricted cases.

In the present paper, we apply the pseudolikelihood method to clustered data samples. When there exist large data samples of a probabilistic model with many parameters, the pseudolikelihood method requires a large number of computational tasks to obtain an accurate estimate. The large number of computational tasks of the pseudolikelihood method originates from the sum over all the data samples. To relax the number of computational tasks, we cluster the data samples and divide the sum into a sum over the clusters and a sum over the data samples in one cluster. In our method, we repeat the sum over the clusters in an iterated evaluation, which reduces the number of computational tasks needed to infer the parameters. We show that the mean-field method and the Thouless–Anderson–Palmer (TAP) equation with clustered data samples are derived from the pseudolikelihood method. There is another method known as the composite likelihood method, which corresponds to a generalization of the pseudolikelihood method [24–26]. Even though the composite likelihood method gives a superior estimation accuracy compared to the pseudolikelihood method, it is inferior with regard to the number of computational tasks required. We apply clustered data samples to the composite likelihood method and obtain a method similar to the Bethe approximation.

2. Theory

Consider a set of N discrete numbers, $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$. This set $\boldsymbol{x}$ is generated by an unknown probability distribution. We want to estimate the parameters in a probabilistic model from the given data $\boldsymbol{x}$. To explain our idea concretely, we prescribe x_i to be binary, i.e., $x_i \in \{-1, +1\}$. The prescription for x_i does not restrict the effectiveness of our method. Then, the set is given by $\boldsymbol{x} \in \{-1, +1\}^N$. We assume that the data $\boldsymbol{x}$ are generated by the following probability distribution:

$$P(\boldsymbol{x}|\{J_{ij}\}, \{h_i\}) = \frac{1}{Z(\{J_{ij}\}, \{h_i\})} \exp\Big(\frac{1}{2}\sum_{i\neq j}^{N} J_{ij}x_i x_j + \sum_{i=1}^{N} h_i x_i\Big), \tag{1}$$

where J_{ij} is a coupling parameter and h_i is a bias [27,28]. The denominator $Z(\{J_{ij}\}, \{h_i\})$ is a partition function, which is introduced to normalize the probability distribution. We consider a graph consisting of N vertices with edges connecting each vertex. The ith element of $\boldsymbol{x}$, i.e., x_i, is located at the ith vertex. There is an edge between the ith and jth vertices in the case where $J_{ij} \neq 0$. For simplicity, we assume that J_{ij} is symmetric, i.e., $J_{ij} = J_{ji}$, and set $J_{ii} = 0$. The above probabilistic model is equivalent to a graph and corresponds to a Boltzmann machine without hidden vertices. The problem is to infer a set of $\{J_{ij}\}$ and $\{h_i\}$ from M sets of $\boldsymbol{x}$, i.e., $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(M)}\}$. For the inference of the parameters in a probabilistic model, the maximum likelihood estimation is useful. Given $\mathcal{D}$, we define a histogram corresponding to the empirical distribution function, such as

$$Q(\boldsymbol{x}) = \frac{1}{M}\sum_{\mu=1}^{M} \delta(\boldsymbol{x}, \boldsymbol{x}^{(\mu)}), \tag{2}$$

where $\delta(\boldsymbol{x}, \boldsymbol{x}')$ is the Kronecker delta, which is equal to unity when $\boldsymbol{x} = \boldsymbol{x}'$ and zero when $\boldsymbol{x} \neq \boldsymbol{x}'$. Using the empirical distribution function, we define the log-likelihood as

$$\begin{aligned}\mathcal{L}(\{J_{ij}\}, \{h_i\}) &= \sum_{\boldsymbol{x}\in\{-1,+1\}^N} Q(\boldsymbol{x}) \ln P(\boldsymbol{x}|\{J_{ij}\}, \{h_i\}) = \frac{1}{M}\sum_{\mu=1}^{M} \ln P(\boldsymbol{x}^{(\mu)}|\{J_{ij}\}, \{h_i\}) \\ &= \frac{1}{M}\sum_{\mu=1}^{M}\Big(\frac{1}{2}\sum_{i\neq j}^{N} J_{ij}x_i^{(\mu)}x_j^{(\mu)} + \sum_{i=1}^{N} h_i x_i^{(\mu)}\Big) - \ln Z(\{J_{ij}\}, \{h_i\}).\end{aligned} \tag{3}$$