Model Selection via Fisher Information Selection Approach

Abstract

One of the most important problem in machine learning is "over-fitting". Various "information criterias", for instance, the Akaike information criteria (AIC) and the Bayesian information criteria (BIC), have been proposed to deal with over-fitting via a penalty term to compensate more complex models. However, the AIC and the BIC assume that each parameter makes equal contribution to the model complexity, i.e. each parameter is equally important to the model, and regard the number of model parameter as the measure of model complexity, which is not suitable for many non-linear models, such as the Boltzmann Machines (BMs). This paper introduce a FIS (Fisher information selection) approach to guide model selection. The FIS is more general since we assume that each parameter has different importance with respect to the model. Specifically, the FIS compensates harder (or softer) to parameters with lower (or higher) fisher information. We apply the FIS in a series of VBM density estimation experiments. Experimental results indicate that the FIS achieves a better performance than the AIC and the BIC.

Introduction

From a Bayesian perspective, the over-fitting problem can be avoided by marginalizing over the model parameters instead of making point estimate of their values (Posada and Buckley 2004). The optimal model is given by the trade-off between the likelihood and model complexity. Specifically, the Bayesian approach achieves the optimal model by maximize the model evidence, i.e. the evidence maximization framework (MacKay and others 1995). Akaike proposed a Akaike information criteria(AIC), which was actually derived from the maximum evidence framework, to guide the model selection (Akaike 1981). Another important Bayesian model selection approach is the Bayesian information criteria(BIC) which penalizes model complexity more heavily than the AIC.

However, the AIC and the BIC assume that each model parameter contributes equally to the model complexity. Both of them use the number of model parameters to represent the model complexity. In many non-linear models, for instance, the Boltzmann machines (BMs), and the recurrent neural networks (RNN) (Williams and Zipser 1989), the above assumption is not correct (e.g. a BM with K free parameters can be topologically different with another BM with K free parameters, thus the complexities of these two BMs can be extremely different).

Inspired by the evidence maximization framework, this paper introduce a fisher information selection (FIS) to guide the model selection. The FIS is more general since we assume that each parameter contributes differently to model complexity, i.e. each parameter is differently important to the model. Specifically, the FIS compensates harder (or softer) to parameters with lower (or higher) fisher information. According to Hou, the confidence of a parameter can be measured by the corresponding fisher information. Our purpose is to save parameters with higher confidence and cut parameters with lower confidence. We apply the FIS in a series of visible Boltzmann machine(VBM) density estimation experiments. Experimental results indicate that the FIS achieves a better model than the AIC and the BIC.

The rest of this paper is organized as follows. Section 2 reviews the evidence maximum framework and information criterias. Then the FIS is proposed in Section 3. Experimental results on VBM density estimation are presented in Section 4.

Evidence Maximization Framework

By adopting a full Bayesian approach, the over-fitting problem can be avoided. However, running a full Bayesian approach is impratical since completely marginalizing over the model space is analytically intractable. Suppose there are a set of candidate models $\{\mathrm{M_j}\}$, where j = 1, 2,..., T and the training data D is generated from one of these models but we uncertain which one. A simple idea is choosing the most probable model, i.e. the model with the maximum evidence $\mathrm{p}\left(D|M_j\right)$, to approximate the overall model space.

Many interesting methods were derived from the evidence maximization framework, e.g. the Akaike information criteria (AIC) and the Bayesian information criteria (BIC) which are given by $\mathrm{AIC} = -2\log p\left(D|\theta\right) + 2K$ and $\mathrm{BIC} = -2\log p\left(D|\theta\right) + Klog\left(N\right)$, where K denotes the number of model parameters , N denotes the sample size and $\theta = (\mathrm{w_1}, \mathrm{w_2}, \ldots, \mathrm{w_K})$ denotes the free parameters.

These information criteria indicators of model performances use the number of model parameter to represent the model complexity. Thus, models with equal number of mod-

el parameters but different topologies suffer the same penalty. However, their complexities and the distributions they represented can be completely different.

In next Section, we will introduce the fisher information selection (FIS). The FIS considers the topology of the specific model. Specifically, according to the fisher information (**?**), we treat each model parameter in different ways.

Model Selection based on Fisher Information

We can get some insight into the over-fitting problem. Suppose a parameter is added into a model with suitable complexity. The model will be tuned to fit the random noise on the training data. And the magnitude of the parameters will typically get larger. As the model complexity increases, the parameters becomes finely tuned to fit the data by developing larger and larger magnitude so that the corresponding distribution matches each of the data point exactly (Bishop 2006). Hence, for a model exhibiting over-fitting, if we cut off a parameter, the magnitude of other parameters should be significantly decreased. But for a model with a suitable complexity, cutting off a parameter only makes little differences to the magnitude of other parameters.

Inspired by this observation, we propose a new model selection method. As we can see in this section, the magnitude of parameters can be accessed to the number of effective parameters. We use the changing of the number of effective parameters, which is indeed caused by the changing of the magnitude of parameters, to guide the model selection.

Number of Effective Parameters

The effective parameters are parameters determined by the training data. In this section, we attempt to estimate the number of effective parameters via the evidence maximization framework which was discussed in section 2. Note that, every model parameter is determined simultaneously by the training data and the prior distribution (or regulation terms). Thus, the number of effective parameters need not to be an integer. We use the number of effective parameters to measure how far the model is determined by the training data and how far the model is determined by the prior distribution (or regulation terms).

Denote the mean of the posterior distribution of model parameters as m. Typically, in order to evaluate the evidence function, we can assume that the variable distribution and the prior distribution of parameter vector θ satisfy $\mathrm{p}\left(X|\theta,\beta\right) = \prod_{n=1}^{N} N\left(X_n|\mu_n,\beta^{-1}\right)$ and $\mathrm{p}\left(\theta|\alpha\right) = \mathrm{p}\left(w|\alpha\right) = N\left(w|0,\alpha^{-1}\right)$, where X is the variable, μ_n and β are the mean and the precision of the variable distribution respectively, α is the precision of the prior distribution of θ. Then, maximizing the evidence function result in the close form solution of the number of effective parameters: $m^T m$ (Bishop 2006).

In order to treat each parameter in different way, our method alternatively assumes the prior distributions of free parameters to be different Gaussian distributions. Specifically, denote the mean of the posterior distribution of the i-th parameter w_i as m_i, we assume the prior distribution of w_i is $\mathrm{p}\left(\theta|\alpha_i\right) = \mathrm{p}\left(w_i|\alpha_i\right) = N\left(w_i|0,\alpha_i^{-1}\right)$ where α_i is the precision of the prior distribution of w_i. Furthermore, we set the precision α_i to be the corresponding eigenvalue of the fisher information matrix F (later in this paper we will discuss the motivations of setting α_i in this way). After a series of derivations, under the evidence maximization framework we obtain the number of the effective parameters $m_N^T F m_N = \sum \alpha_i m_i^2$ where $m_N = (\mathrm{m}_1, \mathrm{m}_2, \ldots, \mathrm{m}_K)$. Assuming the posterior distributions of model parameters are peaked, we can use θ_{MAP} to approximate m_N. Usually, the assumption does not hold. But we can still use it since the variation trend of θ_{MAP} is similar to m_N and we only make use of the variation of $m_N^T F m_N$ to guide the model selection. Furthermore, for a correct selection process, the posterior distribution must be peaked when the searching comes to the final stage.

Fisher Information Selection

Denote the number of effective parameters of the current model as γ_t. Suppose the original model is exhibiting over-fitting and the number of effective parameters of the original model is γ_0. In order to simplify the computation, we can use the fisher information matrix of the original model to evaluate the number of effective parameters of all candidate models since the fisher information matrix only rely on the training data. Hence, one of the approximation of the difference between γ_t and γ_0 is given by the fisher information distance $FID = \left(\theta_0 - \theta_{\mathbf{t}}\right)^T F \left(\theta_0 - \theta_{\mathbf{t}}\right)$. If we cut off a parameter w_c from an over-fitting model, the FID would be influenced by two factors:

- decreasing factor: a positive term $\alpha_c(w_{ct} - w_{c0})^2$ is removed from the FID
- increasing factor: the magnitude of other parameters is significantly decreased, so, for i $\neq$ c , $\alpha_i(w_{it} - w_{i0})^2$ increases

Intuitively, when the current model is still over-fitting, the model tailoring would result in an increasing to the difference between γ_t and γ_0, i.e. the FID is increasing in the early stage of model selection. Thus, the increasing of FID is related with the decreasing of the model complexity. We explore this relation, and give the following model selection method

- (a) travel all parameters w_i of the current model, finding the w_c which gives the largest FID_{new}, where the FID_{new} is the FID of the model given by cutting w_i from the current model;
- (b) cut off w_c;
- (c) repeat (a), (b) until the largest FID_{new} is close to (or even smaller than) the FID of the current model;

We called this model selection method fisher information selection (FIS). Next, we get some insight into the variation of the FID and discuss the advantages of the FIS.

We analyze the variation of FID from the decreasing factor and the increasing factor respectively.

decreasing factor Every cut parameter w_c satisfies that the removed term $\alpha_c(w_{ct} - w_{c0})^2$ is relatively small so that the decreasing factor is small (CIF).

increasing factor When the model is still exhibiting over-fitting, the magnitude of other parameters should be significantly decreased. Thus, the increasing factor is relatively large.

Along with the model tailoring, the increasing of FID becomes insignificant (even decreasing) for two reasons:

- all of the remained parameters are relatively confident, i.e. for all parameters, $\alpha_c(w_{ct} - w_{c0})^2$ is large. Thus, the decreasing factor is large;
- the model is tailored into a suitable complexity, i.e. cutting off an extra parameter only makes small differences in the magnitude of other parameters. Thus, the increasing factor is small.

Once the decreasing factor is larger than the increasing factor, the model selection process breaks. Thus, after the fisher information selection, the remained parameters are confident and the over-fitting problem is no longer severe.

Further Selection

The FIS would break once the model is tailored into a suitable complexity. However, there might be a range of models whose FIDs are close and complexities are different. We alter the FIS so that we can find the model with the optimal complexity. Instead of the FID, we use the KFratio $KFratio = \frac{K}{FID}$ to guide the model selection, where K is the number of parameters of the current model. The model with the minimum KFratio would be finally selected. As we have discussed, if the decreasing factor is close to the increasing factor, i.e. the FID is increasing in a small step, FIS would break. However, using KFratio, if the FID is nearly invariant and the numerator K is still decreasing, the model selection would continue. Thus, the tailoring would continue until the KFratio increases. Actually, KFratio makes a trade-off between FID and K.

Experimental study

In this section, we experimentally investigate the FIS model selection in density estimation tasks for the visible Boltzmann (VBM). We compare the FIS with the AIC, the AICc, and the BIC, and show that in our experiments the model given by the FIS model selection gives better performance.

Experiment Setting

The artificial binary data set: we first randomly select the target distribution $q(x)$, which is randomly chosen from the open probability simplex over the n random variables using the Dirichlet prior. Then the data set with N samples are generated from $q(x)$. Four methods are compared: the AIC, the corrected AIC (Hurvich and Tsai 1993), the BIC and the FIS.

The K-L divergence is used to evaluate the goodness-of-fit of the VBMs selected via the four approaches. In order to accurately and effectively evaluate the K-L divergence, the artificial data set is set to be 10-dimensional, i.e. there are totally 45 connections and 10 biases in the full VBM. For all experiments, we run 20 randomly generated distributions and report the average K-L divergence. The sample size is from 100 to 1000. Note that, for a full VBM with 10 variables, a training set with less than 1000 data points can result in an extremely over-fitting result.

Result

The average K-L divergences between the distributions represented by the models and the target distributions are shown in Table 1. Comparing these four model selection approach and the full VBM, the FIS achieves better result.

sample size	100	300	300	1000
Full VBM	0.95050	0.86965	0.80686	0.70240
AIC	0.89561	0.80697	0.78933	0.68205
corrected AIC	0.87041	0.78463	0.77690	0.67295
BIC	0.88639	0.81802	0.78369	0.69773
FIS	0.69773	0.75911	0.71973	0.67052

Table 1: The averaged K-L divergences

We can see that the model selected by the FIS shows a better performance than the information criterias in fitting the target distributions especially when the original model is extremely over-fitting (see the averaged K-L divergence on data set with 100 data points). Here we explain this observation: there are many unconfident parameters in an extremely complex model. The information criterias does not consider the confidence of the parameters. They assume that each parameter contributes equally to the model complexity. However, intuitively, an unconfident parameter contributes more to the powerless complexity and the model suffers much from cutting a parameter with high confidence. The FIS encourages saving the parameters with high confidence and cutting the parameters with low confidence. Thus the model selected by the FIS saves more confident parameters.

References

Akaike, H. 1981. Likelihood of a model and information criteria. *Journal of econometrics* 16(1):3–14.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

Hurvich, C. M., and Tsai, C.-L. 1993. A corrected akaike information criterion for vector autoregressive model selection. *Journal of time series analysis* 14(3):271–279.

MacKay, D. J., et al. 1995. Ensemble learning and evidence maximization. In *Proc. Nips*, volume 10, 4083. Citeseer.

Posada, D., and Buckley, T. R. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology* 53(5):793–808.

Williams, R. J., and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1(2):270–280.