MEM-NN: A CLASSIFICATION NUERAL NETWORK BASED ON MAXIMUM ENTROPY METHOD

Mincong Luo
Information Centre
China Institute of Atomic Energy
Peking,China
luomincentos@ciae.ac.cn

Li Liu
Reactor Engineering Technology Research Division
China Institute of Atomic Energy
Peking,China
liuli92@ciae.ac.cn

ABSTRACT

In this paper we propose a method call MEM-NN: a classification nueral network based on Maximum Entropy Method. The main idea is that when only partial knowledge about the unknown distribution is grasped, the probability distribution that meets this knowledge but has the largest entropy value should be selected. Which is used to construct a objective function to train a neural network to do images classification. We show it performs well on MNIST dataset compared to baseline.

1 AN INTRODUCTION TO MEM

The principle of maximum entropy is a criterion for selecting the statistical characteristics of random variables that best meets the objective conditions, also known as the principle of maximum information Phillips et al. (2006) Dudik (2007) FIELDING & BELL (1997) Pietra et al. (1996). The probability distribution of random quantities is difficult to measure. Generally, only various mean values (such as mathematical expectation, variance, etc.) or values under certain defined conditions (such as peak value, number of values, etc.) can be measured. The distribution of these values can be measured in a variety of ways, or even infinitely. Usually, one of the distributions has the largest entropy. Selecting this distribution with the largest entropy as the distribution of the random variable is an effective processing method and criterion. Although this method has certain subjectivity, it can be considered as the most suitable choice for objective situations Elith et al. (2006) Elith et al. (2015). When investing, it is often said that you should not put all your eggs in one basket, which can reduce the risk. This principle applies equally in information processing. In mathematics, this principle is called the principle of maximum entropy Charniak (2000) Och & Ney (2002) Och & Ney (2002).

The principle of maximum entropy was proposed by E.T. Jaynes in 1957. The main idea is that when only partial knowledge about the unknown distribution is grasped, the probability distribution that meets this knowledge but has the largest entropy value should be selected. Because in this case, there may be more than one probability distribution that fits the known knowledge. We know that entropy defines the uncertainty of a random variable. When the entropy is maximum, it indicates that the random variable is the most uncertain Och & Ney (2002) Papineni et al. (1998) Liu et al. (2016). In other words, the random variable is the most random, and it is most difficult to accurately predict its behavior.

The entropy for the conditional distribution $P(Y|X)$ is:

$$H(P) = \sum_{x,y} P(y,x) log P(y|x) = \sum_{x,y} \widetilde{P}(x) P(y|x) log P(y|x)$$

If the constraint is satisfied and then the entropy is maximized, where the maximum entropy model P^* is:

$$P^* = arg \max_{P \in C} H(P) \;\; or \;\; P^* = arg \min_{P \in C} -H(P)$$

In summary of above, the formulation of maximum entropy model could be given: given data set $\{(x_i, y_i)\}_{i=1}^{N}$, feature function $f_i(x, y)i = 1, 2, n$,a model set C that satisfies the constraint set based on the empirical distribution can be given:

$$\begin{aligned} &\min_{P \in C} \quad \sum_{x,y} \widetilde{P}(x) P(y|x) log P(y|x) \\ &s.t. \quad E_p(f_i) = E_{\widetilde{P}}(f_i) \\ &\qquad \sum_{y} P(y|x) = 1 \end{aligned}$$

The model is finally formalized as an optimization problem with constraints, which can be transformed into unconstrained optimization problems by Lagrange multiplier method, introducing Lagrange multipliers: $\lambda_0, \lambda_1, , \lambda_n$, define the Lagrange function $L(P, \lambda)$:

$$\begin{aligned} L(P, \lambda) &= -H(P) + \lambda_0 \left(1 - \sum_{y} P(y|x)\right) + \sum_{i=1}^{n} \lambda_i (E_{\widetilde{P}}(f_i) - E_p(f_i)) \\ &= \sum_{x,y} \widetilde{P}(x) P(y|x) log P(y|x) + \lambda_0 \left(1 - \sum_{y} P(y|x)\right) + \sum_{i=1}^{n} \lambda_i \left(\sum_{x,y} \widetilde{P}(x, y) f(x, y) - \sum_{x,y} \widetilde{P}(x) p(y|x) f(x, y)\right) \end{aligned}$$

Now the problem can be formalized as a very small problem that is easy for Lagrangian dual processing: $\min_{P \in C} \max_{\lambda} L(P, \lambda)$.

2 THE THEOREM OF CLASSIFICATION METHOD BASED ON MEM

In this section we will discuss our proposed theorem for classification algorithm based on maximum entropy method and predictive statistical mechanics. we will propose two theorems and the proofs are also given in the following.

In traditional statistical mechanics, the theory is based on the equation of particle motion. In order to apply statistical methods, some assumptions, such as the ergodicity property and the principle of equal probability, are added. The principle of entropy and entropy increase is the result obtained. Jaynes Principle (2014) Dewar & Detering (2010) believes that we can take the opposite approach, that is, entropy theory should be made.

The so-called predictive statistical mechanics is to regard statistical mechanics as statistical inference based on incomplete basic knowledge rather than physical theory, then all the results can be derived from the principle of maximum entropy, that is, the known knowledge as a constraint The maximum value of entropy is obtained to obtain the probability distribution of the system, and then all the thermodynamic quantities are calculated. It should be noticed that the entropy in the principle of maximum entropy refers to information entropy. So we can understand the problem as the following mathematical formulations(discrete case):

Find the maximum:

$$S_I = -\sum_{i} p_i \ln \frac{p_i}{m_i}$$

constraints:

$$\sum_{i} p_i = 1$$

$$< f_r(x) >= \sum_{i} p_i f_r(x)$$

where the $f(x)$ is a known map, p_i is probablity for $x = x_i$, m_i is the measure. we transform this into unconstrained optimization problems by Lagrange multiplier method, the objective function is calculated and let its derivative to p_i be 0:

$$p_i = m_i \exp\{-\lambda_0 - \sum_{r} \lambda_r f_r(x)\}$$

to prove the corresponding S_I to p_i above is the maximum value, we propose the following theorem:

Lemma 1 *For arbitrary two probability distributions* $\{q_i\}$,$\{p_i\}$*:*

$$-\sum_i q_i \ln \frac{q_i}{m_i} \leq -\sum_i q_i \ln \frac{p_i}{m_i}$$

Proof 2.1 *The above is equivalent to* $\sum_i q_i \ln \frac{q_i}{p_i} \geq 0$*, which can be easily proved let* $x = \frac{q_i}{p_i}$ *in* $x \ln(x) \geq x - 1, (x \geq 0)$.

Theorem 1 *the maximum value for the information entropy:*

$$S_{Imax} = \lambda_0 + \sum_r \lambda_r < f_r(x) >$$

Proof 2.2 *now we assume that the* p_i *refers to the solution for* $p_i = m_i \exp\{-\lambda_0 - \sum_r \lambda_r f_r(x)\}$ *and* q_i *refers to any solution. using lemma.1:*

$$\begin{aligned} S_{Iq} &= -\sum_i q_i \ln \frac{q_i}{m_i} \leq -\sum_i q_i \ln \frac{p_i}{m_i} \\ &= -\sum_i q_i \{-\lambda_0 - \sum_r \lambda_r f_r(x)\} \\ &= \lambda_0 \sum_i q_i + \sum_r \lambda_r \sum_i q_i f_r(x_i) \\ &= \lambda_0 + \sum_r \lambda_r < f_r(x) > \end{aligned}$$

in another aspect:

$$\begin{aligned} S_{Ip} &= -\sum_i p_i \ln \frac{p_i}{m_i} \\ &= -\sum_i p_i \{-\lambda_0 - \sum_r \lambda_r f_r(x_i)\} \\ &= \lambda_0 + \sum_r \lambda_r < f_r(x) > \end{aligned}$$

in this way we can get:

$$S_{Iq} \leq S_{Ip}$$

define the partition function:

$$Z(\lambda_r, r = 1, ..., m) = \sum_i m_i \exp\{-\sum_r \lambda_r f_r(x_i)\}$$

then we can get the following by normalized condition:

$$p_i = \frac{m_i}{Z} \exp\{-\sum_r \lambda_r f_r(x_i)\}$$

$$\lambda_0 = \ln Z$$

The next theorem will tell us the average observation $< f_r(x) >$ is the derivative of $\ln Z$ relative to λ_r, which means, $< f_r(x) >$ is the distribution information about x along λ_r.

Theorem 2 *The formula for average observation:*

$$< f_r(x) >= \frac{\partial}{\partial \lambda_r} \ln Z$$

Proof 2.3

$$\begin{aligned}\frac{\partial \ln Z}{\partial \lambda_r} &= \frac{1}{Z}\frac{\partial Z}{\partial \lambda_r} \\ &= \frac{1}{Z}\sum_i m_i \exp\{-\sum_r \lambda_r f_r(x_i)\}(-f_r(x_i)) \\ &= -\sum_i p_i f_r(x_i) = - < f_r(x) >\end{aligned}$$

Now let's have a look at the difference between the cost function for the softmax and the mem, and in the next section we will test the method on MNIST dataset.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \cdot log(p(y^{(i)} = j|x^{(i)}; \theta))\right]$$

$$\begin{aligned}S_{Ip} &= -\sum_i p_i\{-\lambda_0 - \sum_r \lambda_r f_r(x_i)\} \\ &= \lambda_0 + \vec{\lambda} \odot (\vec{p} \otimes f(\vec{x}))\end{aligned}$$

3 EXPERIMENT

The Dateset: We make the experiment on the MNIST dataset Deng (2012): has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

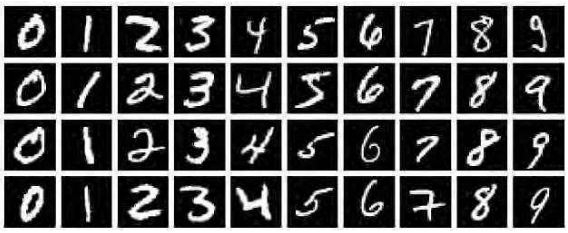


Figure 1: The MNIST database contains 60,000 training images and 10,000 testing images. Half of the training set and half of the test set were taken from NIST's training dataset, while the other half of the training set and the other half of the test set were taken from NIST's testing dataset.

The Baseline Model: The baseline model we use is the softmax method.

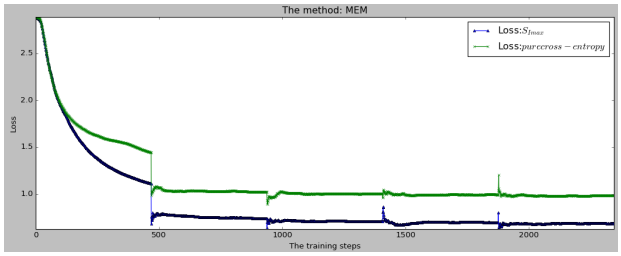
The Result: The result shows that our method (as the target cost function) can better converge more quickly and reach higer accuracy compared to the baseline model.

4 CONCLUSION

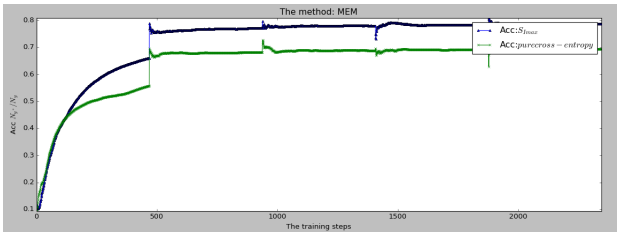
In this paper, we propose a method for images classification called MEM-NN(Neural Network With Maximum Entropy Method). We prove what the S_{Imax} is and use it to construct a objective function to train a neural network to do images classification. We show that the proposed model performs better on the MNIST dataset compared to the baseline model. In the future work, we will extend this method to more machine learning tasks.

REFERENCES

Eugene Charniak. A maximum-entropy-inspired parser. *Proc Naacl*, volume 84(96):132–139, 2000.



(a) Overview



(b) Representation Vector

Figure 2: **(a) : The loss curve of the models**. In the task, our method (as the target cost function) can better converge more quickly. **(b) : The Accuracy curve of the models**. our method can reach higer accuracy compared to the baseline model.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Robert L Dewar and Frank Detering. *Complex physical, biophysical and econophysical systems :*. World Scientific,, 2010.

Miroslav Dudik. *Maximum entropy density estimation and modeling geographic distributions of species*. Princeton University, 2007.

J Elith, Ch Graham, Rp Anderson, M Dudik, S Ferrier, A Guisan, Rj Hijmans, F Huettmann, Jr Leathwick, and A Lehmann. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudk, Yung En Chee, and Colin J. Yates. A statistical explanation of maxent for ecologists. *Diversity and Distributions*, 17(1):43–57, 2015.

Alan H. FIELDING and JOHN F. BELL. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ Conser*, 24(1):38–49, 1997.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. pp. 507–516, 2016.

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Meeting on Association for Computational Linguistics*, pp. 295–302, 2002.

K. A. Papineni, S. Roukos, and R. T. Ward. Maximum likelihood and discriminative training of direct translation models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 189–192 vol.1, 1998.

Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231–259, 2006.

Stephen A. Della Pietra, Stephen A. Della Pietra, and Stephen A. Della Pietra. *A maximum entropy approach to natural language processing*. MIT Press, 1996.

Jaynes' Maximum Entropy Principle. *Jaynes' Maximum Entropy Principle, Riemannian Metrics and Generalised Least Action Bound*. 2014.