A Sample-based Criterion for Unsupervised Learning of Complex Models beyond Maximum Likelihood and Density Estimation

Mani Manavalan[1], **Praveen Kumar Donepudi**[2]

[1]Technical Project Manager, Larsen & Toubro Infotech (LTI), Mumbai, **INDIA**
[2]Enterprise Architect, Information Technology, UST-Global, Inc., Ohio, **USA**

ABSTRACT

Many unsupervised learning processes have the purpose of aligning two probability distributions. Recoding models like ICA and projection pursuit, as well as generative models like Gaussian mixtures and Boltzmann machines, can be seen in this perspective. For these types of models, we offer a new sample-based error measure that can be used even when maximum likelihood (ML) and probability density estimation-based formulations can't be used, such as when the posteriors are nonlinear or intractable. Furthermore, the challenges of approximating a density function are avoided by our sample-based error measure. We show that with an unconstrained model, (1) our technique converges on the correct solution as the number of samples increases to infinity, and (2) our approach's predicted answer in the generative framework is the ML solution. Finally, simulations of linear and nonlinear models on mixtures of Gaussians and ICA issues are used to evaluate our approach. Our method's applicability and generality are demonstrated by the experiments.

Keywords: Unsupervised learning, Maximum likelihood, Gaussian mixtures, Boltzmann machines

12/30/2016 **Source of Support:** Nil, **No Conflict of Interest:** Declared

INTRODUCTION

Many unsupervised learning algorithms can be thought of as attempts to align two probability distributions (Dayan et al., 1995). The generative and recoding models are two well-known types of unsupervised methods that can be cast this way. In a generative unsupervised framework, the environment generates training examples—which we will call observations—by sampling from one distribution while the model embodies the other. Mixed Gaussians (MoG) (Duda and Hart, 1973), factor analysis (Erdogmus and Principle, 2000), and Boltzmann machines (Hinton and Sejnowski, 1986) are examples of generative frameworks. The model translates points from an observation space to an output space in the recoding unsupervised framework, and then compares the output distribution to a reference distribution or a distribution derived from the output distribution. Independent component analysis (leA) (Hyvarinen, 1999) is an example of a method for discovering a representation of vector-valued data that minimizes statistical dependence among the vector elements in the output space.

The output distribution is compared against a factorial distribution derived either from distribution assumptions (e.g., super gaussian) or from a factorization of the output distribution using ICA. Projections methods like projection pursuit (Zhao and Atkeson, 1996) and principal component analysis are other examples within the recoding framework. The goal of unsupervised model learning in each of the cases we've discussed is to align two probability distributions, one or both of which are generated by the model. To enhance the model, we must first construct a measure of the disparity between the two distributions, as well as understand how the model parameters affect the discrepancy (Manavalan, 2014). Using model outputs to build a probability density estimator (PDE) is a natural technique.

The main downside of this method is that the learning procedure's accuracy is heavily dependent on the PDE's quality. The bias-variance trade-off is a problem that PDEs have to deal with. Maximum likelihood (ML), which eliminates PDEs, is a prominent technique for learning generative models. The model's generative distribution is represented analytically in an ML method, making it easy to evaluate the posterior, (data I model), and tweak the model parameters to maximize the likelihood of the data being generated by the model. This restricts the application of machine learning to models with tractable posteriors, which is only true for the most basic models (Dayan et al., 1995; Ghahramani and Hinton, 1998; Hinton and Sejnowski, 1999).

A method that, like ML is used, avoids the development of an explicit PDE while yet allowing the posterior to be expressed analytically. Our method, which we refer to as a sample-based method, assumes a set of samples from each distribution and offers an error measure of disagreement that is defined directly in terms of the samples. As a result, a PDE or analytic expression of the model's density is replaced by a second set of samples selected from the model. The theory of electric fields, which describes the interactions between charged particles, inspired the sample-based technique. Hochreiter and Mozer (2000) is a good place to start to learn more about the metaphor.

Objectives of the Study

In this study, we show that, assuming an unconstrained (maximally flexible) model, our approach converges to the best solution as the sample size grows to infinity. We further show that the ML solution in a generative setting is the expected outcome of our approach. The sample-based strategy works for both linear and nonlinear models, according to empirical evidence.

LITERATURE REVIEW

Unsupervised learning

There's a lot of information in raw data. Images, movies, text, and audio samples all exist in high-dimensional areas and are highly structured, resulting in a signal that is rich and complex (Donepudi, 2014b). Natural photographs have thousands of bits of information, according to Shannon's information theory (Shannon, 1948). In contrast, labels, such as those used in picture categorization, frequently contain only a few bits of information. This drives researchers to develop models that can make use of unlabeled data's information. Yann LeCun put it this way:

”Unsupervised learning accounts for the majority of human and animal learning. Unsupervised learning would be the cake, while supervised learning would be the icing on the cake, if intelligence were a cake. We know how to make the icing, but we're stumped as to how to prepare the cake. Before we can even consider genuine AI, we need to address the unsupervised learning problem.”

— Yann LeCun, NeurIPS 2016 invited talk

Intuition tells us that without labels, we should be able to figure out how the data is organized. Observing images of an unfamiliar object, for example, will enable the reader to recognize it in subsequent photographs but not to name it. This is a fascinating research topic with many yet undiscovered corners since a virtually endless amount of raw data is available for the first time in human history (Bynagari, 2016). Learning from unlabeled data is, however, an ill-posed task, and converting it into a problem that can be solved necessitates specifications. The first is how to choose or enhance a model in the absence of objective aims, and the second is what kind of model can be effective in this situation (Donepudi, 2014a). The research presented in this paper is part of the broad field of unsupervised learning, which seeks to answer these concerns.

While labels only carry a few pieces of information, they are crucial in defining the supervised method. They can be thought of as a condensed overview of how humans perceive parts of the incoming data (Bynagari, 2014). This high-level semantic data on the input serves as a target for iterative model improvement (Donepudi, 2015). This is absent in unsupervised learning, and it is a priori unknown how to design a useful optimization goal. As in a supervised technique, one frequent method is to extract some structure from the raw data and use it as a target label. Self-supervised training is the term for this method. An image, for example, can be chopped into pieces and the correct arrangement utilized as a goal (Doersch et al., 2015, Noroozi and Favaro, 2016).

Generative modelling

Each data point is viewed as the result of a random experiment in generative modeling. Tossing a dice several times, with the data being the collection of values obtained, is a simple example. Natural images gathered by a web crawler on the internet can be viewed as realizations of a more complicated random event. The goal is to learn a probability density model that is as close to the data producing distribution as possible. Direct uses of this type of approach include data compression (Huffman, 1952) and data production. It's also logical to infer that, in order to perform well, such a model would need to provide rich, abstract data representations (Bynagari, 2015).

The ability to predict the outcome of a random experiment with a low level of uncertainty necessitates some level of comprehension. A human with a solid command of the English language, for example, will be able to predict missing letters from a text with far greater accuracy than chance (Shannon, 1951). This is due to the fact that a high degree of English comprehension implicitly presents the reader with a low entropy model. Taking a few patches out of an image and asking a human to predict their content is a comparable experiment. Because of a strong grasp of how natural images are likely to be constructed, an accuracy far better than random will be reached (Manavalan, 2016).

To fit the data-generating distribution, the model must first comprehend the data and construct appropriate representations of it. These representations can then be utilized to address additional problems, such as by training them with large amounts of unlabeled data and then refining them with smaller amounts of tagged data. The principle in this situation is that the abstractions learned by the generative model should be better representations of the data than its raw representation in the input space, and so it is dubbed representation learning; for an overview (Bengio et al., 2012.)

Applications of generative models.

Information compression, which is one of the backbones of telecommunications, is one of the direct practical uses (Donepudi, 2016). Compression algorithms are based on the idea that relatively likely occurrences should be connected with short messages, whereas unusual events should be communicated with longer messages. This necessitates a density model of the data being communicated, which is a common use of density estimation. Recent improvements have made it possible to train such a model on extremely complicated data, which holds the potential of significant compression advantages in the telecommunications industry (Bengio et al., 2012).

Another obvious and valuable application is the generation of realistic-looking fictional data, which is possible if the trained generative model allows for the sampling of new data (Manavalan & Ganapathy, 2014). For example, after training on real situations, the goal could be to generate realistic visual data to populate a virtual environment. This is beneficial in movies and video games, and it could become an important feature of virtual worlds. More realistically, generative modeling is a suitable machine learning sandbox problem. Indeed, there is a nearly unlimited amount of low-cost data available, and over-fitting is unlikely to be an issue in the near future because the signal being fitted is rich and complicated (Manavalan & Bynagari, 2015).

Experimental Methods

The sample-based technique is demonstrated for two common unsupervised learning problems: MoG and ICA. We show that in both circumstances, the sample-based technique works in the linear case. We also look at a nonlinear situation to show how powerful the sample-based technique is.

Mixture of Gaussians

m signifies a mixed component chosen with probability Vm among M components in this generative model framework, and has related model parameters $w_m = (\Omega_m, \mu_m)$. The (linear) model output in the basic MoG model is derived by $x^i = fw_m\,(z^i) = \Omega_m z^i + \mu_m$, where z^i is selected from the Gaussian distribution with zero mean and identity covariance matrix, given a choice of component m. For $fw_m\,(z^i)$, we employed a 3-layer sigmoidal neural network for a nonlinear mixture model (z^i). For our technique, we can develop an update rule for $\Delta v_m = -\epsilon_v \sum_{i=1}^{N_x} (z^i)^T \frac{\partial z^i}{\partial x^i} x^i$, where ϵ_v is step size and $\sum_{m=1}^{M} v_m = 1$ is applied.

Results and Discussion

We used the traditional expected maximization (EM) strategy to train a linear MoG model (using code from [5]), as well as our samplebased approach to train a linear and nonlinear MoG. All models employed a fixed training set of N$_y$ = 100 samples, and all models had M = 10 except one nonlinear model, which had M = 1. Following each training epoch, we generated 100 samples from our model (the X^i) using the sample-based technique. Backpropagation was used to train the nonlinear model.

The outcomes are depicted in Figure 1, that shows (upper panel, left to right) training samples from a ring density, a larger sample from this density, and the solutions obtained from the linear model trained with EM; (lower panels) models trained with the sample-based method (left to right): linear model, nonlinear model, nonlinear model with one component; The sample-based model loses out to the linear ML model. That's not surprising given that ML calculates model probability values analytically (the posterior is tractable), but our technique approximates model probability values using only samples. In each epoch, we only utilized 100 model samples, and the linear sample-based model generated an acceptable answer that isn't much poorer than the ML model. Nonlinear models fit the genuine ring-like distribution better and avoid sharp corners and edges.

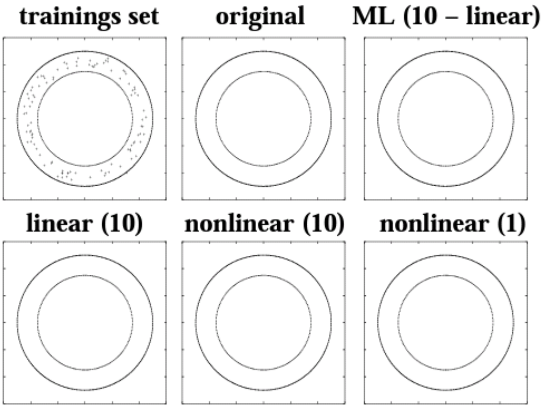

Figure 1: Training samples chosen from a ring density, a larger sample

Independent Component Analysis

We attempted to demix sub-gaussian source distributions with supergaussian modes using a recoding technique. Subgaussian sources are inaccessible to most ICA approaches. The results are practically flawless, as seen in Figure 2. When the mixing and demixing matrices are multiplied, the optimal outcome is a permuted and scaled identity matrix. Hochreiter and Mozer (2000) has further information. We attempted to recover sources from two nonlinear mixings in a subsequent experiment. Because typical rcA methods are built for linear mixings, this problem is impossible to solve with them. Figure 3 illustrates the outcome. Because nonlinear ICA has no single solution, an accurate demixing cannot be expected. Hochreiter and Mozer (2000) has further information. Sources (first row), mixtures (second row), and sources recovered by our method (third row) are projected onto a two-dimensional plane for a three-dimensional linear mixture as shown in Figure 2. When the demixing matrix is multiplied by the mixing matrix, the result is presented in Table 1:

Table 1: Demixing Matrix

Source	Mixture	Recovered Source
-0.0017	0.0010	0.2523
-0.0014	0.1850	-0.0101
-0.1755	0.0003	0.0053

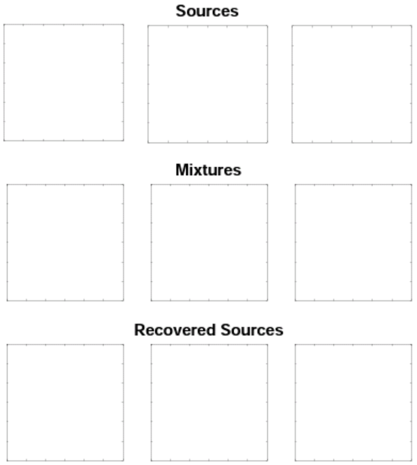

Figure 2: Three-dimensional linear mixture projections of source

Figure 3: For two two-dimensional nonlinear mixing functions

Although our sample-based technique appears to be simple, it has two drawbacks: (1) close samples should be avoided since they result in unbounded gradients; and (2) when computing the force on a data point, all samples must be examined, making the approach computationally demanding. Approximations are proposed in Hochreiter and Mozer (2000) and Gray and Moore (2001), which lower the approach's computing complexity.

We presented simulations in this study that demonstrated the generality and power of our sample-based approach to unsupervised learning issues, as well as two key features of the approach: (1) The technique will find the right answer if certain assumptions are made. (2) The expected result of our technique with an unconstrained model is the ML solution. In conclusion, our sample-based approach can be used to develop unsupervised complicated models in situations when machine learning fails, and our method avoids the shortcomings of PDE approaches.

CONCLUSION

An approach that avoids the creation of an explicit PDE while yet allowing the posterior to be represented analytically, similar to how ML is employed. Our sample-based method assumes a set of samples from each distribution and provides a disagreement error measure that is defined directly in terms of the samples. As a result, a second set of samples chosen from the model replaces a PDE or analytic expression of the model's density. In this paper, we presented simulations that demonstrated the applicability and power of our sample-based approach to unsupervised learning difficulties, as well as two crucial elements of the strategy: (1) If specific assumptions are made, the technique will identify the correct solution. (2) The ML solution is the expected result of our technique with an unconstrained model. Finally, in circumstances where machine learning fails, our sample-based methodology can be utilized to create unsupervised sophisticated models, and our method avoids the drawbacks of PDE approaches.

References

Bengio, Y., Courville, A. C. and Vincent, P. 2012. Unsupervised feature learning and deep learning: A review and new perspectives. CoRR.

Bynagari, N. B. (2014). Integrated Reasoning Engine for Code Clone Detection. *ABC Journal of Advanced Research*, *3*(2), 143-152. https://doi.org/10.18034/abcjar.v3i2.575

Bynagari, N. B. (2015). Machine Learning and Artificial Intelligence in Online Fake Transaction Alerting. *Engineering International*, *3*(2), 115-126. https://doi.org/10.18034/ei.v3i2.566

Bynagari, N. B. (2016). Industrial Application of Internet of Things. *Asia Pacific Journal of Energy and Environment*, *3*(2), 75-82. https://doi.org/10.18034/apjee.v3i2.576

Dayan, P., Hinton, G. E., Neal, R. M. and Zemel. R. S. 1995. The Helmholtz machine. Neural Computation, 7(5):889-904.

Doersch, C., Gupta, A. and Efros. A. A. 2015. Unsupervised visual representation learning by context prediction. ICCV.

Donepudi, P. K. (2014a). Technology Growth in Shipping Industry: An Overview. *American Journal of Trade and Policy*, *1*(3), 137-142. https://doi.org/10.18034/ajtp.v1i3.503

Donepudi, P. K. (2014b). Voice Search Technology: An Overview. *Engineering International*, *2*(2), 91-102. https://doi.org/10.18034/ei.v2i2.502

Donepudi, P. K. (2015). Crossing Point of Artificial Intelligence in Cybersecurity. *American Journal of Trade and Policy*, *2*(3), 121-128. https://doi.org/10.18034/ajtp.v2i3.493

Donepudi, P. K. (2016). Influence of Cloud Computing in Business: Are They Robust?. *Asian Journal of Applied Science and Engineering*, *5*(3), 193-196. Retrieved from https://journals.abc.us.org/index.php/ajase/article/view/1181

Duda R. O. and Hart, P. E. 1973. Pattern Classification and Scene Analysis. Wiley.

Erdogmus D. and Principe J. C. 2000. Comparision of entropy and mean square error criteria in adaptive system training using higher order statistics. In P. Pajunen and J. Karhunen, editors, Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland, pages 75-80. Otamedia, Espoo, Finland, ISBN: 951-22-5017-9, 2000.

Everitt. B. S. 1984. An introduction to latent variable models. Chapman and Hall.

Ghahramani Z. and Hinton G. E. 1996. The EM algorithm for mixtures offactor analyzers. Technical Report CRG-TR-96-1 , University of Toronto, Dept. ofComp. Science, 1996.

Ghahramani Z. and Hinton, G. E. 1998. Hierachical non-linear factor analysis and topographic maps. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, Advances in Neural Information Processing Systems 10, pages 486- 492. MIT Press, 1998.

Gray A. and Moore A. W. 2001. 'N-body' problems in statistical learning. In T. K.Leen, T. Dietterich, and V. Tresp, editors, Advances in Neural Information Processing Systems 13, 2001. In this proceeding.

Hinton G. E. and Sejnowski T. J. 1999. Introduction. In G. E. Hinton and T. J. Sejnowski, editors, Unsupervised Learning: Foundations of Neural Computation, pages VII- XVI. The MIT Press, Cambridge, MA, London, England.

Hinton G. E. and. Sejnowski T. J. 1986. Learning and relearning in Boltzmann machines. In Parallel Distributed Processing, volume 1, pages 282- 317. MIT Press, 1986.

Hochreiter S. and Mozer M. C. 2000. An electric field approach to independent component analysis. In P. Pajunen and J. Karhunen, editors, Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland, pages 45-50. Otamedia, Finland, ISBN: 951-22-5017-9.

Huffman. D. 1952. A method for the construction of minimum-redundancy codes. Proceedings of the I.R.E., pp. 1098-1102.

Hyviirinen A. 1999. Survey on independent component analysis. Neural Computing Surveys, 2:94-128.

Manavalan, M. (2014). Fast Model-based Protein Homology Discovery without Alignment. *Asia Pacific Journal of Energy and Environment*, *1*(2), 169-184. https://doi.org/10.18034/apjee.v1i2.580

Manavalan, M. (2016). Biclustering of Omics Data using Rectified Factor Networks. *International Journal of Reciprocal Symmetry and Physical Sciences*, *3*, 1–10. Retrieved from https://upright.pub/index.php/ijrsps/article/view/40

Manavalan, M., & Bynagari, N. B. (2015). A Single Long Short-Term Memory Network can Predict Rainfall-Runoff at Multiple Timescales. *International Journal of Reciprocal Symmetry and Physical Sciences*, *2*, 1–7. Retrieved from https://upright.pub/index.php/ijrsps/article/view/39

Manavalan, M., & Bynagari, N. B. (2015). A Single Long Short-Term Memory Network can Predict Rainfall-Runoff at Multiple Timescales. *International Journal of Reciprocal Symmetry and Physical Sciences*, *2*, 1–7. Retrieved from https://upright.pub/index.php/ijrsps/article/view/39

Manavalan, M., & Ganapathy, A. (2014). Reinforcement Learning in Robotics. *Engineering International*, *2*(2), 113-124. https://doi.org/10.18034/ei.v2i2.572

Marques G. C. and Almeida L. B. 1999. Separation of nonlinear mixtures using pattern repulsion. In J.-F. Cardoso, C. Jutten, and P. Loubaton, editors, Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation, Aussois, France, pages 277- 282.

Noroozi M. and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. CoRR.

Principe J. C. and Xu D. Information-theoretic learning using Renyi's quadratic entropy. In J.-F. Cardoso, C. Jutten, and P. Loubaton, editors, Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation, Aussois, France, pages 407-412.

Shannon C. E. 1948. A mathematical theory of communication. Bell Syst. Tech. J., vol 27, pp. 379-423, Jul. 1948.

Shannon, C. E. 1951. Prediction and entropy of printed english. Bell Syst. Tech. J., vol 30, pp. 35-4, Jan. 1951, 1951.

Zhao Y. and Atkeson C. G. 1996. Implementing projection pursuit learning. IEEE Transactions on Neural Networks, 7(2):362- 373.

--0--