EVALUATION OF THE GOODNESS-OF-FIT TEST USING STEIN DISCREPANCY

Jiani Xie under the supervision of Susan Wei

School of Mathematics and Statistics, The University of Melbourne

Introduction

Real world data distributions are often quite complicated and in high dimension. When we train models fitting these data samples, it is common that we cannot efficiently compute the normalising constants for our model densities. Now we examine a recent method for evaluating and training unnormalised density models by maximising the Learned Stein Discrepancy(LSD).
Assume given a finite set of samples $\{x_i\}_{i=1}^n$ which come from distribution $p(x)$ and an unnormalised density model $q(x)$, we seek a measure of model fitting the data. In the Goodness-of-Fit testing, we decide between two hypotheses: $H_0 : p = q, H_1 : p \neq q$.
In this project, the performance of Goodness-of-Fit test using Stein Discrepancy is evaluated by looking at how the rejection rate varies with the $\mathcal{L}_2$ penalty parameter λ and RBM perturbation rate. We also try to explore whether the Stein's Identity is violated when the test is carried out.

Stein Discrepancy Basics

First, we introduce a measure called Stein's Identity:

$$\mathbb{E}_{p(x)}[\ \nabla_x \log p(x)^T f(x) + Tr(\nabla_x f(x))]\ = 0$$

where $f : \mathcal{R}_D \to \mathcal{R}_D$ is any function such that $lim_{\|x\|\to\infty} p(x)f(x) = 0$. Such f is referred to as critic. [1]
We propose to parameterise the critic over a bounded space of functions $\mathcal{F}$ with a neural network f_ϕ and optimise its parameters to maximise the quantitative measure known as the Learned Stein Discrepancy

$$LSD(f_\phi, p, q) = \mathbb{E}_{p(x)}[\ \nabla_x \log q(x)^T f_\phi(x) + Tr(\nabla_x f_\phi(x))]$$

. [2] Noticing that the expectation is zero if and only if $p = q$.
In our method, we choose to optimise critic networks within the space of functions $\mathcal{F} = \{f : \mathbb{E}_{p(x)}[f(x)^T f(x)] < \infty\}$, whose squared norm has $finite$ expectation under the data distribution. This constraint is enforced by adding an $\mathcal{L}_2$ regulariser $\mathcal{R}_\lambda = \lambda\mathbb{E}_{p(x)}[f(x)^T f(x)]$ and thus the objective function for training the critic becomes $LSD(f_\phi, p, q) - \mathcal{R}_\lambda(f_\phi)$.

Goodness-of-Fit Testing

The two hypotheses for traditional GoF tests are: $H_0 : p = q, H_1 : p \neq q$.
This two-sample test is hard to be carried out since sampling from the unnormalised model $q(x)$ requires cumbersome work. Secondly, an ideal hypothesis test will reject the null whenever $p \neq q$, even if p and q are quite similar, which becomes challenging when dealing with complicated, high-dimensional data.
By applying LSD, a new simple one-sample location test can be designed for solving Goodness-of-Fit problems.
Assuming we are given data $\mathbf{x} = \{x_i\}_{i=1}^n$ and split $\mathbf{x}$ into training, valuation and testing subsets, represented by $\mathbf{x}_{train}, \mathbf{x}_{val}, \mathbf{x}_{test}$. Our "best" critic f_ϕ is obtained by training its parameter on $\mathbf{x}_{train}$ through some neural network, and $\mathbf{x}_{val}$ is for the model selection procedure. As will soon be introduced, the data subset $\mathbf{x}_{test}$ is for calculating the test statistics for our Stein Discrepancy hypotheses test.

Next, the hypotheses are defined as

$$H_0 : \mathbb{E}_{p(x)}[s_f^q(x)] = 0 \qquad H_1 : \mathbb{E}_{p(x)}[s_f^q(x)] \neq 0$$

where $s_f^q(x) = \nabla_x \log q(x)^T f(x) + Tr(\nabla_x f(x))$. [5]
Then we can carry out the hypothesis test as we generally do: compute the statistic $t = \sqrt{n}\frac{\mu_s}{\sigma_s}$ where μ_s and σ_s are the sample mean and standard deviation of $s_f(x)$ evaluated over the subset $\mathbf{x}_{test}$. For sufficiently large n, the test statistic $t \sim N(0, 1)$ under H_0, thus we should reject the null if $t < \Phi(1 - \alpha)$ where α is the given confidence level and Φ is the inverse CDF of $N(0, 1)$.

Potential Problem

Under the class of functions $\mathcal{F}$ mentioned above, the optimal critic takes the following form [4]

$$f_\phi(x) = \frac{1}{2\lambda}(\nabla_x \log q(x) - \nabla_x \log p(x)). \qquad (*)$$

Since the penalty parameter λ is involved in the critic, its value potentially has influence on the fitted model and the performance of the Goodness-of-Fit testing.
We also suspect whether the Stein's Identity is actually enforced during the testing. Stein's Identity is one important assumption that we construct our algorithms on, so if it is violated covertly, we would question the accuracy and reliability of the one-sample testing method and further investigations and experiments are needed.

Experiments

We examine the Goodness-of-Fit testing in its ability to determine whether or not a set of samples was drawn from a given Gaussian-Bernoulli Restricted Boltzmann Machine[3]. The Gaussian-Bernoulli RBM is an unnormalised latent-variable model whose density can be expressed as

$$p(x, h) = \frac{1}{Z}\ exp\ (\frac{1}{2}x^T Bh + b^T x + c^T h - \frac{1}{2}\|x\|^2)$$

with parameters B, b, c, visible dimension x, and latent dimension h.
Note that $\nabla_x \log p(x) = b - x + B \cdot tanh(B^T x + c)$ is easy to calculate.
The distribution p in our experiment is a RBM with parameters designed and controlled by us, from which we randomly sample and obtain $n = 1000$ data points $\{x_i\}_{i=1}^n$. We then perturb the weights of the model with Gaussian noise of standard deviation (also called the 'perturbation rate') in $[0, 0.02, 0.04, 0.06]$. Our test needs to determine if the samples were drawn from this model $q(x)$. We perform the test with three RBMs of different dimensions $(x, h) \in \{(50, 40), (100, 80), (200, 100)\}$ in order to gain a wider vision of statistical results.
To find out how the values of penalty parameter λ effect the significance level of our test, we vary $\lambda \in \{0.1, 1, 10\}$ when running the experiments. In total, there are $4 \times 3 \times 3 = 36$ combinations of parameter configuration.
Concurrently, we check whether the Stein's Identity is satisfied by constructing a t-test having the Stein's Identity as the null hypothesis. We substitute the best critic f_ϕ, each trained through every $n_iter = 100$ iterations, into the left-hand side of the Stein's Identity and obtain another test statistics $H_\phi = \mathbb{E}_{p(x)}[\ \nabla_x \log p(x)^T f_\phi + Tr(\nabla_x f_\phi)]$.
In both tests, the number of rejection is recorded and thus the rejection rate can be calculated.

Results and Analysis

Goodness-of-Fit testing results. Test confidence level $\alpha = 0.10$

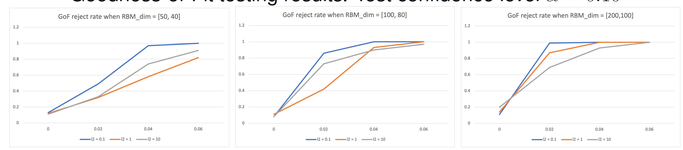

Figure: Perturbation magnitude on the x-axis, rejection rate on the y-axis. Number of datapoints for each experiment is $n = 1000$. Ideal behaviour is a 10% rejection when perturbation is 0 and close to 100% rejection otherwise.

From $(*)$, we know when λ is very large, f can be too small to be able to demonstrate the difference between distributions p and q through Stein Discrepancy. The test would lose its power since the null hypothesis will then be falsely accepted more than it should, even when the model does not fit the data well.
The plots above coincide with our conjecture. Compared the situation with small $\lambda = 0.1$, the rejection rates with a larger $\lambda = 10$ are lower under various perturbation rates and RBM dimensions.
As for the validation of the Identity, the rejection rates collected are generally in between $[0.05, 0.11]$. As we have set the significance level $\alpha = 0.1$, it is quite reasonable to acknowledge that, for most of the time our null hypothesis is well accepted. Thus, given that the base principle is not violated, we can be more confident in our experiment outcomes.

Acknowledgements

I would like to express my gratitude towards my supervisor Susan Wei for introducing me to the interesting topic and guiding me with great support throughout this project. My goals cannot be achieved without the help of my supervisor who has been very kind to share her enlightening ingenious ideas.
Also say a special thank you to the School of Mathematics and Statistics at the University of Melbourne for offering me this invaluable opportunity that shows me a brilliant insight into the world of mathematical research.
This research was supported by Spartan, an extraordinary High Performance Computing system operated by the University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

[1] C.Stein. *A bound for the error in the normal approxi- mation to the distribution of a sum of dependent random variables. In Proceedings of the Sixth Berkeley Sympo- sium on Mathematical Statistics and Prob.* The Regents of the University of California, 1972.

[2] J. Gorham and L. Mackey. *In Proceedings of the 34th International Confer- ence on Machine Learning.* Vol. 70. Tanzanian Mathematical Society, 2017, pp. 1292–1301.

[3] K.H.Cho, T.Raiko, and A.Ilin. "Gaussian-bernoulli deep boltzmann machine". In: *The 2013 International Joint Conference on Neural Networks (IJCNN)* (2013), pp. 1–7.

[4] T.Hu et al. "Stein neural sampler". In: *arXiv preprint arXiv:1810.03545* (Jan. 2018).

[5] W.Grathwohl et al. "Learning the Stein Discrepancy for Training and Evaluating Energy-Based Models without Sampling". In: *International Conference on Machine Learning* (Feb. 2020).