Baltzer Journals July 2, 1995

Measurements of Generalisation Based on Information Geometry

Huaiyu Zhu and Richard Rohwer
Neural Computing Research Group
Department of Computer Science and Applied Mathematics,
Aston University, Birmingham B4 7ET, UK
E-mail: zhuh@aston.ac.uk

Neural networks are statistical models and learning rules are estimators. In this paper a theory for measuring generalisation is developed by combining Bayesian decision theory with information geometry. The performance of an estimator is measured by the information divergence between the true distribution and the estimate, averaged over the Bayesian posterior. This unifies the majority of error measures currently in use. The optimal estimators also reveal some intricate interrelationships among information geometry, Banach spaces and sufficient statistics.

1 Introduction

A neural network (deterministic or stochastic) can be regarded as a parameterised statistical model $P(y|x,w)$, where $x \in X$ is the input, $y \in Y$ is the output and $w \in W$ is the weight. In an environment with an input distribution $P(x)$, it is also equivalent to $P(z|w)$, where $z := [x,y] \in Z := X \times Y$ denotes the combined input and output as data [11]. Learning is the task of inferring w from z. It is a typical statistical inference problem in which a neural network model acts as a "likelihood function", a learning rule as an "estimator", the trained network as an "estimate" and the data set as a "sample". The set of probability measures on sample space Z forms a (possibly infinite dimensional) differential manifold $\mathcal{P}$ [2, 16]. A statistical model forms a finite-dimensional submanifold $\mathcal{Q}$, composed of representable distributions, parameterised by weights w acting as coordinates.

To infer w from z requires additional information about w. In a Bayesian framework such auxiliary information is represented by a prior $P(p)$, where p is the true but unknown distribution from which z is drawn. This is then combined with the likelihood function $P(z|p)$ to yield the posterior distribution $P(p|z)$ via the Bayes formula $P(p|z) = P(z|p)P(p)/P(z)$.

An estimator $\tau : Z \to \mathcal{Q}$ must, for each z, fix one $q \in \mathcal{Q}$ which in a sense approximate p. [1] This requires a measure of "divergence" $D(p,q)$ between $p,q \in \mathcal{P}$ defined independent of parameterisation. General studies on divergences between probability distributions are provided by the theory of information geometry (See [2, 3, 7] and further references therein). The main thesis of this paper is that generalisation error should be measured by the posterior expectation of the information divergence between true distribution and estimate. We shall show that this retains most of the mathematical simplicity of mean squared error theory while being generally applicable to any statistical inference problems.

2 Measurements of Generalisation

The most natural "information divergence" between two distribution $p,q \in \mathcal{P}$ is the δ-divergence defined as [2] [2]

$$D_\delta(p,q) := \frac{1}{\delta(1-\delta)}\left(1 - \int p^\delta q^{1-\delta}\right), \qquad \forall \delta \in (0,1). \tag{1}$$

The limits as δ tends to 0 and 1 are taken as definitions of D_0 and D_1, respectively. Following are some salient properties of the δ-divergences [2]:

$$D_\delta(p,q) = D_{1-\delta}(q,p) \geq 0. \qquad D_\delta(p,q) = 0 \iff p = q. \tag{2}$$

$$D_0(q,p) = D_1(p,q) = K(p,q) := \int p \log \frac{p}{q}. \tag{3}$$

$$D_{1/2}(p,q) = D_{1/2}(q,p) = 2\int \left(\sqrt{p} - \sqrt{q}\right)^2. \tag{4}$$

$$D_\delta(p, p+\Delta p) \approx \frac{1}{2}\int \frac{(\Delta p)^2}{p} \approx \frac{1}{2}\left\langle (\Delta \log p)^2 \right\rangle. \tag{5}$$

The quantity $K(p,q)$ is the Kullback-Leibler divergence (cross entropy). The quantity $D_{1/2}(p,q)$ is the Hellinger distance. The quantity $\int (\Delta p)^2/p$ is usually called the χ^2 distance between two nearby distributions.

Armed with the δ-divergence, we now define the generalisation error

$$E_\delta(\tau) := \int_p P(p) \int_z P(z|p) D_\delta(p, \tau(z)), \qquad E_\delta(q|z) := \int_p P(p|z) D_\delta(p,q), \tag{6}$$

where p is the true distribution, τ is the learning rule, z is the data, and $q = \tau(z)$ is the estimate. A learning rule τ is called δ-optimal if it minimises $E_\delta(\tau)$. A

[1] Some Bayesian methods give the entire posterior $P(p|z)$ instead of a point estimate q as the answer. They will be shown later to be a special case of the current framework.

[2] This is essentially Amari's α-divergence, where $\alpha \in [-1,1]$, re-parameterised by $\delta = (1-\alpha)/2 \in [0,1]$ for technical convenience, following [6].

probability distribution q is called a δ-optimal estimate, or simply a δ-estimate, from data z, if it minimises $E_\delta(q|z)$. The following theorem is a special case of a standard result from Bayesian decision theory.

Theorem .1 (Coherence)
A learning rule τ is δ-optimal if and only if for any data z, excluding a set of zero probability, the result of training $q = \tau(z)$ is a δ-estimate.

Definition .2 (δ-coordinate)
Let $\mu := 1/\delta$, $\nu := 1/(1-\delta)$. Let L_μ be the Banach space of μth power integrable functions. Then L_μ and L_ν are dual to each other as Banach spaces. Let $p \in \mathcal{P}$. Its δ-coordinate is defined as $\overset{\delta}{l}(p) := p^\delta/\delta \in L_\mu$ for $\delta > 0$, and $\emptyset 0 l(p) := \log p$ [2]. Denote by $\overset{1/\delta}{l}$ the inverse of $\overset{\delta}{l}$.

Theorem .3 (δ-estimator in $\mathcal{P}$)
The δ-estimate $\widehat{q} \in \mathcal{P}$ is uniquely given [14] by $\widehat{q} \sim \overset{1/\delta}{l}(\int P(p|z)\overset{\delta}{l}(p))$.

3 Divergence between Finite Positive Measures

One of the most useful properties of the least mean square estimate is the so called $MSE = VAR + BIAS^2$ relation, which also implies that, for a given linear space W, the LMS estimate of w within W is given by the projection of the posterior mean $\widehat{w}$ onto W. This is generalised to the following theorem [16], applying the generalised Pythagorean Theorem for δ-divergences [2].

Theorem .4 (Error decomposition in $\mathcal{Q}$)
Let $\mathcal{Q}$ be a δ-flat manifold. Let $P(p)$ be a prior on $\mathcal{Q}$. Then $\forall q \in \mathcal{Q}$, $\forall z \in Z$,

$$E_\delta(q|z) = E_\delta(\widehat{p}|z) + D_\delta(\widehat{p}, q), \tag{7}$$

where $\widehat{p}$ is the δ-estimate in $\mathcal{Q}$.

To apply this theorem it is necessary to extend the definition of δ-divergence to $\widetilde{\mathcal{P}}$, the space of finite positive measures, which is δ-flat for any δ for a finite sample space Z [2], following suggestions in [2].

Definition .5 ***(δ-divergence on $\widetilde{\mathcal{P}}$)***
The δ-divergence on $\widetilde{\mathcal{P}}$ is defined by

$$D_\delta(p,q) := \frac{1}{\delta(1-\delta)} \int (\delta p + (1-\delta)q - p^\delta q^{1-\delta}) \tag{8}$$

This definition retains most of the important properties of δ-divergence on $\mathcal{P}$, and reduces to the original definition when restricted to $\mathcal{P}$. It has the additional advantage of being the integral of a positive measure, making it possible to attribute the divergence between two measures to their divergence over various events [16]. In particular, the generalised cross entropy is [16]

$$K(p,q) := \int \left(q - p + p \log \frac{p}{q} \right). \tag{9}$$

The δ-divergence defines a differential structure on $\widetilde{\mathcal{P}}$. The Riemannian geometry and the δ-affine connections can be obtained by the Eguchi relations [2, 7] The most important advantage of this definition is that the following important theorem is true and can be proved by pure algebraic manipulation [16].

Theorem .6 ***(Error Decomposition on $\widetilde{\mathcal{P}}$)***
Let $P(p)$ be a distribution over $\widetilde{\mathcal{P}}$. Let $q \in \widetilde{\mathcal{P}}$. Then

$$\langle D_\delta(p,q) \rangle = \langle D_\delta(p,\widehat{p}) \rangle + D_\delta(\widehat{p},q), \tag{10}$$

where $\widehat{p}$ is the δ-average of p given by $\widehat{p}^\delta := \langle p^\delta \rangle$.

Theorem .7 ***(δ-estimator in $\widetilde{\mathcal{P}}$)***
The δ-estimate $\widehat{p} = \tau_\delta(z)$ in $\widetilde{\mathcal{P}}$ is given by $\widehat{p}^\delta = \langle p^\delta \rangle_z$. In particular, the 1-estimate is the posterior marginal distribution $\widehat{p} = \langle p \rangle_z$.

Theorem .8 ***(δ-estimator in $\mathcal{Q}$)***
Let $\mathcal{Q}$ be an arbitrary submanifold of $\widetilde{\mathcal{P}}$. The δ-estimate $\widehat{q}$ in $\mathcal{Q}$ is given by the δ-projection of $\widehat{p}$ onto $\mathcal{Q}$, where $\widehat{p}$ is the δ-estimate in $\widetilde{\mathcal{P}}$.

4 Examples and Applications to Neural Networks

Explicit formulas are derived for the optimal estimators for the multinomial [15] and normal distributions [14].

Example 1
Let $m \in \mathbb{N}^n$, $p \in \mathcal{P} = \Delta^{n-1}$, $a \in \mathbb{R}_+^n$. Consider multinomial family of distributions $M(m|p)$ with a Dirichlet prior $D(p|a)$. The posterior is also a Dirichlet distribution $D(p|a+m)$. The δ-estimate $\widehat{p} \in \widetilde{\mathcal{P}}$ is given by $(\widehat{p}_i)^\delta = (a_i + m_i)_\delta / (|a+m|)_\delta$, where $|a| := \sum_i a_i$ and $(a)_b := \Gamma(a+b)/\Gamma(a)$. In particular, $\widehat{p}_i = (a_i + m_i)/|a+m|$ for $\delta = 1$, and $\widehat{p}_i = \exp(\Psi(a_i + m_i) - \Psi(|a+m|))$ for $\delta = 0$, where Ψ is the the digamma function. The δ-estimate $\widehat{q} \in \mathcal{P}$ is given by normalising $\widehat{p}$.

Example 2
Let $z, \mu \in \mathbb{R}$, $h \in \mathbb{R}_+$, $a \in \mathbb{R}$, $n \in \mathbb{R}_+$. Consider the Gaussian family of distributions $f(z|\mu) = N(z-\mu|h)$, with fixed variance $\sigma^2 = 1/h$. Let the prior be another Gaussian $f(\mu) = N(\mu - a|nh)$, Then the posterior after seeing a sample z of size k, is also a Gaussian $f(\mu|z) = N(\mu - a_k|n_k h)$, where $n_k = n+k$, $a_k = (na + \sum z)/n_k$, which is also the posterior least squares estimate. The δ-estimate $\widehat{q} \in \mathcal{P}$ is given by the density $f(z'|\widehat{q}) = N\left(z' - a_k \middle| h/(1+\delta/n_k)\right)$.

The entities $|a|$ for the multinomial model and n for the Gaussian model are effective previous sample sizes, a fact known since Fisher's time. In a restricted model, the sample size might not be well reflected, and some ancillary statistics may be used for information recovery [2].

Example 3
In some Bayesian methods, such as the Monte Carlo method [10], no estimator is explictly given. Instead, the posterior is directly used for sampling p. This produces a prediction distribution on test data which is the posterior marginal distribution. Therefore these methods are implicitly 1-estimators.

Example 4
Multilayer neural networks are usually not δ-convex for any δ, and there may exist local optima of $E_\delta(\cdot|z)$ on $\mathcal{Q}$. A practical learning rule is usually a gradient descent rule which moves w in the direction which reduces $E_\delta(q|z)$. The 1-divergence can be minimised by a supervised learning rule, the Boltzmann machine learning rule [1]. The 0-divergence can be minimised by a reinforcement learning rule, the simulated annealing reinforcement learning rule for stochastic networks[13].

$$\min_q K(p,q) \iff \Delta w \sim \langle \partial_w \o 0 l(q) \rangle_p - \langle \partial_w \o 0 l(q) \rangle_q \tag{11}$$

$$\min_q K(q,p) \iff \Delta w \sim \langle \partial_w \o 0 l(q),\ \o 0 l(p) - \o 0 l(q) \rangle_q \tag{12}$$

5 Conclusions

The problem of finding a measurement of generalisation is solved in the framework of Baysian decision theory, with machinery developed in the theory of information geometry.

By working in the Bayesian framework, this ensures that the measurement is internally coherent, in the sense that a learning rule is optimal if and only if it produces optimal estiamtes for almost all the data. By adopting an information geometric measurement of divergence between distributions, this ensures that the theory is independent of parameterisation. This resolves the controversy in [8, 12, 9].

To guarantee a unique and well-defined solution to the learning problem, it is necessary to generalise the concept of information divergence to the space of finite positive measures. This development reveals certain elegant relations between information geometry and the theory of Banach spaces, showing that the dually-affine geometries of statistical manifolds are in fact intricately related to the dual linear geometries of Banach spaces.

In a computational model, such as a classical statisitical model or a neural network, the optimal estimator is the projection of the ideal estimator to the model. This theory generalises the theory of linear Gaussian regression to general statistical estimation and function approximation problems. Further research may lead to Kalman filter type learning rules which are not restricted to linear and Gaussian models.

Acknowledgement

We are grateful to Prof. S. Amari for clarifying a point of information geometry. We would like to thank many people in the Neural Computing Research Group, especially C. Williams, for useful comments and practical help. This work was partially supported by EPSRC grant (GR/J17814).

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cog. Sci.*, 9:147–169, 1985.

[2] S. Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Springer Lecture Notes in Statistics*. Springer-Verlag, New York, 1985.

[3] S. Amari. Differential geometrical theory of statistics. In Amari et al. [4], chapter 2, pages 19–94.

[4] S. Amari, O. E. Barndoff-Nieldon, R. E. Kass, S. L. Lauritzen, and C. R. Rao, editors. *Differential Geometry in Statistical Inference*, volume 10 of *IMS Lecture Notes Monograph*.

IMS, Hayward, CA, 1987.

[5] S. J. Hanson, J. D. Cowan, and C. L. Giles, editors. *Advances in Neural Information Processing Systems*, volume 5, San Mateo, CA, 1993. Morgan Kaufmann.

[6] R. E. Kass. Canonical parameterization and zero parameter effects curvature. *J. Roy. Stat. Soc., B*, 46:86–92, 1984.

[7] S. L. Lauritzen. Statistical manifolds. In Amari et al. [4], chapter 4, pages 163–216.

[8] D. J. C. MacKay. *Bayesian Methods for Adaptive Models.* PhD thesis, California Institute of Technology, Pasadena, CA, 1992.

[9] D. J. C. MacKay. Hyperparameters: Optimise, or integrate out? Technical report, Cambridge, 1993.

[10] R. M. Neal. Bayesian learning via stochastic dynamics. In Hanson et al. [5], pages 475–482.

[11] H. White. Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1(4):425–464, 1989.

[12] D. H. Wolpert. On the use of evidence in neural neworks. In Hanson et al. [5], pages 539–546.

[13] H. Zhu. *Neural Networks and Adaptive Computers: Theory and Methods of Stochastic Adaptive Computations.* PhD thesis, Dept. of Stat. & Comp. Math., Liverpool University, 1993. `ftp://archive.cis.ohio-state.edu/pub/neuroprose/Thesis/zhu.thesis.ps.Z`.

[14] H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation for continuous distributions. Technical Report NCRG/4352, Dept. Comp. Sci. & Appl. Math., Aston University, Aug. 1995. `ftp://cs.aston.ac.uk/neural/zhuh/continuous.ps.Z`.

[15] H. Zhu and R. Rohwer. Bayesian invariant measurements of generalisation for discrete distributions. Technical Report NCRG/4351, Dept. Comp. Sci. & Appl. Math., Aston University, Aug. 1995. `ftp://cs.aston.ac.uk/neural/zhuh/discrete.ps.Z`.

[16] H. Zhu and R. Rohwer. Information geometric measurements of generalisation. Technical Report NCRG/4350, Dept. Comp. Sci. & Appl. Math., Aston University, Aug. 1995. `ftp://cs.aston.ac.uk/neural/zhuh/generalisation.ps.Z`.