Confidence in prediction by neural networks

Liat Ein-Dor and Ido Kanter
Minerva Center and Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel
(Received 17 August 1998)

The idea that a trained network can assign a confidence number to its prediction, indicating the level of its reliability, is addressed and exemplified by an analytical examination of a perceptron with discrete and continuous output units. Results are derived for both Gibbs and Bayes scenarios. The information gain by the confidence number is estimated by various entropy measurements. [S1063-651X(99)06606-4]

PACS number(s): 87.10.+e, 84.35.+i, 02.50.−r, 05.20.−y

Statistical physics methods have contributed greatly to the theory of learning in recent years [1–3]. Analytical methods were developed to investigate the learning of a rule from randomized data by large neural networks. The quality of the learning is measured by the averaged generalization error that quantifies the average amount of disagreement between the *student*, the trained network, and known rules.

There are basically two lines of approach in the investigation of the task of learning a rule from random examples. In the first approach, known as a batch learning, the examples are stored and can be provided at any given moment of the learning process. For a given training set, the learning trail gains from the quenched fluctuations in the examples provided, and therefore the analytical treatment is based on the replica method [2]. The second line of research concerns the physics of so-called on-line learning processes [4] and was initiated in Refs. [5,6]. From a practical point of view, on-line learning is particularly attractive since it uses only the latest example from the training set. This obviously reduces the storage needs in comparison with memory based batch prescriptions. Furthermore, this property makes it possible to investigate analytically a variety of on-line learning scenarios, where the learning dynamics is described exactly in terms of coupled differential equations [7].

In both learning scenarios the major analytical activity concentrates on the study of teacher and student networks with the same architecture and with continuous adjustable weights; the size of the input is N and the size of the training set (number of random examples) is defined by αN; see [2] and references therein. The generalization error, the average amount of disagreement between the teacher and student predictions on a new example, was found to scale *asymptotically* with $1/\alpha$ for binary output units and to scale as $e^{-\alpha}$ for networks with continuous output units [1,2].

The traditional question in the learning theory is to find the learning prescription which minimizes *asymptotically* the generalization error or maximizes the similarity between the student and the teacher in the case of a realizable learning rule. In this paper we address and examine the following orthogonal question. A network is trained by a given learning algorithm on a set of random examples, the training set. A new question (a new input) is then presented, where it is clear that a deterministic student gives a well-defined answer (an output). A practical question one may now ask is, to what extent can we rely on the student's answer? More precisely, a confidence number must be assigned to each answer to indicate the level of its reliability. In general, the confidence number can take any value in a definite range. However, it is more convenient to give a probabilistic interpretation to the confidence number, and hence, the natural range is [0,1]. In such a case the reliability of the answer is represented by the confidence number, a probability between zero and 1. The average confidence number is nothing else but $1-\epsilon_g$, where ϵ_g is the average generalization error.

In this paper we would like first to address and then to examine analytically on some limited architectures, the following questions: (i) Can one assign for each input a different confidence number, such that in some of the questions the confidence number is greater or smaller than the the average confidence number, $1-\epsilon_g$? (ii) What are the parameters of the question (input) which cause the confidence number to be above or below the averaged one? (iii) What is the quantitative interplay between these parameters and the resulting confidence number? (iv) How does one measure qualitatively the information gain by assigning a different confidence number for each input? And does the information gain depend on the architecture of the network or the prescription of the learning? (v) Can the dependence of the confidence number on the ''quality'' of the input be extended to the case of continuous output units? A similar idea was previously addressed and examined to improve the learning process. The generalization error is reduced by rejecting examples that lie within a given neighborhood to the decision boundary, namely, examples with low reliability [8].

In order to simplify the discussion, the above-mentioned questions are addressed and examined within the framework of a realizable learning rule, where both teacher and student have the same prototypical architecture, a perceptron with a binary output unit. The N input units are $\{S_i\}$ $i=1, 2, \ldots, N$, and the weights of teacher and student are defined respectively by $\{W_i\}$ and $\{J_i\}$. The output of the teacher, for instance, is given by

$$\text{out}=\text{sgn}\left(\sum_{i=1}^{N} W_i S_i\right). \tag{1}$$

This architecture is extended later to a perceptron with a continuous output unit, $\text{out}=\hat{O}(\Sigma_{i=1}^{N} W_i S_i)$, where for simplicity of discussion we choose the common activation function $\hat{O}=\tanh$.

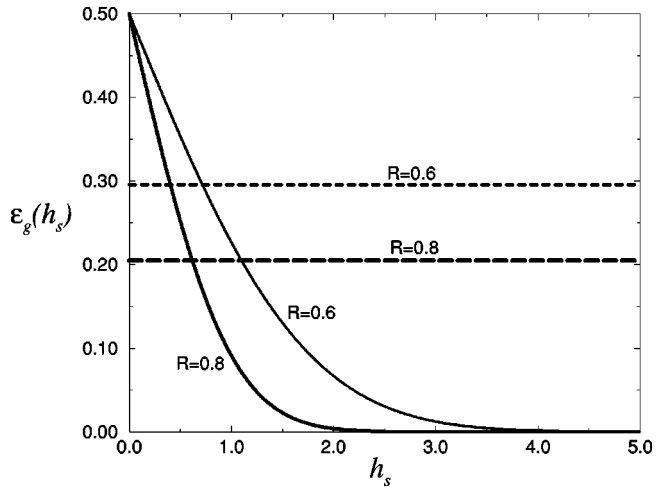

FIG. 1. $\epsilon_g(h_s)$ vs the positive field of the student h_s, Eq. (3), for $R=0.6$ and 0.8. The horizontal lines are the values of the averaged generalization error ϵ_g, Eq. (2).

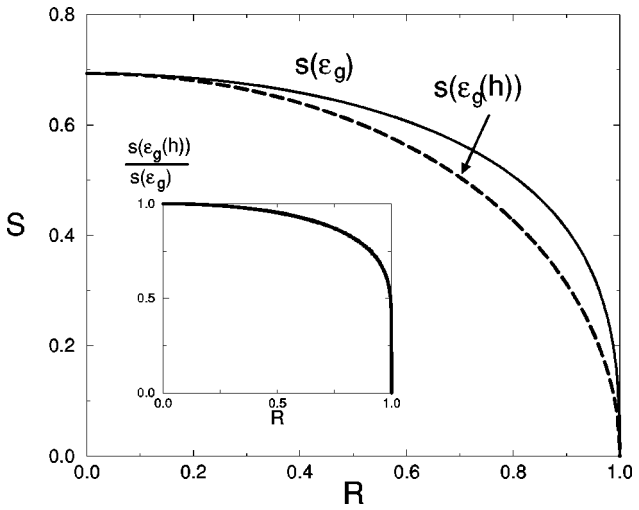

FIG. 2. The entropy derived from Eq. (6), $S(\epsilon_g)$, vs R (solid line) and the entropy derived from the distribution of the generalization error as a function of the local fields of the student Eq. (5), $S[\epsilon_g(h)]$, vs R (dashed line). At $R=0$ both entropies are equal to $\ln 2$. Inset: the ratio $S[\epsilon_g(h)]/S(\epsilon_g)$ vs R.

For a perceptron with a binary output unit and random inputs, Eq. (1), the generalization error of the student depends only on $R=W\cdot J/\|W\|\|J\|$ [2], and is given explicitly by

$$\epsilon_g=\frac{\cos^{-1}(R)}{\pi}. \tag{2}$$

In a similar spirit, one can define the *average* generalization error for an input that induces a local field $h_s=\Sigma_i J_i S_i$ on the output of the student. Note that the average is over all possible teachers obeying an overlap R with a given student. In such a case and where the output of the student is $+1$, one can find explicitly [2,8]

$$\epsilon_g(h_s)=\frac{1}{2}\mathrm{erfc}[Rh_S/\sqrt{2(1-R^2)}], \tag{3}$$

where $\mathrm{erfc}(x)\equiv 2/\sqrt{\pi}\int_x^\infty \exp(-x^2)$. Note that the fraction of such inputs with h_s in the case of random inputs is $\exp(-h_s^2/2)/\sqrt{2\pi}$ and for a negative output of the teacher one has to replace $h_s\to|h_s|$ in Eq. (3). A typical result for ϵ_g and $\epsilon_g(h)$ for $h>0$ and $R=0.6,0.8$ is presented in Fig. 1. It is clear that $\epsilon_g(0)=1/2$ is independent of R, since an orthogonal input to the weights of the student, J, does not contain any information regarding the local field of the teacher, which is an unbiased Gaussian. Hence, the sign of the output of the student is uncorrelated with the output of the teacher. Similarly, $\epsilon_g(h)$ is a decreasing function of R (>0), since for a *given input* with h_s and R, the probability that the local field of the teacher is h_t is

$$P(h_t|h_s,R)=\sqrt{\frac{1}{2\pi(1-R^2)}}e^{-(h_t-Rh_s)^2/2(1-R^2)}. \tag{4}$$

The center of the Gaussian, Eq. (4), is at Rh_s, and it increases with h_s and therefore the weight of the negative tail, the error, decreases. Similarly, for a given h_s, the generalization error decreases with R since the standard deviation of the Gaussian, $\sqrt{1-R^2}$, decreases with R.

It is now clear that we can do better than the average confidence number. For a student who develops a similarity $(R>0)$ with the teacher, inputs with large h_s with high probability also have $h_t>0$, and their confidence number is higher than the average one, where for $h_s\to 0$ the confidence number is 1/2. The values of large h_s and the exact confidence numbers as a function of h_s and R are given by Eqs. (2)–(4).

The quantitative measure of the information gain by using a confidence number as a function of h_s, $1-\epsilon_g(h_s)$, can be deduced from a comparison between the entropy of the answers (outputs) of the student which is averaged over an infinite number of random questions (inputs)

$$S[\epsilon_g(h)]=-\int_{-\infty}^{\infty}\sqrt{\frac{1}{2\pi}}dh e^{(-h^2/2)}\{\epsilon_g(h)\ln[\epsilon_g(h)] + [1-\epsilon_g(h)]\ln[1-\epsilon_g(h)]\} \tag{5}$$

and the entropy of the averaged confidence number

$$S(\epsilon_g)=-[\epsilon_g\ln(\epsilon_g)+(1-\epsilon_g)\ln(1-\epsilon_g)], \tag{6}$$

where ϵ_g and $\epsilon_g(h)$ are given as a function of R by Eqs. (2) and (3). Results for $S(\epsilon_g)$ and $S[\epsilon_g(h)]$ as a function of R are presented in Fig. 2, where it is clear that the two curves should coincide at $R=0,1$ where independent of $h_s(\neq 0)$, the confidence number is 1/2,1. Clearly, a lower entropy indicates a better knowledge regarding the teacher's outputs.

Although as $R\to 1$ both $S(\epsilon_g)$, $S[\epsilon_g(h)]\to 0$, see Fig. 2, the information gain can be deduced from the rate of the convergence of the entropies to zero. In the inset of Fig. 2, the ratio between $S[\epsilon_g(h)]/S(\epsilon_g)$ versus R is calculated numerically from Eqs. (5) and (6). It is clear that the information gain increases with R, and asymptotically as $R\to 1$ one can show that

$$S[\epsilon_g(h)]/S(\epsilon_g)=\frac{4C\sqrt{\pi}}{|\ln(1-R)|}=\frac{2C\sqrt{\pi}}{\ln(\alpha)} \tag{7}$$

where $C=-\int_0^\infty [x\ln x+(1-x)\ln(1-x)]dy\sim 0.638$, $x=0.5\,\mathrm{erfc}(y)$ and the last equality is derived from the asymptotic generalization error for the Gibbs case $\epsilon_g \sim 0.62/\alpha$ [2]. The decreasing of $S[\epsilon_g(h)]/S(\epsilon_g)$ with R, is a result of the behavior of $\epsilon_g(h)$, Eq. (5), indicating that as R increases, roughly speaking, the inputs can be split into two classes. In the first class with $h_s > O(\sqrt{1-R^2})$, the generalization error is almost zero ($\ll \epsilon_g$), whereas in the second class $h_s < O(\sqrt{1-R^2})$ the generalization error is close to 1/2. In such a developed bimodal distribution as a function of R, the width of the distribution of the generalization error around the average, ϵ_g, increases and similarly the information gain increases.

The above-mentioned results can apply also to the Bayes algorithm [2], where the main idea is to use the distribution of the splitting of the normalized version space by a new input to y and $1-y$ (for details, see Eq. (24) in Refs. [9] and [2]). The average entropy, $S_{Bayes}(\epsilon_g)$, is given again by Eq. (6), but $\epsilon_g=\cos^{-1}(\sqrt{R})/\pi$ instead of Eq. (2), and $\epsilon_g(y)=y$ for $y\leqslant 1/2$ [2]. Similar to the Gibbs entropies, Eqs. (5) and (6), one can show that

$$S_{Bayes}[\epsilon_g(y)]=-\frac{2}{\gamma}\int_0^{1/2} dy e^{-(t^2/2)(1-\gamma^2)} \times[y\ln y+(1-y)\ln(1-y)],$$

where $\gamma=\sqrt{R/(1-R)}$ and t is determined through the relation $y=0.5\,\mathrm{erfc}(t\gamma/\sqrt{2})$. One can show now that asymptotically $S_{Bayes}(\epsilon_g)=-\sqrt{1-R}\ln(1-R)/2\pi\sim 0.44\ln(\alpha)/\alpha$ and $S_{Bayes}[\epsilon_g(h)]=2C\sqrt{1-R}/\sqrt{\pi}\sim 2\sqrt{\pi}0.44C/\alpha$. Furthermore, asymptotically one can find that $S_{Bayes}[\epsilon_g(h)]/S_{Bayes}(\epsilon_g)=2C\sqrt{\pi}/\ln(\alpha)$ and

$$\frac{S_{Bayes}[\epsilon_g(y)]}{S_{Gibbs}[\epsilon_g(h)]}=\frac{S_{Bayes}(\epsilon_g)}{S_{Gibbs}(\epsilon_g)}=\frac{0.44}{0.62}, \tag{8}$$

which may indicate a universal property.

It is interesting now to extend these concepts, $\epsilon_g(h)$ and $S[\epsilon_g(h)]$, to a perceptron with a continuous output unit that is more realistic in many learning tasks,

$$\mathrm{out}=\hat{O}\left(\sum_{i=1}^{N} W_i S_i\right), \tag{9}$$

where we concentrate on the common choice, $\hat{O}=\tanh$. The averaged generalization error is defined by

$$\epsilon_g=\left\langle\left\langle \frac{1}{4}[\tanh(W\cdot S)-\tanh(J\cdot S)]^2\right\rangle\right\rangle, \tag{10}$$

where $\langle\langle\cdots\rangle\rangle$ stands for the average over the input space. The generalization error can be expressed as a function of two order parameters; R as was defined earlier, and we assume now that $\|W\|=1$ and the additional order parameter $Q=\|J\|$ measuring the length of the weights of the student. Note that although $\epsilon_g\in[0:1]$, it does not have an interpretation of a probability.

The main question now is to find a criterion for an agreement between the student and the teacher. If we adopt a strict measure that the two networks ''agree'' only when their outputs are *exactly* the same, we will find that the probability for such an event is zero for any $R\neq 1$. A more practical definition for an agreement is that the output of the student is *not too far* from that of the teacher. Following Eq. (10), a natural way to define an agreement with an error Δ is

$$\frac{1}{4}[\tanh(W\cdot S)-\tanh(J\cdot S)]^2\leqslant\Delta, \tag{11}$$

where Δ is a given parameter that can be fixed, for instance, by the required resolution of the user. From Eq. (11) one can find that

$$h_s^-<h_t<h_s^+, \tag{12}$$

with $h_s^\pm=\tanh^{-1}[\tanh(h_s)\pm 2\sqrt{\Delta}]$. The generalization error for a given h_s and Δ is then given by

$$1-\epsilon_g(h_s)=\frac{1}{2}[\mathrm{erfc}(H_s^-)-\mathrm{erfc}(H_s^+)], \tag{13}$$

where $H_s^\pm=(h_s^\pm-Rh_s/Q^2)/\sqrt{2(1-R^2/Q^2)}$ and the averaged generalization error is given by $\epsilon_g=1/\sqrt{2\pi}\int dh_s e^{-h_s^2/2}\epsilon_g(h_s)$. Similar to Eqs. (5) and (6), one can define the entropies, $S(\epsilon_g)$ and $S[\epsilon_g(h)]$. Results of these two entropies as a function of $R\epsilon[0:1]$ for $Q=1$ and for some typical values of Δ are presented in Fig. 3. Note that for some values of Δ, the entropy does not monotonically decrease with R, as was found for binary output units, a result which requires an explanation. For learning with binary units, the generalization error is always $\leqslant 1/2$, since in the worst case one can choose a random output (as for $R=0$).

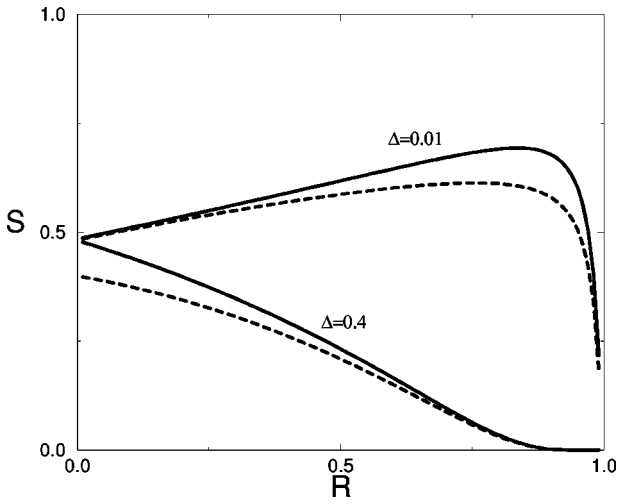

FIG. 3. The entropies for a perceptron with a continuous output unit. $S(\epsilon_g)$ [Eq. (6)], vs R (solid lines) and $S[\epsilon_g(h)]$ [Eq. (5)], vs R (dashed lines).

In the case of continuous units one should distinguish between two different scenarios. In the first scenario, similar to the binary case, for a given Δ and $R=0$ the student knows with a probability $\geqslant 1/2$ the ''true'' answer of the teacher. Hence, as R increases, both the generalization error and the entropy decreases toward zero. In the second scenario, the student knows at $R=0$ with probability $>1/2$ that his answer

is ''wrong,'' which is the case where Δ is small enough. In such a learning process, as R increases, ϵ_g first decreases toward 1/2 and the entropy increases towards ln(2). As R increases beyond that point, both the entropy and ϵ_g decay to zero. A similar scenario holds for $S[\epsilon_g(h)]$ but the value is lower than ln(2).

Results for a perceptron with a continuous output unit can be developed further. Assume that the averaged agreement, $\overline{\Delta}$, between teacher and student is required, for instance, by the user. On some inputs of the student with local field h_s we count the answer to be correct although the difference square [see Eq. (11)] between the teacher and student outputs is large, $\Delta(h_s)>\overline{\Delta}$, where with other inputs we have a more restricted criterion for a correct answer, $\Delta(h_s)<\overline{\Delta}$. This global constraint can be summarized for the case of random inputs by

$$\overline{\Delta}=\int_{-\infty}^{\infty} dh \frac{e^{-h^2/2}}{\sqrt{2\pi}}\Delta(h). \tag{14}$$

It is clear that this global constraint does not uniquely determine the function of $\Delta(h)$. The question raised now is to find *the best criterion for an agreement*. More precisely, the best criterion is the one that minimizes the entropy with respect to all possible distributions of $\Delta(h)$. Hence, one has to minimize Eq. (5) under the global constraint (14). A trivial solution with zero entropy always exists. For $h_s\in[-\infty;h_0]$ $\Delta(h)=0$ such that $\epsilon_g=1$ and for $h_s\in[h_0;\infty]$ $\Delta(h)=[1+\tanh(h)]^2/4$ such that $\epsilon_g=0$. However, since we would like to maximize the agreement between teacher and student, $\Delta(h_s)$ is bounded such that $\epsilon_g[\Delta(h)]\leqslant 1/2$ for all h_s. The minimization of Eq. (5) under constraint (14) was carried out by the Monte Carlo method. In Fig. 4 results for $\Delta(h)$, which minimizes the entropy under the global constraint that $\overline{\Delta}=0.5$, are presented for $h>0$, where $\Delta(-h)=\Delta(h)$. The curve $[1+\tanh(h)]^2/4$ represents the upper bound for $\Delta(h)$, which gives $\epsilon_g(h)=0$. For small h, $\Delta(h)$ is chosen such that ϵ_g is almost 0 and increases with h. For $\Delta(h)=\overline{\Delta}$, the entropies obtained from Eqs. (6) and (5) are $S(\epsilon_g)\sim 0.011$ and $S[\epsilon_g(h)]\sim 0.0098$. These numbers should be compared with the enhancement of the entropies after the optimization of $\Delta(h)$, as presented in Fig. 3. The results are $S[\epsilon(g)]\sim 0.0018$ and $S[\epsilon_g(h)]\sim 0.0017$, which are around five times smaller than the entropies before the optimization. For smaller values of $\overline{\Delta}$, the optimal $\Delta(h)$ decreases with positive h and will be discussed elsewhere.

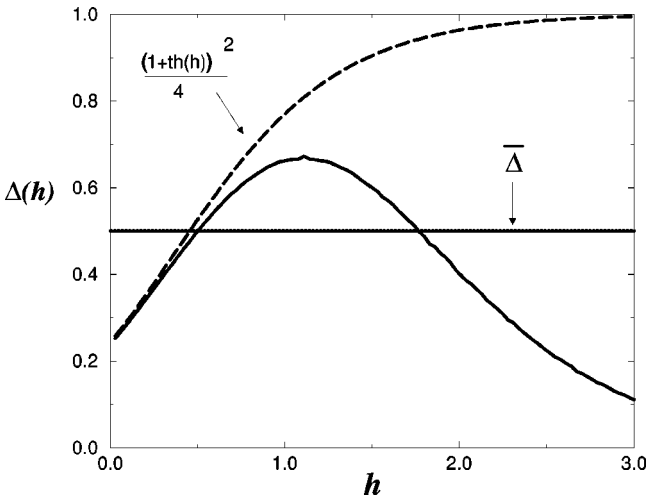

FIG. 4. The optimal $\Delta(h)$ vs h for $\overline{\Delta}=0.5$. The dashed line $[1+\tanh(h)]^2/4$, indicates the upper bound for $\Delta(h)$, where $\epsilon_g(h)=0$.

All of the above-mentioned concepts and results can be generalized for multilayered networks; however, the equations are more involved and will be discussed elsewhere. Nevertheless, we would like to conclude with a result for a committee machine with nonoverlapping receptive fields and with K hidden units. If the overlap between the weights of the teacher and that of the student for each one of the hidden units is equal to R, one can show that $S_{MLN}[\epsilon_g(h)]/S_{MLN}(\{\epsilon_g\})\propto(1-R)^{(K-1)/2}/|\log(1-R)|$, which indicates an enhancement in comparison to the perceptron case.

We thank W. Kinzel and M. Opper for fruitful discussions. I.K. acknowledges the partial support of the Israel Academy of Science.

[1] T.L.H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).
[2] M. Opper and W. Kinzel, in *Models of Neural Networks III*, edited by E. Domany, J.L. van Hemmen, and K. Schulten (Springer, Berlin, 1996).
[3] J.A. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).
[4] S. Amari, IEEE Trans. Electron. Comput. **EC-16**, 299 (1967).
[5] W. Kinzel and P. Ruján, Europhys. Lett. **13**, 473 (1990).
[6] O. Kinouchi and N. Caticha, J. Phys. A **26**, 6243 (1992).
[7] M. Rosen-Zvi, M. Biehl, and I. Kanter, Phys. Rev. E **58**, 3606 (1998).
[8] J.M. Parrondo and C. Van den Broeck, Europhys. Lett. **22**, 319 (1993).
[9] M. Opper and D. Haussler, in *IVth Annual Workshop on Computational Learning Theory*, Sanata Cruz, 1991 (Morgan Kaufmann, San Mateo, CA, 1991), pp. 75–87.