ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 169-174

Bayesian online learning in the perceptron

Ole Winther and Sara A. Solla
CONNECT,The Niels Bohr Institute,Blegdamsvej 17
2100 Copenhagen Ø, Denmark
winther@connect.nbi.dk and solla@research.att.com

Abstract. In a Bayesian approach to online learning a simple approximate parametric form for posterior is updated in each online learning step. Usually in online learning only an estimate of the solution is updated. The Bayesian online approach is applied to two simple learning scenarios, learning a perceptron rule with respectively a spherical and a binary weight prior. In the first case we rederive the results for the optimal Hebb-type online algorithm for spherical input distribution.

1. Introduction

Recently there has been a lot of interest in studying online learning scenarios within the statistical mechanics setting (see e.g. [1]-[5]). One of the main reasons for this is the simplicity of the analysis compared to the analysis of batch learning. Furthermore it turns out that the generalization performance of the online learning algorithms in many cases are not much worse than batch algorithms. Thus one may find a good approximate solution through much less computional effort.

The kind of Bayesian approach to online learning we will use in this paper is to approximate the posterior of weights of the network with a simple parametric form. For each online learning step one may apply Bayes theorem to find the new value of the parameters of the posterior from the current parameters and the new training example. M. Opper [5] has done exactly that approximating the posterior with a Gaussian when learning a rule which is a continuous function of its parameters. In this case the approximate posterior becomes exact when the training set grows large.

However in most cases not all the information contained in the training examples can be contained in the approximate posterior. This is the reason for the inferior generalization performance of the Bayesian online algorithm compared to a pure Bayesian batch algorithm.

The rest of the paper is organized as follows. In section 2. the Bayesian approach to classification is presented. In section 3. we consider a specific scenario: learning a simple perceptron rule when the input data is uncorrelated. We will consider both a spherical and binary parameter prior. We conclude in section 4. and point out some possible extensions to this work.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 169-174

2. Bayesian Classification

To explain the Bayesian approach to classification [7] consider a training set of m input-output pairs $D_m = \{(\mathbf{s}^\mu, \tau^\mu), \mu = 1, \ldots, m\}$ where the input $\mathbf{s}^\mu$ is a N-dimensional vector and the output $\tau^\mu = \pm 1$ is a binary classification label. The examples are assumed to be drawn independently from the same distribution. The probability of the training set given the unknown rule parameterized by $\mathbf{w}$ is $p(D_m|\mathbf{w}) = \prod_\mu [p(\mathbf{s}^\mu)p(\tau^\mu|\mathbf{w}, \mathbf{s}^\mu)]$ where we have assumed that the input is independent of the rule. The total knowledge about the rule after observing m examples is expressed by the *posterior* which is found using Bayes rule

$$p(\mathbf{w}|D_m) = \frac{p(\mathbf{w}) \prod_\mu p(\tau^\mu|\mathbf{w}, \mathbf{s}^\tau)}{Z} , \tag{1}$$

where $p(\mathbf{w})$ is the *prior* over rule parameters and $Z = \int d\mathbf{w} p(\mathbf{w}) \prod_\mu p(\tau^\mu|\mathbf{w}, \mathbf{s})$ is a normalization constant. As will be explained below to make Bayes optimal predictions one need to perform certain averages over the posterior. In general the posterior is a very complicated distribution for which it is not possible to to do the averages analytically and it is very computationally costly to perform them numerically in a high dimensional rule space.

One may also observe from the posterior that the information in all the examples are needed. This is not what we want in *online learning*. In online learning we will disregard the training example once it has been used to update the parameters we keep track of. Usually the parameters represent the an estimate of the rule only. To explain Bayesian online learning we add a new example to get a recursive relation for the posterior

$$p(\mathbf{w}|D_{m+1}) = \frac{p(\mathbf{w}|D_m)p(\tau^{m+1}|\mathbf{w}, \mathbf{s}^{\tau+1})}{\int d\mathbf{w} p(\mathbf{w}|D_m p(\tau^{m+1}|\mathbf{w}, \mathbf{s}^{\tau+1})} \tag{2}$$

This posterior is exact. It still depends on all the whole training set explicitly. However we can approximate $p(\mathbf{w}|D_m)$ with a simpler distribution $p(\mathbf{w}|A_m)$ where A_m is shorthand for the parameters that characterize the distribution, e.g. the first two moments of $\mathbf{w}$ in the Gaussian case. The updated moments A_{m+1} is then obtained from the distribution

$$p(\mathbf{w}|A_m, (\mathbf{s}^{m+1}, \tau^{m+1})) = \frac{p(\mathbf{w}|A_m)p(\tau^{m+1}|\mathbf{w}, \mathbf{s}^{\tau+1})}{\int d\mathbf{w} p(\mathbf{w}|A_m)p(\tau^{m+1}|\mathbf{w}, \mathbf{s}^{\tau+1})} \tag{3}$$

This framework is a straight forward Bayesian extension to usual online learning given a recursive relation for distributions of rules rather than just for a single estimate of rule.

The parameters of the true function which are picked at random with probability given by the prior $p(\mathbf{w})$ should have non-zero a priori probability in the approximate scheme. This suggests that we should choose the approximate posterior $p(\mathbf{w}|A_m)$ such that $p(\mathbf{w}|A_0) = p(\mathbf{w})$ may be fulfilled.

In a forthcoming paper [6] we will show from a purely information theoretical argument how to choose A_{m+1} optimally (as a function of A_m and

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 169-174

$(\mathbf{s}^{m+1}, \tau^{m+1}))$. We can therefore prove that the Bayesian approach makes optimal use of the information contained in A_m and $(\mathbf{s}^{m+1}, \tau^{m+1})$ and thus gives a lower bound for the generalization error of any algorithm using the same amount of information.

Using the posterior distribution $p(\mathbf{w}|A_m)$ (or $p(\mathbf{w}|D_m)$ in the batch case) we can calculate the *predictive probability* [8] of an output label τ given the input $\mathbf{s}$

$$p(\tau|\mathbf{s}) = \int d\mathbf{w} p(\mathbf{w}|A_m) p(\tau|\mathbf{w}, \mathbf{s}) \equiv \langle p(\tau|\mathbf{w}, \mathbf{s}) \rangle \tag{4}$$

The Bayes optimal classification algorithm says that one should choose the label with highest probability $\tau^{\text{Bayes}} = \text{argmax}_\tau\, p(\tau|\mathbf{s})$. For binary ± 1 classification this may be written as $\tau^{\text{Bayes}} = \text{sign}(2p(\tau|\mathbf{s}) - 1)$. This will minimize the error rate because we will only make an error when the predictive probability for the correct output is less than one half. Since $p(\tau|\mathbf{s})$ itself gives the probability for the output label the average error on the specific input will be $\sum_{\tau=\pm 1} p(\tau|\mathbf{s})\Theta(1 - 2p(\tau|\mathbf{s}))$ where $\Theta(x)$ is the step function ($\Theta(x) = 1$ for $x > 0$ and 0 otherwise)[1]. Averaging over the input distribution we get the generalization error of the Bayes algorithm

$$\epsilon^{\text{Bayes}} = \int d\mathbf{s} p(\mathbf{s}) \sum_{\tau=\pm 1} p(\tau|\mathbf{s})\Theta(1 - 2p(\tau|\mathbf{s})) \ . \tag{5}$$

In the following we will calculate $p(\tau|\mathbf{s})$ and ϵ^{Bayes} for a specific learning scenario.

3. Learning a perceptron rule

In this section we will consider learning a perceptron rule $\tau = \text{sign}(\mathbf{w} \cdot \mathbf{s})$. The output will be flipped with a probability of κ, thus

$$p(\tau|\mathbf{w}, \mathbf{s}) = (1 - \kappa)\Theta(\tau \mathbf{w} \cdot \mathbf{s}) + \kappa\Theta(-\tau \mathbf{w} \cdot \mathbf{s}) \ . \tag{6}$$

We will consider the thermodynamic limit $N \to \infty$ and spherical zero mean unit variance inputs, i.e. $\overline{s_i} = 0$ and $\overline{s_i s_j} = \delta_{ij}$. [9] derived the predictive probability using a cavity argument [10] which is expected to become exact in this limit. Though $\mathbf{w}$ might not be Gaussian the projections on the random new direction $\mathbf{s}$ will according to the central limit theorem add up to a Gausian variable with mean $\langle \mathbf{w} \rangle \cdot \mathbf{s}$ and variance $\rho \equiv \sum_{i,j} s_i s_j (\langle w_i w_j \rangle - \langle w_i \rangle \langle w_j \rangle) = \langle \mathbf{w} \cdot \mathbf{w} \rangle - \langle \mathbf{w} \rangle \cdot \langle \mathbf{w} \rangle$, where we in the last step made an order $\frac{1}{N}$ error by replacing $s_i s_j$ with $\overline{s_i s_j} = \delta_{ij}$. Using the normality of $\mathbf{w} \cdot \mathbf{s}$ it is straight forward to write down the predictive probability

$$p(\tau|\mathbf{s}) = (1 - 2\kappa) H\left(-\tau \frac{\langle \mathbf{w} \rangle \cdot \mathbf{s}}{\sqrt{\rho}}\right) + \kappa \ , \tag{7}$$

[1] For comparison consider the so called Gibbs algorithm in which one chooses the output label in proportion to its probability. The average error is $\sum_{\tau=\pm 1} p(\tau|\mathbf{s})(1 - p(\tau|\mathbf{s}))$ which is clearly higher than the error of the Bayes algorithm.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 169-174

where $H(x) = \int_x^\infty e^{-t^2/2} dt/\sqrt{2\pi}$. This will allow us to derive the well known results for the Bayes classifier [11] $\tau^{\text{Bayes}} = \text{sign}\left(\langle \mathbf{w} \rangle \cdot \mathbf{s}\right)$ and the Bayes error rate [12]

$$\epsilon^{\text{Bayes}} = \kappa + (1 - 2\kappa)\frac{1}{\pi} \arccos\left(\sqrt{\frac{\langle \mathbf{w} \rangle \cdot \langle \mathbf{w} \rangle}{\langle \mathbf{w} \cdot \mathbf{w} \rangle}}\right) . \tag{8}$$

The result for the Bayes classifier tells us that we only need the posterior mean of the weights to make Bayesian predictions. Now we will go on to study two specific weight priors and corresponding choices for the approximate distribution $p(\mathbf{w}|A_m)$.

Spherical prior

For a spherical Gaussian prior $p(\mathbf{w}) = \frac{1}{\sqrt{2\pi}^N} e^{-\mathbf{w}\cdot\mathbf{w}/2}$ an obvious choice for $p(\mathbf{w}|A_m)$ is a Gaussian distribution. In [5] the case of a general Gaussian approximation to the posterior is discussed. However, for the scenario studied here the situation simplifies [5]. Because both the prior and inputs are spherical the off-diagonal of the covariance matrix will vanish in the thermodynamic limit. The diagonal elements will furthermore be non-fluctuating quantities equal to ρ/N. Due to the scale invariance of $p(\tau|\mathbf{w}, \mathbf{s})$ the prior will fix the scale of the weights: $\langle \mathbf{w} \cdot \mathbf{w} \rangle = N$. The final update rule for the posterior mean is

$$\langle \mathbf{w} \rangle_{m+1} = \langle \mathbf{w} \rangle + \frac{\rho}{N} \frac{\partial}{\partial \langle \mathbf{w} \rangle} \ln p(\tau^{m+1}|\mathbf{s}^{m+1}) \tag{9}$$

where $\rho = N - \langle \mathbf{w} \rangle \cdot \langle \mathbf{w} \rangle$ and $p(\tau^{m+1}|\mathbf{s}^{m+1})$ is the predictive probability from eq. (7). Note that this update rule is identical to the one found in [1] which was derived using a variational principle for maximizing the average generalization gain in each step.

To get a recursion relation for the average generalization error eq. (8) (expressed by the order parameter $\langle \mathbf{w} \rangle \cdot \langle \mathbf{w} \rangle$) we take the dot product of eq. (9) with itself. Since the order parameter is expected to be self averaging in the thermodynamic limit we can average the recursive relation over the joint probability distribution of the input and output $p(\mathbf{s})p(\tau|\mathbf{s})$. Again we obtain the same results as in [1].

Binary prior

For the binary prior $p(\mathbf{w}) = \prod_i \left[\frac{1}{2}\delta(w_i - 1) + \frac{1}{2}\delta(w_i + 1)\right]$ we will choose a biased binary distribution $p(\mathbf{w}|A_m) = \prod_i \left[\frac{1+\langle w_i \rangle}{2}\delta(w_i - 1) + \frac{1-\langle w_i \rangle}{2}\delta(w_i + 1)\right]$. To derive the recursion relation for the posterior mean of the i'th weight we have to apply a cavity argument to relate the posterior mean without the i'th weight to the full posterior [13]. Doing this we get the following fixpoint equations

$$\langle w_i \rangle_{m+1} = \frac{\langle w_i \rangle + \tanh(x_i)}{1 + \langle w_i \rangle \tanh(x_i)} \tag{10}$$

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 169-174

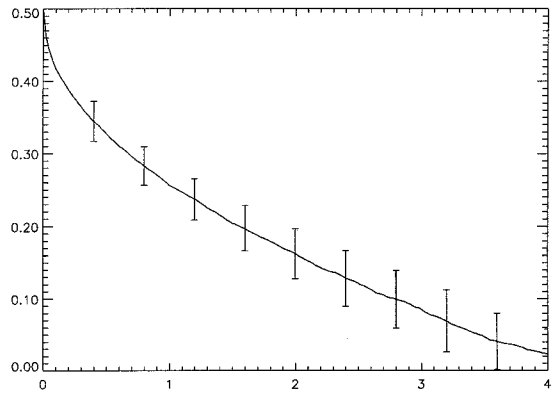

Figure 1: The learning curve ϵ^{Bayes} versus m/N for binary weight prior with $N = 50$ and $\kappa = 0$.

with

$$x_i = \frac{\partial}{\partial \langle w_i \rangle} \ln p(\tau^{m+1}|\mathbf{s}^{m+1}) - \langle w_i \rangle_{m+1} \frac{\partial^2}{\partial \langle w_i \rangle^2} \ln p(\tau^{m+1}|\mathbf{s}^{m+1}) \qquad (11)$$

where $p(\tau^{m+1}|\mathbf{s}^{m+1})$ is the predictive probability eq. (7). The second term in the equation for x_i represents the correction due to expressing everything in terms of full posterior averages. Note that we get fixpoint equations in this case because the correction term depends on $\langle w_i \rangle_{m+1}$. In the simulations we will avoid this problem by setting $\langle w_i \rangle_{m+1} = \langle w_i \rangle$ on the rhs. We expect $\langle w_i \rangle_{m+1} - \langle w_i \rangle$ to be order $1/\sqrt{N}$. The error we make should therefore be of order $1/N$ The results from averaging the learning curve (ϵ^{Bayes} versus m/N) over 50 independent runs is plotted in figure 1. We observe that in some cases the algorithm get trapped in a non-optimal solution. In most cases however a phase transition to the true rule occurs for m/N between 3 and 4.

4. Conclusion and outlook

We have presented a framework for Bayesian online learning. In usual online learning a new estimate of the solution to the learning problem is found from the current estimate and the new example. In the Bayesian online approach we choose an approximate posterior and find the new estimate of the parameters of the posterior from the current approximate posterior and the new example. This approach will fail if the distribution we choose exclude some of the a priori possible solutions. It is therefore natural to choose the parametric form of the approximate posterior such that it initially may be set equal to the prior of weights. We have studied two thermodynamic limit perceptron scenarios – learning a simple perceptron rule with respectively Gaussian and binary weight prior. In the first case we approximated the posterior with a Gaussian and rederived the optimal Hebb-type algorithm [1]. This has also been anticipated

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 169-174

[5]. In the second case we approximated the posterior with a biased binary distribution.

There are two directions in which this work could be extended. First of all setting up a general criteria for how in each online step to make optimal use of the training information. Secondly the straight forward extension within the statistical mechanics framework of this approach to multilayer neural networks scenarios and the derivation of theoretical learning curve for the binary perceptron scenario.

Acknowledgements

This research is supported by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center (CONNECT).

References

[1] O. Kinouchi and N. Caticha, J. Phys. A: Math. Gen. **25** 6243 (1992).

[2] M. Biehl and P. Riegler, Europhys. Lett. **28**, 525 (1994).

[3] D. Saad and S. A. Solla, Phys. Rev. Lett. **74**, 4337 (1995).

[4] C. Van den Broeck and P. Reimann, Phys. Rev. Lett. **76** 2188 (1996).

[5] M. Opper, Phys. Rev. Lett. **77**, 4671 (1996).

[6] O. Winther and S. A. Solla, In preperation (1997).

[7] J. O. Berger; *Statistical Decision theory and Bayesian Analysis*, Springer-Verlag, New York (1985).

[8] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag (1982).

[9] M. Opper and O. Winther, Phys. Rev. Lett. **76**, 1964 (1996),

[10] M. Mézard, G. Parisi and M. A. Virasoro. *Spin Glass Theory and Beyond*, Lecture Notes in Physics, 9, World Scientific (1987).

[11] T. L. H. Watkin, Europhys. Lett. **21**, 871 (1993).

[12] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991) and
M. Opper and D. Haussler, in *IVth Annual Workshop on Computational Learning Theory (COLT91)*, Morgan Kaufmann (1991).

[13] M. Mézard, J. Phys. A **22**, 2181 (1989).