ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 161-166

Knowledge and generalisation in simple learning systems

David Barber, David Saad

Department of Physics,University of Edinburgh
Edinburgh EH9 3JZ, U.K.
D.Barber@ed.ac.uk

Abstract. We examine the effect of an increase in knowledge/constraints on the generalisation error of adaptive learning systems, specifically the linear perceptron. For a constraint which is then tightened, the new version space becomes a subset of the original, and we examine to what extent it is necessarily true that the average generalisation error must decrease. We show that in the case of the linear perceptron, increasing the knowledge such that the teacher space remains convex is *not* a sufficient condition for a reduction in average generalisation error.

1. Introduction

In this paper we deal with the scenario of learning from examples (see *e.g.*,[1]), in which knowledge about the problem we are trying to learn is contained in both the presented examples and in additional a priori assumptions about the form of the problem. We examine the effect an increase in such knowledge has on the generalisation performance of simple learning systems, concentrating in particular on the linear perceptron. We assume that a training set of input/output pairs is generated by some teacher function, and the task is to find a student whose outputs match closely the outputs of the teacher function on the training set. It is well known that without any constraints on the teacher function that generates the training set, it is an impossible task to find a student that generalises to unseen examples [4]. A priori assumptions are therefore made as to the form of the teacher, that is, restrictions are imposed on the space of teacher functions. Throughout this paper, we assume that the spaces of which both the teacher and student functions are members, are the same. The learning problem is then realisable in the sense that amongst the student space, there is a student that will match perfectly the output of the teacher on all possible inputs. We denote the teacher/student space of functions by $F(\Psi)$, and a particular mapping as $y = f(x, \theta)$ for $f \in F(\Psi)$ and $\theta \in \Psi$ where the output is denoted by y, and the input by x. A particular mapping that a function performs is then represented by the point θ in the parameter space Ψ. We assume that a single teacher θ^0 generates the set of training data $\mathcal{L} = \{x^\sigma, f(x^\sigma, \theta^0)\}$. In the learning problem, one attempts to

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 161-166

find a student $f(x,\theta)$ that matches the teacher $f(x,\theta^0)$ on the training set. To measure the extent to which the student has learnt the teacher, an error measure $\epsilon(\theta,\theta^0,x)$ is defined. All the information the student has about the teacher is then represented by the space $\Theta \subset \Psi$, alternatively, the student is constrained to lie within the space Θ. In Section 2. we review the general learning theory and introduce the concept of reducivity to deal with the issue of constraint increases. In Section 3., we examine reducivity in higher dimensions, using the linear perceptron as the function space $F(\Psi)$. In Section 4. we conclude with a summary and discussion of the main results of the paper.

2. General theory

2.1. The generalisation function

To measure how well the student performs on the training set, the training energy is formed, $E_{tr} \propto \sum_{\sigma=1}^{p} \epsilon(\theta,\theta^0,x^\sigma)$. The student is found by minimising the training error with respect to the parameter θ, whilst adhering also to additional a priori constraints. This is typically achieved by stochastic gradient descent, resulting in a post training Gibbs distribution of students, $P(\theta|\mathcal{L}) \propto P^{pri}(\theta)\exp(-E_{tr}/T)$ where the temperature, T, controls the randomness of the stochastic algorithm (see *e.g.*,[3]). $P^{pri}(\theta)$ is the prior on the student, expressing the a priori constraints on the students. In the limit of zero T, the distribution of students becomes uniform over those that have zero training error and satisfy the a priori constraints; this space of student functions is known as the version space [3], which we denote by Θ. For the rest of the paper, zero T is implied. To find the expected error the student makes on a random example input, termed the generalisation function, we average the error over the input distribution, $P(x)$, giving $\epsilon_f(\theta,\theta^0) = \int dx P(x)\epsilon(\theta,\theta^0,x)$. Hence, if we know the teacher, we can find the expected error for any student from Θ. If, however, all the knowledge we have about the student and teacher is that contained in the training set and prior, we know only that the teacher and student both lie in Θ.

2.2. The generalisation error

The average performance of a random student selected from the version space is termed the generalisation error, which one expects to improve as the number of training examples increases. In the scenario that we have so far been considering, Θ expresses all the knowledge we have about the student, after presentation of none, or many examples, and we assume that this is also all the knowledge we have about the teacher, defining the generalisation error accordingly,

$$\epsilon_g(\Theta) = \left\langle \epsilon_f(\theta,\theta^0) \right\rangle_{\theta\in\Theta,\theta^0\in\Theta}$$

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 161-166

where $\langle..\rangle_{\theta^0\in\Theta}$ and $\langle..\rangle_{\theta\in\Theta}$ represent averages over the version space Θ. We write $\epsilon_g(\Theta)$ to emphasize that the generalisation error is a function of the version space.

Intuitively, one expects that any further restrictions or a priori assumptions, resulting in a smaller version space, must necessarily reduce the generalisation error. To test this intuition, we make the following definition.

- $F(\Theta')$ is an 'error reduced' function space of $F(\Theta)$ if $\epsilon_g(\Theta') < \epsilon_g(\Theta)$ for $\Theta' \subset \Theta$, and we say that reducivity holds.

In this paper we examine which subsets Θ' of Θ are error reducing, according to the preceding definition. We mention briefly that one can also consider the generalisation error for a fixed teacher, $\epsilon_g(\theta^0, \Theta) = \langle \epsilon_f(\theta, \theta^0) \rangle_{\theta\in\Theta}$, and one could also check reducivity in light of this. We show in a later section, however, that the main results of this paper also hold for $\epsilon_g(\theta^0, \Theta)$, and concentrate accordingly on $\epsilon_g(\Theta)$.

3. The linear perceptron

For the noise free linear perceptron, the inputs are represented by n dimensional real vectors, $\mathbf{x} \in \Re^n$, and the output is a single valued real variable, $y \in \Re$ [1]. The inputs $\mathbf{x}$ are assumed drawn independently and identically, from a zero mean, unit covariance matrix Gaussian distribution. The teacher outputs are given by $f(\mathbf{x}, \mathbf{w}^0) = \mathbf{w}^0 \cdot \mathbf{x}/\sqrt{n}$. Similarly, the student outputs are $f(\mathbf{x}, \mathbf{w}) = \mathbf{w} \cdot \mathbf{x}/\sqrt{n}$. We also impose the additional a priori spherical constraint on both the student and teacher parameters, $\mathbf{w} \cdot \mathbf{w} = \mathbf{w}^0 \cdot \mathbf{w}^0 = n$. The error measure is taken to be proportional to the squared difference between the teacher and student outputs, $\epsilon(\mathbf{w}, \mathbf{w}^0, \mathbf{x}) = (\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^0 \cdot \mathbf{x})^2/2n$.

3.1. A two dimensional version space

We look now at the three dimensional linear perceptron. A point on the surface of a three dimensional sphere of radius $r = \sqrt{3}$ is given by the ordered pair (ϕ, θ), which represents the usual spherical polar coordinate parameterisation.

We assume that the version space is given by: $\Theta = \{(\phi, \theta), \phi \in [a, b], \theta \in [c, d]\}$.

A straightforward calculation gives

$$\epsilon_g(\Theta) = 1 - \frac{1}{(d-c)^2}\left(\lambda\left(\cos(d) - \cos(c)\right)^2 + \left(\sin(d) - \sin(c)\right)^2\right),$$

where $\lambda = 2\left(1 - \cos(b-a)\right)/(b-a)^2$. To violate reducivity we look for regions, for example, such that $\partial\epsilon_g(\Theta)/\partial c > 0$, and we plot one such region in figure(1a). Indeed, there are infinitely many pairs (Θ, Θ') that violate reducivity. At this point, the reader may well conjecture that reducivity would be guaranteed for convex regions Θ and $\Theta' \subset \Theta$. (In general, a region is convex if the geodesic connecting any two points lies wholly within the region itself). Perhaps somewhat surprisingly, we demonstrate in the next section that convexity is not a sufficient condition for reducivity.

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 161-166

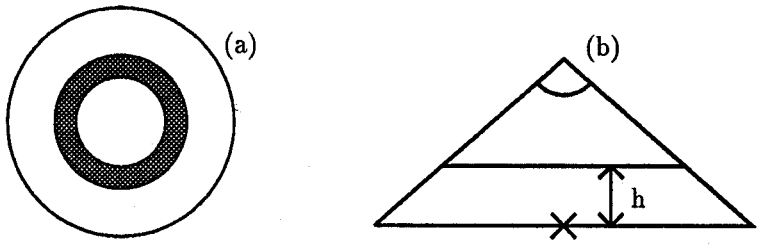

Figure 1: Version spaces that violate reducivity. (a) View from above the pole of a sphere of radius $\sqrt{3}$. The shaded region represents the version space, $\Theta = \{\theta \in [0.4, 0.6], \phi \in [0, 2\pi]\}$. Making Θ smaller by pushing the inner boundary towards the outer boundary does not result in a reduction in generalistion error. (b) Counter example used to show that convexity is not a sufficient condition for reducivity. We take the hypotenuse to have length 2.

3.2. Euclidean approximation to the version space

For simplicity, we now consider the approximation in which the version space is small enough such that the region can be considered Euclidean. For the linear perceptron described above, this corresponds to a region small enough such that the curved surface of the hypersphere appears 'flat'. By writing $\mathbf{w} = \mathbf{c} + \tilde{\mathbf{w}}$, and $\mathbf{w}^0 = \mathbf{c} + \tilde{\mathbf{w}}^0$, where $\mathbf{c}$ lies in the space Θ, we write the generalisation error as

$$\epsilon_g(\tilde{\Theta}) = \frac{1}{2n} \left\langle \left(\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^0\right)^2 \right\rangle_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^0 \in \tilde{\Theta}},$$

where $\tilde{\Theta}$ is the approximately flat region on the sphere. Notice that this can be written in the form,

$$\epsilon_g(\tilde{\Theta}) = \frac{1}{n} \left(\langle \tilde{\mathbf{w}}^2 \rangle_{\tilde{\mathbf{w}} \in \tilde{\Theta}} - \langle \tilde{\mathbf{w}} \rangle^2_{\tilde{\mathbf{w}} \in \tilde{\Theta}} \right).$$

We now consider an infinitesimal decrease in the space $\tilde{\Theta}' = \tilde{\Theta} - \Delta$. For a uniform distribution over the space, and ignoring terms in Δ^2, we can write, with a slight abuse of notation,

$$\epsilon_g(\tilde{\Theta}') - \epsilon_g(\tilde{\Theta}) \approx \frac{\Delta}{n\tilde{\Theta}} \left(\langle \tilde{\mathbf{w}}^2 \rangle_{\tilde{\mathbf{w}} \in \tilde{\Theta}} - \langle \tilde{\mathbf{w}}^2 \rangle_{\tilde{\mathbf{w}} \in \Delta} \right), \tag{1}$$

where Δ and $\tilde{\Theta}$ are the surface contents of Δ and $\tilde{\Theta}$ respectively. In eq.(1), we have assumed w.l.o.g. that $\langle \tilde{\mathbf{w}} \rangle_{\tilde{\mathbf{w}} \in \tilde{\Theta}} = 0$, *i.e.*,that the origin, $\mathbf{c}$, is taken to be the centroid of $\tilde{\Theta}$. Reducivity holds then for the condition

$$\langle \tilde{\mathbf{w}}^2 \rangle_{\tilde{\mathbf{w}} \in \Delta} > \langle \tilde{\mathbf{w}}^2 \rangle_{\tilde{\mathbf{w}} \in \tilde{\Theta}}. \tag{2}$$

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 161-166

Note that this is a general condition, holding for any dimension. Using this, we can show that convexity (for the linear perceptron at least) is not a sufficient condition for reducivity to hold.

3.2.1. Convexity is not sufficient for reducivity

Eq.(2) will not be satisfied for regions, Δ, sufficiently close to the centroid. This observation leads to the following two dimensional counter example. Let the convex region $\tilde{\Theta}$ be the larger triangle as shown on figure(1b). By explicit calculation, one finds ϵ_g(tri)=4/9 for the angle shown a right angle. We now take $\tilde{\Theta}'$, a convex subset of $\tilde{\Theta}$, to be the trapezium as shown, for which, in the limit $h \rightarrow 0$, ϵ_g(trap)=2/3. Hence $\epsilon_g(\tilde{\Theta}') > \epsilon_g(\tilde{\Theta})$, demonstrating the insufficiency of convexity as a condition for reducivity.

At this point we refer back to Section(2.2.) and note that we can readily find an example of a fixed teacher for which an increase in the students knowledge results in an increase in $\epsilon_g(\theta^0, \Theta)$. In the above trapezium/triangle example, consider a very flat triangle, for which the marked angle tends to π. We take the teacher to be positioned at the cross marked in figure(1b), for which, $\epsilon_g(\times, tri) = 1/6$. Taking again, $\tilde{\Theta}'$ to be the infinitely thin trapezium, we have $\epsilon_g(\times, trap) = 1/3$, which is larger than $\epsilon_g(\times, tri)$.

The previous arguments have been aimed at infinitesimal, local alterations to $\tilde{\Theta}$, and we consider briefly an example of global enlargement. We envisage situations in which the boundary of $\tilde{\Theta}$ can be expressed in a spherical coordinate system, $r = r(\phi, \theta, ..)$, which is the case for convex regions. The enlarged version space can then be defined by a new boundary, $r' = \lambda(\phi, \theta, ..)r(\phi, \theta, ..)$, for some $\lambda(\phi, \theta, ..) > 1$. Assuming we can bound λ by some extremum values, $\lambda_{min} < \lambda(\phi, \theta, ..) < \lambda_{max}$, it is straightforward to form an inequality such that the generalisation error of the larger version space is greater than the generalisation error of the smaller. For an enlargement which preserves the origin as the centroid, the condition in two dimensions is $\lambda_{min}^2 > \lambda_{max}$ (sufficient, but by no means necessary).

By examining eq.(1), we note that the greatest decrease in generalisation error is to be found for a region Δ furthest away from the centroid of the set. This is in line with the intuitive notion that we can improve generalisation most by increasing our knowledge about the teacher in those regions that contribute most to the generalisation error. One way to obtain this knowledge is to choose an input x such that the reply from the teacher will give us information about the teacher in the desired region; this is the concept of query learning (see *e.g.*, [2]).

4. Summary and discussion

We have examined the effect of constraints on the generalisation error of simple learning systems, concentrating in particular on the linear perceptron. Assuming that both the student and teacher lie in the version space of constraints, we

ESANN'1995 proceedings - European Symposium on Artificial Neural Networks
Brussels (Belgium), 19-20-21 April 1995, D-Facto public., ISBN 2-9600049-3-0, pp. 161-166

studied what effect increasing the constraint, by decreasing the version space, has on the generalisation error. We gave a simple example of a two dimensional version space for which decreasing the version space does not decrease the generalisation error. Furthermore, convexity of the version spaces is not a sufficient condition for the smaller version space to have lower generalisation error. In general it is a non-trivial problem to predict whether reducing the version space will reduce the generalisation error, and each case must be treated explicitly.

References

[1] A. Hertz, J. Krogh and G. Palmer. *Introduction to the theory of Neural Computation.* Addsion-Wesley, Redwood City California, 1991.

[2] P. Sollich. Query construction, entropy and generalization in neural network models. *Physical Review E*, 49:4637–4651, 1994.

[3] Rau A. Biehl M. Watkin, T. L. H. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65:499–556, 1993.

[4] D. H. Wolpert. On the connection between in-sample testing and generalisation error. *Complex Systems*, 6:47–94, 1992.