Regularization by Early Stopping in Single Layer Perceptron training

Sarunas Raudys[#] and Tautvydas Cibas[*]

[#]Department of Informatics, Vytautas Magnus University, Kaunas, Lithuania
e-mail:raudys@ktl.mii.lt
[*]Université de Paris-Sud, LRI, Bât.490, F-91405 Orsay Cedex, France
e-mail:cibas@lri.lri.fr

Abstract. Adaptative training of the non-linear single-layer perceptron can lead to the Euclidean distance classifier and later to the standard Fisher linear discriminant function. On the way between these two classifiers one has a regularized discriminant analysis. That is equivalent to the "weight decay" regularization term added to the cost function. Thus early stopping plays a role of regularization of the network.

1 Introduction and notations

A artificial feedforward neural networks (ANN) were found to be a solid tool in pattern recognition. Because of their nonlinearity and complex structures, their behavior is difficult to analyse analytically. To facilitate this task we analyse the single-layer perceptron (SLP) which is a key elementary grain in modern feedforward networks and has a strong tie with a classical discriminant analysis.

In classical statistical discriminant analysis, to discriminate between two pattern classes one uses a discriminant function

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \tag{1}$$

and performs a classification of the *p*-variate vector $\mathbf{x} = \left(x_1, x_2, \ldots, x_p\right)^T$ according to its sign. $\mathbf{w}^T$ denotes the vector transposition. The weights w_0, $\mathbf{w}=\left(w_1, w_2, \ldots, w_p\right)^T$ in (1) of the form

$$\mathbf{w}^F = \mathrm{S}^{-1}\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right), \qquad w_0 = -\tfrac{1}{2}\,\mathbf{w}^T\left(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}\right) \tag{2}$$

are the weights of a *Fisher discriminant function* (Fisher, 1936),

where $\mathrm{S} = \frac{1}{N_1+N_2-2}\sum_{i=1}^{2}\sum_{j=1}^{N_i}\left(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)}\right)\left(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}^{(i)}\right)^T$ is a sample pooled covariance matrix, $\bar{\mathbf{x}}^{(i)} = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{x}_j^{(i)}$ is a sample mean vector, $\mathbf{x}_j^{(i)}$ is the *j*-th learning vector from the *i*-th class and N_i is the learning vector's number from class *i*.

When one omits the covariance matrix S in (2) then on has the *Euclidean distance classifier*.

The SLP itself can form linear discriminant hyperplane in a highdimensional feature space and discriminate complicated objects. In the notations used in the

present paper, the SLP has p inputs and one output which is calculated by $output = o\left(\mathbf{w}^T\mathbf{x} + w_0\right)$, where $o(net)$ is a non-linear (activation) function. The weights $w_0, \mathbf{w}$ of the discriminant function will be learned during the training process. For two class classification problem we minimize the following mean squares cost function :

$$cost = \frac{1}{2\left(N_1+N_2\right)} \sum_{i=1}^{2} \sum_{j=1}^{N_i} \left(d_j^{(i)} - o\left(\mathbf{w}^T\mathbf{x}_j^{(i)} + w_0\right)\right)^2 \tag{3}$$

by adapting the weight vector according to a standard global gradient delta learning rule

$$\mathbf{w}_{(t+1)} = \mathbf{w}_{(t)} - \eta \frac{\partial cost_{(t)}}{\partial \mathbf{w}_{(t)}} \tag{4}$$

where $d_j^{(i)}$ is the desired output for input $\mathbf{x}_j^{(i)}$, $\mathbf{w}_{(t)}$ is a weight vector at instant t and η is a learning step.

It is known (Koford & Groner, 1966; Gallinari *et al.*, 1991), that when the number of training vectors from two pattern classes is the same (N_1=N_2), training of cost function (3) with linear transfer function $o(net)$=net (which becomes ADALINE, a prototype of modern SLP) leads to a weight vector $\mathbf{w}$ which is equivalent to (2), the weights of the standard Fisher linear discriminant function, who is asymptotically optimal when the classes are Gaussian with common covariance matrix.

The minimization of the cost function (3) with an additional "weight decay" regularization term $\lambda\mathbf{w}^T\mathbf{w}$ gives weights of the form (Raudys *et al.*, 1994)

$$\mathbf{w}^{FR} = \left(\mathrm{S} + \lambda \mathrm{I}\right)^{-1}\left(\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)}\right)k, \qquad w_0 = -\tfrac{1}{2}\,\mathbf{w}^T\left(\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)}\right) \tag{5}$$

where λ is a positive regularization constant, I is a $p \times p$ identity matrix and k denotes a scalar which does not depend on $\mathbf{x}$, the vector to be classified. Vector $\mathbf{w}^{FR}$ is in fact a weight vector of the *regularized discriminant analysis* (see e.g. Friedman, 1989; McLachlan, 1992).

In (Raudys, 1995) it was indicated that in a case N_1=N_2 and with centred data in a way that $\overline{\mathbf{x}}^{(2)} = -\overline{\mathbf{x}}^{(1)}$, after the first back-propagation learning step in a batch mode of the SLP initialized with zero weights, one obtains the weights equivalent to the Euclidean distance classifier weight vector. On the way between the Euclidean distance classifier and the Fisher classifier we have regularized discriminant analysis.

An objective of the present paper is, in comparison of SLP training with the classical statistical discriminant analysis, to show that the *regularization of SLP by "weight decay" can be replaced by early stopping* in cost function (3) minimization case. We present here a more strict proof when in (Raudys, 1995) to demonstrate that in some conditions a SLP at the beginning of its training reacts like the Euclidean distance classifier and continuing its training until it recovers the functions of a Fisher classifier, it discriminates between two classes like in the regularized discriminant analysis case. We will show that *the criterion used to find weights of the SLP changes during the training process.*

2 Regularized discriminant analysis in SLP design

Consider the SLP's batch mode training rule where one uses a gradient of cost function (3) to make weight changes according to equation (4). Let's use an activation function of type $o(net)=\frac{e^{net}-e^{-net}}{e^{net}+e^{-net}}$. Suppose that $d_j^{(1)}=1$, $d_j^{(2)}=-1$, data is centred in a way that $\bar{\mathbf{x}}^{(2)}=-\bar{\mathbf{x}}^{(1)}$ and N_1=N_2=N.

For very small initial weights $\mathbf{w}^T\mathbf{x}_j^{(i)}+w_0$ is close to zero, therefore $\frac{\partial o(\mathbf{x})}{\partial \mathbf{x}}=d\mathbf{x}$ and $o(net)$=net. With a condition $\bar{\mathbf{x}}^{(2)}=-\bar{\mathbf{x}}^{(1)}$ and N_1=N_2 we have $\frac{\partial cost_{(t)}}{\partial w_{0(t)}}=w_{0(t)}$, so we will interesting only in the weights $\mathbf{w}$ evolution during the training.

From (3) we have

$$\frac{\partial cost_{(t)}}{\partial \mathbf{w}_{(t)}}=\frac{1}{2N}\sum_{i=1}^{2}\sum_{j=1}^{N}\left(d_j^{(i)}-\mathbf{w}_{(t)}^T\mathbf{x}_j^{(i)}\right)\left(-\mathbf{x}_j^{(i)}\right)=-\frac{1}{2}\Delta\bar{\mathbf{x}}+\frac{1}{2}\mathrm{K}\mathbf{w}_{(t)} \tag{6}$$

where $\Delta\bar{\mathbf{x}}=\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}$ and $\mathrm{K}=\frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{N}\mathbf{x}_j^{(i)}\left(\mathbf{x}_j^{(i)}\right)^T$.

When the prior weights are very small one may assume $\mathbf{w}_{(0)}$=0. Then after the first learning iteration for the weight vector $\mathbf{w}$ one obtains

$$\mathbf{w}_{(1)}=\mathbf{w}_{(0)}-\eta\frac{\partial cost_{(0)}}{\partial \mathbf{w}_{(0)}}=\frac{\eta}{2}\left(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\right) \tag{7}$$

It is the weight vector of the Euclidean distance classifier designed for centred data.

Now we will analyse the changes of the weight vector in the second and following iterations.

The use of total gradient adaptation rule (4) with (7) and gradient (6) results in:

$$\mathbf{w}_{(2)}=\mathbf{w}_{(1)}-\eta\frac{\partial cost_{(1)}}{\partial \mathbf{w}_{(1)}}=\frac{\eta}{2}\Delta\bar{\mathbf{x}}-\eta\left(-\frac{1}{2}\Delta\bar{\mathbf{x}}+\frac{1}{2}\mathrm{K}\mathbf{w}_{(1)}\right)$$

$$=\left(\mathrm{I}-\left(\mathrm{I}-\frac{\eta}{2}\mathrm{K}\right)^2\right)\mathrm{K}^{-1}\Delta\bar{\mathbf{x}}$$

and further,

$$\mathbf{w}_{(t)}=\left(\mathrm{I}-\left(\mathrm{I}-\frac{\eta}{2}\mathrm{K}\right)^t\right)\mathrm{K}^{-1}\Delta\bar{\mathbf{x}} \tag{8}$$

where K was defined above. By definition matrix K is not singular, so it has an inverse. The use of the first terms of the expansion $\left(\mathrm{I}-\frac{\eta}{2}\mathrm{K}\right)^t=\mathrm{I}-t\frac{\eta}{2}\mathrm{K}+\frac{1}{2}t(t-1)\left(\frac{\eta}{2}\right)^2\mathrm{K}^2-\ldots$ in (8) for small η and t results in

$$\mathbf{w}_{(t)}\approx\left(t\frac{\eta}{2}\mathrm{K}-\frac{1}{2}t(t-1)\left(\frac{\eta}{2}\right)^2\mathrm{K}^2\right)\mathrm{K}^{-1}\Delta\bar{\mathbf{x}}$$

$$=t\frac{\eta}{2}\left(\mathrm{I}-\frac{1}{2}(t-1)\frac{\eta}{2}\mathrm{K}\right)\Delta\bar{\mathbf{x}}.$$

Further, the use of the first terms of the expansion $\left(\mathrm{I}-\beta\mathrm{K}\right)^{-1}=\mathrm{I}+\beta\mathrm{K}+\ldots$, with supposition that η and t are small, gives

$$\mathbf{w}_{(t)} = t\,\frac{\eta}{2}\left(\mathrm{I} + \frac{1}{2}(t-1)\,\frac{\eta}{2}\,\mathrm{K}\right)^{-1}\Delta\bar{\mathbf{x}}$$

$$= t\,\frac{\eta}{2}\left(\mathrm{I} + (t-1)\,\frac{\eta}{2}\left(\frac{N-1}{N}\mathrm{S} + \frac{1}{4}\Delta\bar{\mathbf{x}}\Delta\bar{\mathbf{x}}^{T}\right)\right)^{-1}\Delta\bar{\mathbf{x}}. \qquad (9)$$

Assuming matrix $\mathrm{I}\lambda+\mathrm{S}$ is not singular, after some matrix algebra (see e.g. the use of Bartlett formula in Raudys *et al.,* 1994) from (9) we obtain

$$\mathbf{w}_{(t)} = \left(\mathrm{I}\,\frac{2}{(t-1)\eta}\,\frac{N}{(N-1)} + \mathrm{S}\right)^{-1}\Delta\bar{\mathbf{x}}\,\frac{tN}{(t-1)(N-1)}\,k. \qquad (10)$$

Weight vector $\mathbf{w}_{(t)}$ in (10) is equivalent to the weight vector $\mathbf{w}^{FR}$ of *regularized discriminant analysis* (5) with regularization parameter $\lambda = \frac{2N}{(t-1)\eta(N-1)}$; k is a same scalar as in (5). Equation (10) indicates explicitly that with an increase in t, the number of iterations, the weight vector $\mathbf{w}_{(t)}$ moves from an Euclidean distance classifier (when t=1, see eq.(7)) to a Fisher classifier, since $\lambda \to 0$ when $t \to \infty$.

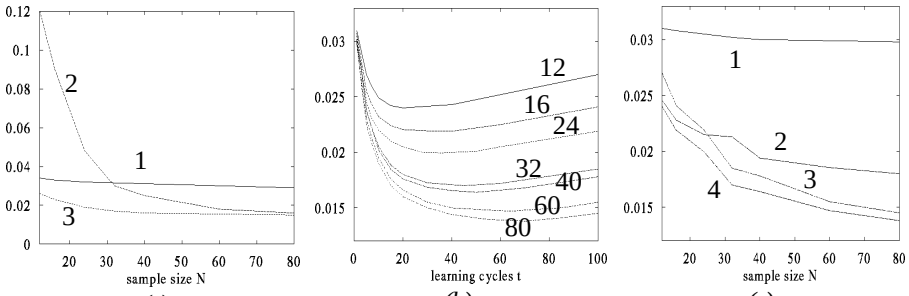
3 Regularization and classifier complexity

In statistical pattern recognition it is well known that the sensitivity of a classifier to the learning sample size depends on the complexity of the classifier (Raudys, 1970). For example, in the case of classification of individuals into one of two multivariate Gaussian classes with different means μ_1, μ_2 and sharing common covariance matrix Σ it is known that for small learning sets it is preferable to use the simple structured Euclidean distance classifier while for larger sets it is better to use the complex structured Fisher classifier (Raudys, 1970; Jain & Chandrasekaran, 1982; Raudys & Jain, 1991).

The use of the regularized discriminant analysis (5) offers a great number of other algorithms. They differ in λ, the optimal value of which depends on N, the learning sample size: λ decreases as N increases (see e.g. Raudys *et al.,* 1994).

In our simulations we have used two class 20-variate Gaussian centred data with unit variance for all variables, correlation between all the variables was ρ=0.1495, and $(\mu_1\text{-}\mu_2)^T$=(0.0125, 0.0778,...,1.1867, 1.2519)^T.

In Fig.(a) we present the curves of the generalization errors versus sample size calculated theoretically for three classifiers (Euclidean distance, Fisher and regularized discriminant analysis), performed by 120 repetitions of the experiment with different randomly chosen learning sets of size $12 \le N \le 80$. Regularized discriminant analysis was performed with the optimal value of the regularization parameter for each individual learning set. As we see in Fig.(a) it is evident that in finite learning-set cases the regularized discriminant analysis with optimal λ (curve-3 in Fig.(a)) outperforms the both here used statistical classifiers.



(a) *(b)* *(c)*

Figures: generalization error versus learning set size in (a),(c) and versus the number of learning iteration in (b). (a): (1) - Euclidean dist.classifier, (2) - Fisher classifier, (3) - regularized discr.analysis. (b): SLP training with N=12,16,24,32,40,60,80. (c): (1) - SLP after one training it. (⇔ Euclidean dist. classif.), (2) - SLP after 13 training it., (3) - SLP after 100 training it. and (4) - SLP after opt.number of training it. to obtain min.of EP_N

4 Simulation with the single layer perceptron

Numerous simulation experiments (see e.g. previous section, as well as Friedman, 1988; McLachlan, 1992) indicate that regularized discriminant analysis with optimal value of the smoothing parameters outperforms other parametric statistical classification rules. This is absolutely true when the pattern classes are Gaussian and one uses statistical classifiers designed for Gaussian models. For real non-Gaussian data, however, the classical parametric classifiers can fail to obtain good generalization. In this sense "nonparametric" classifiers based on single-layer and multi-layer perceptrons can appear to be better candidates. Experiments with feedforward ANN confirm this. Moreover, optimaly regularized networks usually help to reduce generalization error.

In section 2 it was shown that, in the sense of classical pattern recognition, while training the SLP, on the way between the Euclidean distance classifier and the Fisher classifier one discovers regularized discriminant analysis. Therefore one can expect that the generalization error of the optimally stopped SLP be lower than that at the very beginning and at the end of the iterative learning process. In Fig.(b) we present the evolution of the generalization error of SLP during learning for seven different learning-set sizes. Each of these curves is a mean value of 120 experiment with 120 different randomly chosen learning sets. We use the same data as in previous section. We initialized all weights to zero and used learning step η=1. For all graphs we found a minimum which depends on the sample size *N*: it increases with *N*.

In Fig.(c) we present the generalization error of the SLP versus the sample size obtained after a different number of learning iterations. Here we observe the same effect which is well known in statistical pattern recognition: for small sample sizes one needs to use simple structured classification rules (curve-2 in Fig.(c)) and only for large sample sizes one can use the complex ones (curve-3). The use an optimal number of learning iterations individually for each sample size (curve-4) results in

the smallest generalization error which is very close to the generalization error of the regularized discriminant analysis with value λ_{opt} (curve 3 in Fig.(a)).

The above observations agree fairly well with the theory discussed in the section 2: *the number of learning iterations plays a role of the regularization parameter and thus helps reduce the generalization error*.

5 Conclusions

Regularization of statistical classifiers and neural networks is a powerful tool which helps to obtain the classifier of optimal complexity and to reduce the generalization error. The number of learning steps in the single-layer perceptron training as well as the step size play a role of the regularization parameter. The influence of these parameters however can be compensated by other regularization methods such as addition of a "weight decay" regularization term to the cost function, a change in the target value, noise injection, etc.

Thus one regularization method can be compensated by another one. The optimal values of the regularization parameters (including the number of learning iterations) depend on the data (the pattern classification problem to be solved), the complexity of the classifier (the number of inputs in the SLP design), and, it is very important to stress, the learning-set size. In the experimental results reported in this paper we have presented only mean values obtained from 120 repetitions of the experiments. It is worth to mention that in individual learning sets we observe significant deviations of the results and the optimal values of the regularization parameters (here, the number of learning iterations) varies with each individual learning set.

References

Fisher R.A.(1936). The use of multiple measurements in taxonomic problems. *Ann.of Eugenics*, Vol.7, No.2, London, 179-188.Record, Part 4, 96-104.

Friedman J.M.(1989). Regularized discriminant analysis. *J.American Statistical Association*, Vol.84, 165-175.

Gallinari P., Thyria S., Badran F., Fogelman-Soulie F.(1991). On relations between discriminant analysis and multilayer perceptrons. *Neural Networks*, Vol.4, 349-360.

Jain A., Chandrasekaran B.(1982). Dimensionality and size considerations in pattern recognition practice. *Handbook of Statistics*, Vol.2, 835-855, North Holland.

Koford J.S., Groner G.F.(1966). The use of an adaptive threshold element to design a linear optimal pattern classifier. *IEEE Trans.on Inf. Theory*, Vol.IT-12, 42-50.

McLachlan G.J.(1992). *Discriminant Analysis and Statistical Pattern Recognition*, Willey.

Raudys S.(1970). On the problems of sample size in pattern recognition. Proc. *2nd All-Union Conf.Statist.Methods in Control Theory*, Moscow, Nauka, 64-67 (in Russian).

Raudys S.(1995). A negative weight decay or antiregularization. Proc.of *ICANN'95*, Paris, Vol.2., 449-454.

Raudys S., Jain A.K.(1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans.on Pattern Analysis and Machine Intelligence*, Vol.13, 252-264.

Raudys S., Skurichina M., Cibas T., Gallinari P.(1994). Optimal regularization of linear and nonlinear perceptrons. Proc.*Int.Conf.Neural Networks and Applications*, Marseilles.