Statistical Risk Analysis for Classification and Feature Extraction by Multilayer Perceptrons

Chanchal Chatterjee and Vwani Roychowdhury
School of Electrical and Computer Engineering
Purdue University, West Lafayette Indiana 47907-1285

Abstract

We investigate the training of multilayer perceptrons with the commonly used mean square error (MSE) criterion, and demonstrate a number of novel connections between the neural network operations and the Bayes risk analysis. Although previous research shows a number of connections from seemingly different criteria, we establish a common statistical framework to derive a generalized version of most, if not all, of these results, and also present several new results. We discuss the following: (1) We present two equivalent cost functions, and show that the MSE at the network output is equivalent to these cost functions for large samples. (2) We show that if the network performs a weighted classification, then the network output estimates the conditional risk. (3) We next show that if the final layer of the network is linear, then minimizing the MSE at the output, also maximizes a generalized criterion for nonlinear discriminant analysis (NDA). (4) We show that for a network with linear output layer, the outputs sum to one, and behave like probabilities. This new result allows us to estimate conditional risks at the network output, and also perform NDA at the final hidden layer. (5) Results for the uniform costs show that the MSE at the output is a tight upper bound of the error probability of the Bayes decision rule.

1. Introduction

We investigate the training of multilayer perceptrons with the commonly used mean square error (MSE) criterion, and demonstrate a number of novel connections between the neural network operations and the Bayes risk analysis. Previous research into this area has shown a number of results on the use of the network as a probability estimator [5,6,8,9,10], and as a nonlinear discriminant analysis feature extractor [4,11]. Each one of these results are derived from a seemingly different viewpoint, and using different cost functions. In this study, we establish a general framework for statistical analysis of Bayes risks for multi-class random data. This framework helps us derive a generalized version of most, if not all, of these known results, and also allows us to derive several new results. In this framework, we first investigate how the network behaves as a conditional risk estimator. We next investigate the feature extraction ability of neural networks, specifically the ability to perform nonlinear discriminant analysis.

We first present two equivalent cost functions, both based upon the MSE criterion. The first cost function called, the *conditional risk criterion*, is the MSE between the network output and the conditional risk. The second cost function, called the *Bayes cost criterion*, is the MSE between the class conditional outputs and the *Bayes cost for misclassification*. Both these cost functions are minimized to obtain the optimum networks for classification and feature extraction.

In the context of classification by the neural networks, we demonstrate that the MSE criterion for multilayer perceptrons, is equivalent to the above cost functions when the number of training samples becomes arbitrarily large. It is well-known [5,6,8,9,10] that the multilayer perceptrons performing a one of m (one output unity, all others zero) classification, approximate the Bayesian a posteriori probabilities at the network output. Based upon the Bayes risk analysis, we demonstrate a more general result, that is: if the network classification is weighted by the Bayes cost for misclassification, then the network approximates the conditional risk functions at the output. For the commonly used *uniform cost*, we obtain the previously known result.

We next extend the two cost functions, discussed above, to demonstrate the feature extraction ability of the neural network. In particular, we explore the network's ability to perform nonlinear discriminant analysis (NDA). Based on the Bayes risk interpretation of the network operation, we generalize the use of the network to perform NDA. A number of previously known [4,11] results are now easily derived and explained by the new analyses. We demonstrate that if the final layer is linear with thresholds at the output nodes, then minimizing the MSE at the network output, also maximizes a generalized criterion for NDA. Thus, the network not only estimates the

0-7803-3210-5/96 $4.00©1996 IEEE

conditional risks at the output, but also performs NDA at the final hidden layer.

Since the network approximates, under some conditions of Bayes costs, the a posteriori probabilities at the network outputs, these probabilities should sum to one. A common method [5,8] to achieve this results is by adding a final layer to normalize the outputs. We show that a network with linear output layer not only sums the outputs to one, but also performs NDA at the space spanned by the final hidden units. Furthermore, we obtain a tight upper bound of the error probability of the Bayes decision rule (i.e., Bayes error e^*). In addition, we demonstrate that under some conditions, we can also obtain a tight lower bound of the Bayes error e^*. Furthermore, we prove that the MSE criterion under some conditions is equal to the average conditional quadratic entropy criterion.

2. The Statistical Model

Let there be a finite m-set of pattern classes $\omega=\{\omega_1,...,\omega_m\}$, with a priori probability P_i, $i=1,...,m$, conditional probability density $p(\mathbf{x}/\omega_i)$, $i=1,...,m$, and a posteriori probability $p(\omega_i/\mathbf{x})$, $i=1,...,m$; where $\mathbf{x}\in\mathbb{R}^d$ is a pattern vector whose mixture distribution is given by $p(\mathbf{x})$. In traditional Bayesian risk analysis, we consider the cost of misclassification. Let $D=[d_{ij}]$ for $i,j=1,...,m$, be the matrix of costs of *misclassification* with d_{ji}, the cost of assigning to class ω_i a pattern which actually belongs to ω_j. We shall assume $0\le d_{ij}\le 1$, which is discussed later. However, for the network and the results that follow, it is convenient to define the *modified costs* $c_{ij}=1-d_{ij}$. Let $C=[c_{ij}]$ for $i,j=1,...,m$, be the matrix of modified costs.

Let $\mathbf{e}_i$ be the i^{th} standard basis vector with one in the i^{th} place and zero elsewhere. We define $\mathbf{p}(\omega/\mathbf{x})$ as the m-dimensional vector whose i^{th} component is the a posteriori probability of class ω_i given the pattern $\mathbf{x}$ i.e., $\mathbf{p}(\omega/\mathbf{x}) = [p(\omega_1/\mathbf{x}) \ldots p(\omega_m/\mathbf{x})]^T$. The *conditional risk* $\rho_i(\mathbf{x})$ of deciding in favor of class ω_i for a given pattern $\mathbf{x}$ is given by:

$$\rho_i(\mathbf{x}) = \sum_{j=1}^{m} d_{ji}\, p(\omega_i / \mathbf{x}) = \mathbf{e}_i^T D\mathbf{p}(\omega / \mathbf{x}) \quad \text{for } i=1,...,m.$$

We define a *modified conditional risk* $\sigma_i(\mathbf{x})$ in terms of the modified cost matrix C as follows:

$$\sigma_i(\mathbf{x}) = \sum_{j=1}^{m} c_{ji}\, p(\omega_i / \mathbf{x}) = \mathbf{e}_i^T C\mathbf{p}(\omega / \mathbf{x}) = 1 - \rho_i(\mathbf{x}) \quad \text{for } i=1,...,m. \tag{1}$$

The *modified conditional risk vector* $\sigma(\mathbf{x})$ is defined as follows:

$$\sigma(\mathbf{x}) = \left[\sigma_1(\mathbf{x})\cdots\sigma_m(\mathbf{x})\right]^T = C^T\mathbf{p}(\omega / \mathbf{x}).$$

From (1), the *Bayes conditional risk* $r(\mathbf{x})$ is given by:

$$r(\mathbf{x}) = \min_i \rho_i(\mathbf{x}) = 1 - \max_i \sigma_i(\mathbf{x}) \quad \text{for } i=1,...,m. \tag{2}$$

One cost matrix of special interest to us is the so-called *equal or uniform* cost matrix $C_0=I$, where I is the mXm identity matrix. For this cost matrix,

$$\sigma_i(\mathbf{x}) = p(\omega_i / \mathbf{x}), \text{ and } r(\mathbf{x}) = 1 - \max_i p(\omega_i / \mathbf{x}). \tag{3}$$

Let e^* denotes the *error probability of the Bayes decision rule*, also called the Bayes error. From (3), we have:

$$e^* = 1 - E\left\{\max_i p(\omega_i / \mathbf{x})\right\}. \tag{4}$$

Define $b(\omega/\mathbf{x})$ as the simple *Bayesian distance* [1,2]:

$$b(\omega / \mathbf{x}) = E\left\{\left\|p(\omega / \mathbf{x})\right\|^2\right\}. \tag{5}$$

We now give the constraints on costs c_{ij}, which are: $0\le c_{ij}\le 1$ and $\sum_{i=1}^{m} c_{ij} \ge 1$. For an m-class problem, it is well-known [1,2] that the Bayes error e^* is within $[0,(m-1)/m]$. In order to obtain easily interpretable numerical values for the average Bayes risk $E\{r(\mathbf{x})\}$, it is desirable [1] to keep this value within the same range. The sufficient conditions for $0\le E\{r(\mathbf{x})\}\le 1$ are: $c_{ij}\ge 0$ and $\sum_{i=1}^{m} c_{ij} \ge 1$, for $i,j=1,...,m$. In addition, we impose the constraint: $c_{ij}\le 1$, for $i,j=1,...,m$, which greatly simplifies our analyses. Devijver [1] has shown that any arbitrary cost matrix can be made to satisfy these conditions by a simple procedure.

Note that in this study, for the network and the results that follow, for convenience, we shall refer to $\sigma_i(\mathbf{x})$ as the conditional risks, instead of the traditional $\rho_i(\mathbf{x})$.

3. Classification by Multilayer Perceptrons

Let $\mathbf{f}(\mathbf{x},\mathbf{w})$ be the output of the multilayer perceptrons, where $\mathbf{w}$ is the weight vector, and $\mathbf{x}$ is the pattern to be classified. Define the criterion $J_1(\mathbf{w};C)$ as below, which is minimized to find the weight vector $\mathbf{w}$:

$$J_1(\mathbf{w};C) = E\left\{\left\|\mathbf{f}(\mathbf{x},\mathbf{w}) - \boldsymbol{\sigma}(\mathbf{x})\right\|^2\right\}. \tag{6}$$

Clearly, J_1 minimizes the distance of the network output from the conditional risk, which is an intuitively appealing criterion. However, since the exact expression of the risk vector $\boldsymbol{\sigma}(\mathbf{x})$ is generally unknown, this formulation of the criterion is not very realistic in practical applications. Nevertheless, the analysis of J_1 will provide us some results that are useful. We call this criterion the *conditional risk criterion*.

The next criterion $J_2(\mathbf{w};C)$ is more useful for applications and is defined as:

$$J_2(\mathbf{w};C) = \sum_{i=1}^{m} P_i E_i\left\{\left\|\mathbf{f}(\mathbf{x},\mathbf{w}) - C^T\mathbf{e}_i\right\|^2\right\}, \tag{7}$$

where $E_i\{\cdot\}$ denotes the conditional expectation for class ω_i. We call this criterion the *Bayes cost criterion*.

We next consider the MSE criterion for multilayer perceptrons that perform a one of m (i.e., one output unity, all others zero) classification. We shall generalize this traditional MSE criterion to include the Bayes costs for classification by defining it as follows, for finite samples:

$$\hat{J}_{net}(\mathbf{w};C) = \frac{1}{n}\sum_{i=1}^{m}\sum_{\mathbf{x}\in\omega_i}\left\|\mathbf{f}(\mathbf{x},\mathbf{w}) - C^T\mathbf{e}_i\right\|^2, \tag{8}$$

where n is the total number of training samples. Note that for the uniform cost matrix $C_0=I$, for a network performing a one of m classification, $\hat{J}_{net}(\mathbf{w};C_0)$ is equivalent to the traditional finite sample MSE at the network output.

Lemma 1. *The mean square error criterion of the multilayer perceptrons (see (8)) converges with probability one to the Bayes cost criterion $J_2(\mathbf{w};C)$ as the number of training samples n_i from each class becomes arbitrarily large.*

Proof. From (8), we obtain:

$$\hat{J}_{net}(\mathbf{w};C) = \sum_{i=1}^{m}\frac{n_i}{n}\frac{1}{n_i}\sum_{\mathbf{x}\in\omega_i}\left\|\mathbf{f}(\mathbf{x},\mathbf{w}) - C^T\mathbf{e}_i\right\|^2.$$

Using the strong law of large numbers, for independent and identically distributed random variables, as $n_i \to \infty$ for $i=1,\ldots,m$, $\hat{J}_{net}(\mathbf{w};C) \to J_2(\mathbf{w};C)$ with probability one. The above analysis also holds for the more general weighted MSE case. ■

Lemma 2. *The conditional risk criterion J_1 is equivalent to the Bayes cost criterion J_2 for solving the network weight vector $\mathbf{w}$.*

Proof. Notice that:

$$\begin{aligned}
J_2(\mathbf{w};C) &= \sum_{i=1}^{m} P_i E_i\left\{\left\|\mathbf{f}(\mathbf{x},\mathbf{w}) - C^T\mathbf{e}_i\right\|^2\right\} \\
&= \sum_{i=1}^{m} P_i \int\left(\left\|\mathbf{f}(\mathbf{x},\mathbf{w})\right\|^2 + \left\|C^T\mathbf{e}_i\right\|^2 - 2\mathbf{f}^T C^T\mathbf{e}_i\right)p(\mathbf{x}/\omega_i)d\mathbf{x} \\
&= \int\left\|\mathbf{f}(\mathbf{x},\mathbf{w})\right\|^2\sum_{i=1}^{m}P_i p(\mathbf{x}/\omega_i)d\mathbf{x} + \int\sum_{i=1}^{m}\left\|C^T\mathbf{e}_i\right\|^2 P_i p(\mathbf{x}/\omega_i)d\mathbf{x} - 2\int\mathbf{f}^T C^T\sum_{i=1}^{m}\mathbf{e}_i P_i p(\mathbf{x}/\omega_i)d\mathbf{x}.
\end{aligned}$$

Note that $\sum_{i=1}^{m} P_i p(\mathbf{x}/\omega_i) = p(\mathbf{x})$, and $P_i p(\mathbf{x}/\omega_i) = p(\omega_i/\mathbf{x})p(\mathbf{x})$. Therefore,

$$\begin{aligned}
J_2(\mathbf{w};C) &= \int\left\|\mathbf{f}(\mathbf{x},\mathbf{w})\right\|^2 p(\mathbf{x})d\mathbf{x} + \int\left\|\boldsymbol{\sigma}(\mathbf{x})\right\|^2 p(\mathbf{x})d\mathbf{x} - 2\int\mathbf{f}^T\boldsymbol{\sigma}(\mathbf{x})p(\mathbf{x})d\mathbf{x} + \int\left(\sum_{i=1}^{m}\left\|C^T\mathbf{e}_i\right\|^2 p(\omega_i/\mathbf{x}) - \left\|\boldsymbol{\sigma}(\mathbf{x})\right\|^2\right)p(\mathbf{x})d\mathbf{x} \\
&= \int\left\|\mathbf{f}(\mathbf{x},\mathbf{w}) - \boldsymbol{\sigma}(\mathbf{x})\right\|^2 p(\mathbf{x})d\mathbf{x} + \int\left(\sum_{i=1}^{m}\left\|C^T\mathbf{e}_i\right\|^2 p(\omega_i/\mathbf{x}) - \left\|\boldsymbol{\sigma}(\mathbf{x})\right\|^2\right)p(\mathbf{x})d\mathbf{x}.
\end{aligned}$$

The first term in the above expression is J_1, and the second term is independent of $\mathbf{w}$. Thus, minimizing J_2 with respect to $\mathbf{w}$ also minimizes J_1. ■

Combining Lemmas 1 and 2, we see that if we minimize the network MSE $\hat{J}_{net}(\mathbf{w};C)$, then the network output approximates the conditional risk vector $\boldsymbol{\sigma}(\mathbf{x})$ for large samples. This is a generalization of the previously known result that for networks performing a one of m classification, the output approximates the Bayes a posteriori probabilities. The one of m classification is same as the uniform cost matrix $C_0=I$, which gives us the previously known result from these Lemmas. The above Lemmas can be easily extended to a more general weighted MSE case.

4. Feature Extraction by Multilayer Perceptrons

In this section, we shall analyze the feature extraction abilities of the multilayer perceptrons, specifically the network's ability to perform nonlinear discriminant analysis (NDA). It is well-known [3] that the a posteriori probabilities $p(\omega_i/\mathbf{x})$, $i=1,...,m$, are sufficient statistic and carries all information for classification in the Bayes sense. Since the a posteriori probabilities sum to one, only $m-1$ of these functions are linearly independent. Thus, $\{p(\omega_i/\mathbf{x}),\ i=1,...,m-1\}$ is the ideal feature set for classification. The Bayes classifier in this feature space is a piecewise bisector classifier [3] which is its simplest form.

It is now clear that the optimum feature set are those that can be classified by a linear classifier. If the network performs a NDA, then the data in this feature space can be classified by a linear classifier, which can be implemented by a single layer that is linear. We shall, therefore, place a final linear layer in our network design.

Let $\mathbf{y}(\mathbf{w},\mathbf{x})=[y_1(\mathbf{w},\mathbf{x}),\ ...,\ y_m(\mathbf{w},\mathbf{x})]^T$ be the output at the final hidden layer. Every $y_i(\mathbf{w},\mathbf{x})$ is assumed to have a second moment with respect to $p(\mathbf{x})$. We need to approximate the conditional risks $\sigma_i(\mathbf{x})$ for $i=1,...,m$, by a linear transform of the form:

$$\mathbf{f}(\mathbf{w},\mathbf{x}) = A^T\mathbf{y}(\mathbf{w},\mathbf{x}) + \mathbf{t}, \tag{9}$$

where $\mathbf{f}$ is the final output of the network, A is the weight matrix for the final linear layer, and $\mathbf{t}$ is the threshold vector at the final nodes. We now look into the two criteria discussed above to obtain the network weights $\mathbf{w}$ and A.

Applications of the MSE design criterion consists of finding the weight matrix A and threshold $\mathbf{t}$, for the final layer which minimizes:

$$J_1(\mathbf{w},A,\mathbf{t};C) = E\left\{\left\|A^T\mathbf{y}(\mathbf{w},\mathbf{x}) + \mathbf{t} - \boldsymbol{\sigma}(\mathbf{x})\right\|^2\right\}. \tag{10}$$

It is known that $J_1(\mathbf{w},A,\mathbf{t};C)$ is a convex function of A and $\mathbf{t}$, and therefore, has a unique minimum. Setting the gradient of (10) with respect to $\mathbf{t}$ to zero, we obtain:

$$\mathbf{t} = E\{\boldsymbol{\sigma}(\mathbf{x})\} - A^T E\{\mathbf{y}(\mathbf{w},\mathbf{x})\} = \bar{\boldsymbol{\sigma}} - A^T\bar{\mathbf{y}}, \tag{11}$$

where $\bar{\boldsymbol{\sigma}}$ and $\bar{\mathbf{y}}$ are the expectations of $\boldsymbol{\sigma}(\mathbf{x})$ and $\mathbf{y}(\mathbf{w},\mathbf{x})$ respectively with respect to $p(\mathbf{x})$. Replacing the above estimate of $\mathbf{t}$ in (10), and setting the gradient of (10) with respect to A to zero, we obtain:

$$E\left\{(\mathbf{y}-\bar{\mathbf{y}})(\mathbf{y}-\bar{\mathbf{y}})^T\right\}A = E\left\{(\mathbf{y}-\bar{\mathbf{y}})(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})^T\right\} = E\left\{(\mathbf{y}-\bar{\mathbf{y}})\mathbf{p}(\boldsymbol{\omega}/\mathbf{x})^T\right\}C. \tag{12}$$

We denote: $E\{(\mathbf{y}-\bar{\mathbf{y}})(\mathbf{y}-\bar{\mathbf{y}})^T\} = S_m$, the mixture covariance of $\mathbf{y}$; and $E\{(\mathbf{y}-\bar{\mathbf{y}})\mathbf{p}(\boldsymbol{\omega}/\mathbf{x})^T\} = M$, a matrix of the class conditional means of $\mathbf{y}$. Clearly, for nonsingular matrix S_m:

$$A = S_m^{-1}MC. \tag{13}$$

We now compute $J_1(\mathbf{w};C)$ by placing A and $\mathbf{t}$ as given in (13) and (11) respectively, into (10). This yields:

$$J_1(\mathbf{w};C) = E\left\{\left\|C^TM^TS_m^{-1}(\mathbf{y}-\bar{\mathbf{y}}) - (\boldsymbol{\sigma}(\mathbf{x})-\bar{\boldsymbol{\sigma}})\right\|^2\right\}. \tag{14}$$

After simplifying (14), we obtain:

$$J_1(\mathbf{w};C) = E\left\{\left\|\boldsymbol{\sigma}(\mathbf{x})-\bar{\boldsymbol{\sigma}}\right\|^2\right\} - \mathrm{tr}\left\{C^TM^TS_m^{-1}MC\right\}. \tag{15}$$

It is now clear from (15), that minimizing $J_1(\mathbf{w};C)$ with respect to $\mathbf{w}$ is same as maximizing $\mathrm{tr}\{C^TM^TS_m^{-1}MC\}$. We note that $\mathrm{tr}\{C^TM^TS_m^{-1}MC\}$ is a criterion for NDA.

We next investigate the MSE due to the Bayes cost criterion J_2 with A and $\mathbf{t}$ given in (13) and (11) respectively. We obtain from (7):

$$J_2(\mathbf{w};C) = \sum_{i=1}^{m} P_i \left\| C^T(\mathbf{e}_i - \overline{\mathbf{p}}) \right\|^2 - \mathrm{tr}\left\{ C^T M^T S_m^{-1} M C \right\}, \text{ where } \overline{\mathbf{p}} = E\{\mathbf{p}(\omega / \mathbf{x})\}. \quad (16)$$

Clearly, minimizing $J_2(\mathbf{w};C)$ is same as maximizing $\mathrm{tr}\{C^T M^T S_m^{-1} M C\}$. Since Lemma 1 shows that the MSE of the network output is equivalent to J_2 for large samples, we conclude that minimizing the MSE at the network output also maximizes a criterion for NDA at the final hidden layer. We summarize this result in the following Lemma.

Lemma 3. *If the final layer of the network is linear with a threshold at the output nodes, then minimizing the MSE at the network output is equivalent to maximizing a generalized criterion for NDA at the final hidden layer.* ■

Note that this result is a generalization of previous results obtained by Webb and Lowe [11], and by Gallinari *et al.* [4]. Our criterion also includes the Bayes cost matrix C. Further note that $\mathrm{tr}\{C^T M^T S_m^{-1} M C\} = \mathrm{tr}\{S_m^{-1} M C C^T M^T\}$. Thus, maximizing this criterion is also equivalent to maximizing the usual discriminant analysis criterion $\mathrm{tr}(S_m^{-1} \tilde{S}_b)$, where the between class scatter matrix $\tilde{S}_b$ is redefined as $MCC^T M^T$. The traditional between class scatter matrix S_b can be obtained from $\tilde{S}_b$ for $C = \mathrm{diag}(P_1^{-1/2}, \ldots, P_m^{-1/2})$. This analysis explains the so-called "weighted between class scatter matrix" obtained by Webb and Lowe [11], which is the same as $\tilde{S}_b$ above, for the uniform cost matrix $C_0=I$.

The Case of Uniform Costs of Classification i.e. $C_0=I$

Since J_2 is equivalent to the MSE at the network output, we shall investigate this criterion under the uniform cost matrix $C_0=I$, which is the most common formulation for $\hat{J}_{net}$. We shall also study this criterion with respect to the Bayes error e^*. This reveals the following results. From (15), we obtain:

$$J_1(\mathbf{w};C_0) = E\left\{ \left\| \mathbf{p}(\omega / \mathbf{x}) - \overline{\mathbf{p}} \right\|^2 \right\} - \mathrm{tr}\left\{ M^T S_m^{-1} M \right\} = b(\omega / \mathbf{x}) - \left\| \overline{\mathbf{p}} \right\|^2 - \mathrm{tr}\left\{ M^T S_m^{-1} M \right\}. \quad (17)$$

From (16), we obtain:

$$J_2(\mathbf{w};C_0) = \sum_{i=1}^{m} P_i \left\| \mathbf{e}_i - \overline{\mathbf{p}} \right\|^2 - \mathrm{tr}\left\{ M^T S_m^{-1} M \right\} = 1 - \left\| \overline{\mathbf{p}} \right\|^2 - \mathrm{tr}\left\{ M^T S_m^{-1} M \right\}. \quad (18)$$

Since $J_1(\mathbf{w};C_0) \geq 0$, we have from (17):

$$b(\omega / \mathbf{x}) \geq \left\| \overline{\mathbf{p}} \right\|^2 + \mathrm{tr}\{M^T S_m^{-1} M\}.$$

From (18), we obtain: $J_2(\mathbf{w};C_0) \geq 1 - b(\omega/\mathbf{x})$. Devijver [1,2] has shown that: $1 - b(\omega/\mathbf{x}) \geq e^*$, where e^* is the error probability of the Bayes decision rule. Therefore,

$$J_2(\mathbf{w};C_0) \geq 1 - b(\omega / \mathbf{x}) \geq e^*. \quad (19)$$

Thus, the commonly used MSE criterion for the network also produces an upper bound of the Bayes error e^*.

5. Normalizing the Network Output for Uniform Costs

As discussed above, under the uniform cost matrix $C_0=I$, the network outputs approximate the a posteriori probabilities. For these to be true probabilities, the outputs should sum to one. This will require that the optimization problem be changed to one of constrained optimization. A common method of achieving this is by adding a layer to the existing network to perform the normalization. In this section, we shall present a novel solution to this problem.

We shall continue with our network design that has a final linear layer with weight matrix A. The following Lemma summarizes our result.

Lemma 4. *For a network with a final linear layer, the outputs sum to one, under the uniform cost matrix $C_0=I$.*

Proof. Let us assume that the output vector $\mathbf{y}$ in the final hidden layer has dimension $p \leq m$. As before, the network has m outputs $\mathbf{f} = [f_1 \ldots f_m]^T$. We shall assume that the mixture covariance S_m of $\mathbf{y}$ is nonsingular. Then from (13), we have $A = S_m^{-1} M$. Define a vector $\mathbf{1}=[1\ldots1]^T$. Note that $A\mathbf{1} = S_m^{-1} M\mathbf{1} = 0$, since $M\mathbf{1} = 0$. From (11), we have $\mathbf{t} = \overline{\mathbf{p}} - A^T \overline{\mathbf{y}}$, where $\overline{\mathbf{p}}$ is defined in (16). Notice that $\mathbf{t}^T\mathbf{1} = \overline{\mathbf{p}}^T\mathbf{1} - \mathbf{y}^T A\mathbf{1} = \overline{\mathbf{p}}^T\mathbf{1} = 1$. Therefore, $\mathbf{f}^T\mathbf{1} = (A^T\mathbf{y} + \mathbf{t})^T\mathbf{1} = \mathbf{y}^T A\mathbf{1} + \mathbf{t}^T\mathbf{1} = 1$. Thus, the network outputs sum to one. ■

Lemma 4 shows that if the network architecture is modified such that the final layer is linear, then the outputs of the network sum to one. Under this condition, one case of particular theoretical interest is when $\mathbf{p}(\omega/\mathbf{x}) = A^T\mathbf{y}(\mathbf{w},\mathbf{x}) + \mathbf{t}$. In this situation, $J_1(\mathbf{w};C_0) = 0$, and $b(\omega / \mathbf{x}) = \|\overline{\mathbf{p}}\|^2 + \mathrm{tr}\{M^T S_m^{-1} M\}$. From (18), we have:

$$J_2(\mathbf{w};C_0) = 1 - b(\omega / \mathbf{x}). \tag{20}$$

Note that:

$$1 - b(\omega / \mathbf{x}) = E\left\{\sum_{i=1}^{m} p(\omega_i / \mathbf{x})(1 - p(\omega_i / \mathbf{x}))\right\} = h(\omega / \mathbf{x}).$$

In this equation, $h(\omega/\mathbf{x})$ is commonly known as the *conditional quadratic entropy* [1]. We, therefore, have:

$$J_2(\mathbf{w};C_0) = h(\omega / \mathbf{x}) = E\left\{\sum_{i=1}^{m} p(\omega_i / \mathbf{x})(1 - p(\omega_i / \mathbf{x}))\right\}. \tag{21}$$

From (21), we obtain: $b(\omega / \mathbf{x}) = 1 - J_2(\mathbf{w};C_0)$, which gives us strict upper and lower bounds of the Bayes error e^* as shown by Devijver [2], and given below:

$$\frac{m-1}{m}\left(1 - \sqrt{1 - \frac{mJ_2(\mathbf{w};C_0)}{m-1}}\right) \le e^* \le 1 - b(\omega / \mathbf{x}). \tag{22}$$

In summary, if the network outputs are the a posteriori probabilities, then the MSE at the network output $J_2(\mathbf{w};C_0)$ is equal to the conditional quadratic entropy $h(\omega/\mathbf{x})$, and gives us efficient bounds of the Bayes error e^*.

References

[1] P.A.Devijver, "Relationships between Statistical risks and the least-mean-square-error design criterion in Pattern Recognition", in *Proceedings First International Conference on Pattern Recognition*, Washington D.C., pp. 139-148, 1973.

[2] P.A.Devijver, "On a New Class of Bounds on Bayes Risk in Multihypothesis Pattern Recognition", *IEEE Transactions on Computers*, Vol. C-23, No. 1, pp. 70-80, 1974.

[3] K.Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd Ed., New York: Academic Press, 1990.

[4] P.Gallinari, S.Thiria, F.Badran, F.Fogelman-Soulie, "On the Relations Between Discriminant Analysis and Multilayer Perceptrons", *Neural Networks*, Vol. 4, pp. 349-360, 1991.

[5] H.Gish, "A probabilistic approach to the understanding and training of neural network classifiers", in *Proceedings IEEE Conference on Acoustics Speech and Signal Processing*, pp. 1361-1364, 1990.

[6] S.Miyake and F.Kanaya, "A Neural Network Approach to a Bayesian Statistical Decision Problem", *IEEE Transactions on Neural Networks*, Vol. 2, No. 5, pp. 538-540, 1991.

[7] J.D.Patterson, T.J.Wagner, B.F.Womack, "A Mean-Square Performance Criterion for Adaptive Pattern Classification Systems, *IEEE Transactions on Automatic Control*, Vol. AC-12, pp. 195-197, 1967.

[8] M.D.Richard and R.P.Lippmann, "Neural Network Classifiers Estimate Bayesian a posteriori Probabilities", *Neural Computation*, Vol. 3, pp. 461-483, 1991.

[9] D.W.Ruck, S.K.Rogers, M.Kabrisky, M.E.Oxley, and B.W.Suter, "The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function", *IEEE Transactions on Neural Networks*, Vol.1, No.4, pp.296-298, 1990.

[10] E.A.Wan, "Neural Network Classification: A Bayesian Interpretation", *IEEE Transactions. on Neural Networks*, Vol. 1, No. 4, pp. 303-305, 1990.

[11] A.R.Webb and D.Lowe, "The Optimised Internal Representation of Multilayer Classifier Networks Performs Nonlinear Discriminant Analysis", *Neural Networks*, Vol. 3, pp. 367-375, 1990.