A PARZEN WINDOW BASED DERIVATION OF MINIMUM CLASSIFICATION ERROR FROM THE THEORETICAL BAYES CLASSIFICATION RISK

Erik McDermott & Shigeru Katagiri

NTT Communication Science Laboratories, NTT Corporation
Hikari-dai 2-4, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
mcd@cslab.kecl.ntt.co.jp

ABSTRACT

This article shows that the Minimum Classification Error (MCE) criterion function commonly used for discriminative design of speech recognition systems is equivalent to a Parzen window based estimate of the theoretical Bayes classification risk. In this analysis, each training token is mapped to the center of a Parzen kernel in the domain of a suitably defined random variable. The kernels are summed to produce a density estimate; this estimate in turn can easily be integrated over the domain of incorrect classifications, yielding the risk estimate. The expression of risk for each kernel can be seen to correspond directly to the usual MCE loss function. The resulting risk estimate can be minimized by suitable adaptation of the recognition system parameters that determine the mapping from training token to kernel center. This analysis provides a novel link between the MCE empirical cost measured on a finite training set and the theoretical Bayes classification risk.

1. INTRODUCTION

The Minimum Classification Error (MCE) framework is an approach to discriminative training for pattern classification that explicitly incorporates classification performance into the training criterion. Given discriminant functions for each category, MCE defines a loss function that is a smoothed approximation of the recognition error rate and uses this function as the criterion function for optimization [3][5][6][7]. Through minimization of this criterion function, MCE is aimed directly at minimizing classification error rather than at learning the true data probability distributions, the target of Maximum Likelihood Estimation (MLE) via Baum-Welch or Viterbi training. MCE has been used successfully to train Hidden Markov Models for speech recognition tasks [1][7][8].

Here, a new theoretical perspective on MCE is presented. This addresses the nature of the smoothness of the MCE loss function, as well as the relationship between minimization of an overall MCE loss summed over a finite set of training data and minimization of the theoretical classification risk measured over the continuous densities underlying the classification problem. We show that the continuous, 0-1 MCE loss function can be derived from an estimate of the theoretical Bayes classification risk, using Parzen estimation of the density of a suitably defined variable, the misclassification measure. In this analysis, the specific kernel type used for Parzen estimation leads to a specific type of MCE loss function, and vice versa; the width of the Parzen kernel directly corresponds to the steepness of the MCE loss function, and vice versa. Minimization of the MCE loss function is seen to correspond to the minimization of a Parzen estimate of the theoretical classification risk.

2. THE MINIMUM CLASSIFICATION ERROR FRAMEWORK

The MCE framework has been described in several publications [3][6][7]. For each training token, MCE uses a three-step definition, mapping a training pattern token \( x \) and the system parameters \( \Lambda \) (e.g., all the HMM means and covariances) to a 0-1 loss function reflecting classification error. The pattern \( x \) could be a single pattern vector or a sequence of, e.g., speech-derived feature vectors, \( x = x_1^T, \ldots, x^T \). The formalism assumes that discriminant functions \( g_j(x, \Lambda) \) can be defined for each string category \( C_j \), and uses a misclassification measure \( d_k(x, \Lambda) \) to compare the match between the training token to the correct category \( C_k \) with the match to the best incorrect categories.

The loss function is typically a sigmoid,

\[
\ell(d_k(x, \Lambda)) = \frac{1}{1 + e^{-\alpha d_k(x, \Lambda)}}, \tag{1}
\]

but can be any of a variety of continuous 0-1 functions. The total loss \( L(X, \Lambda) \) is the local loss summed over the \( M \) categories and \( N_k \) tokens in each category \( C_k \) making up the training data \( X \):

\[
L(X, \Lambda) = \frac{1}{N} \sum_k \sum_{i=1}^{N_k} \ell(d_k(x_{ik}, \Lambda)), \tag{2}
\]
where \( x_{ik} \) denotes the \( i \)th training token in category \( C_k \).

The overall loss function can be minimized using several different approaches [7].

### 3. A NOVEL ANALYSIS OF THE SMOOTHNESS OF THE MCE LOSS FUNCTION

#### 3.1. Bayes risk & Bayes decision rule

Essentially, we show that the smooth MCE loss function can be derived from a particular approach to modeling the true theoretical error, the Bayes risk [2]:

\[
R = \int_{X} R(\alpha(x)|x)p(x)dx. \tag{3}
\]

Here \( \alpha(x) \) represents the classification decision (the choice of one out of \( M \) categories) made for the input pattern \( x \).

\( R(\alpha(x)|x) \) represents the risk involved in making specific classifications:

\[
R(\alpha_i|x) = \sum_{j=1}^{M} \lambda(\alpha_i|C_j)P(C_j|x). \tag{4}
\]

Here \( \lambda(\alpha_i|C_j) \) denotes the cost of mis-classifying a member of category \( C_j \) as category \( C_i \). Typically this cost is 1 whenever \( i \) and \( j \) are different. It is easy to show [2] that the overall risk (Equ. (3)) is minimized by the Bayes decision rule:

\[
declare C_j \text{ if } P(C_i|x) > P(C_j|x) \text{ for all } j \neq i. \tag{5}
\]

Clearly, this requires knowledge of the posterior probabilities \( P(C_j|x) \).

#### 3.2. Linking the Bayes risk to discriminant functions

In an actual classifier system, we do not know the true posterior probabilities \( P(C_j|x) \), and hence cannot directly implement the optimal decision rule. In the following, we start with the above definitions of error and arrive at the MCE approach to classifier design, in a manner that is different from previous expositions of MCE and that links the empirical error measured on a given, finite training set, and the theoretical error given above.

The first step is to link Equ. (3) to an actual pattern recognition system. Using the expression for the risk incurred by an individual classification decision, given by Equ. (4), we can rewrite Equ. (3) as

\[
R = \sum_{j=1}^{M} \int_{X} \lambda(\alpha(x)|C_j)p(C_j,x)dx. \tag{6}
\]

The link between this expression of error and an actual classification system is embodied by the decisions \( \alpha(x) \) that the system takes. In a system where the classification decision is to choose the category with the largest discriminant function value \( g_j(x) \), we can express the categorical cost \( \lambda(\alpha(x)|C_j) \) as

\[
\lambda(\alpha(x)|C_j) = 1((-g_j(x) + \max_{i \neq j} g_i(x)) > 0). \tag{7}
\]

If \( C_j \) has the largest discriminant function value, the system will choose \( C_j \) and the indicator function \( 1((-g_j(x) + \max_{i \neq j} g_i(x)) > 0) \) is 0; otherwise the system does not choose \( C_j \) and the indicator function is 1. Note that the discriminant functions depend on the modifiable parameters, \( \Lambda \); where space allows we write them as \( g_j(x, \Lambda) \). We can now rewrite the overall error in terms of the indicator function just introduced:

\[
R = \sum_{j=1}^{M} \int_{X} 1((-g_j(x) + \max_{i \neq j} g_i(x)) > 0)p(C_j,x)dx. \tag{8}
\]

#### 3.3. Defining Bayes risk in a new domain

The misclassification measure used in the following is:

\[
m_j = d_j(x) = -g_j(x) + \max_{i \neq j} g_i(x). \tag{9}
\]

As with the discriminant functions, the dependence of the misclassification measure on the tunable system parameters \( \Lambda \) could be emphasized by writing \( d_j(x, \Lambda) \). The overall error can then be written as

\[
R = \sum_{j=1}^{M} \int_{X} 1(d_j(x) > 0)p(C_j,x)dx. \tag{10}
\]

In turn, this is equivalent to integrating over a reduced part of the pattern space \( X' \):

\[
R = \sum_{j=1}^{M} \int_{X_j} p(C_j,x)dx, \tag{11}
\]

where \( X_j \) is given by

\[
X_j = \{ x \in X \mid d_j(x) > 0 \}. \tag{12}
\]

If \( X \) and \( g_1(X), ..., g_M(X) \) are continuous random variables, \( M_j = d_j(X) \) is also a continuous random variable. From the cumulative distribution technique for finding the distributions of functions of random variables, we can express the integral for each category \( C_j \) over the reduced space \( X_j \) with an integral over the positive domain of the misclassification measure \( m_j \):

\[
\int_{X_j} p(C_j,x)dx = P[d_j(x) > 0, C_j] = \int_0^{\infty} p(C_j, m_j)dm_j. \tag{13}
\]
We can now express the overall error expressed in Equ. (11) as:
\[ R = \sum_{j=1}^{M} P(C_j) \int_{0}^{\infty} p(m_j|C_j) \, dm_j. \] (14)
This is a new expression for the Bayes risk that is equivalent to the original expression given in Equ. (3).

A new approach to pattern classifier design is suggested by Equ. (14). We can try to express the density \( p(m_j|C_j) \) in a way that reflects the dependence of \( m_j \) on both \( x \) and the system parameters \( \mathbf{A} \) that go into the definition of the discriminant functions \( q_i() \), and then adjust the system parameters so as to minimize an estimate of Equ. (14). The following sections outline this new approach.

### 3.4. Parzen estimate of risk

In order to explicitly relate the empirical risk to the theoretical cost, we use the following Parzen estimate of the density \( p(m_j|C_j) \):
\[ p_{N_j}(m_j|C_j) = \frac{1}{N_j} \sum_{i=1}^{N_j} \phi \left( \frac{m_j - d_j(x_{ij}, \mathbf{A})}{h} \right). \] (15)
Here \( \phi((m_j - d_j(x_{ij}, \mathbf{A}))/h) \) is a window or kernel centered on the data point \( d_j(x_{ij}, \mathbf{A}) \) and with width \( h \). With the estimate \( p_{N_j}(m_j|C_j) \) defined for any value of \( m_j \), we can now define an estimate of the true overall risk expressed in Equ. (14):
\[ R_N = \sum_{j=1}^{M} P(C_j) \int_{0}^{\infty} p_{N_j}(m_j|C_j) \, dm_j. \] (16)
Expanding this using the estimate \( p_{N_j}(m_j|C_j) \) given by Equ. (15) yields
\[ R_N = \sum_{j=1}^{M} P(C_j) \int_{0}^{\infty} \frac{1}{N_j} \sum_{i=1}^{N_j} \phi \left( \frac{m_j - d_j(x_{ij}, \mathbf{A})}{h} \right) \, dm_j. \] (17)
Rearranging, and using \( N_j/N \) (where \( N = \sum_{j=1}^{M} N_j \)) for \( P(C_j) \), gives:
\[ R_N = \frac{1}{N} \sum_{j=1}^{M} \sum_{i=1}^{N_j} \frac{1}{h} \int_{0}^{\infty} \phi \left( \frac{m_j - d_j(x_{ij}, \mathbf{A})}{h} \right) \, dm_j. \] (18)
We have thus arrived at a Parzen estimate of classification error, defined in terms of an integral over positive values of the misclassification measure. Parzen estimation theory [2] tells us that this estimate converges to the theoretical risk as the number of data points approaches infinity and the kernel width is reduced. Furthermore, we have some control over the value of \( R_N \) via the system parameters \( \mathbf{A} \) that are implicit in the definition of misclassification measure, and that therefore affect the specific values of the window anchor points \( d_j(x_{ij}, \mathbf{A}) \).

### 3.5. Parzen estimate of risk for specific kernels

In this section we show the utility of the Parzen estimate of risk (Equ. (18)) by considering simple but perfectly valid Parzen kernel types. For these kernels, we can easily find the closed form of the integral in Equ. (18).

#### 3.5.1. Uniform Parzen kernel leads to piece-wise linear MCE loss

Considering first the one-dimensional uniform kernel,\[
\phi(u) = \begin{cases} 
1 & \text{if } |u| \leq 1/2, \\
0 & \text{otherwise.} 
\end{cases}
\] (19)
Therefore, \( \phi((m_j - d_j(x_{ij}, \mathbf{A}))/h) \) will be one if \( m_j \) falls within the segment of length \( h \) centered at \( d_j(x_{ij}, \mathbf{A}) \), and zero otherwise.

We can now integrate Equ. (18) for this specific choice of kernel. Focusing first on a single data point \( x_{ij} \), and abbreviating \( d = d_j(x_{ij}, \mathbf{A}) \), we have\[
\frac{1}{h} \int_{0}^{\infty} \phi \left( \frac{m_j - d}{h} \right) \, dm_j = \begin{cases} 
0 & \text{if } d < h, \\
\frac{d}{h} & \text{if } \frac{h}{2} \leq d \leq \frac{h}{2}, \\
1 & \text{if } d > \frac{h}{2}.
\end{cases}
\] (20)
The integral over \( m_j \), therefore yields a piece-wise linear function that rises monotonically from 0 to 1. Writing this expression as \( \ell_1() \), we can express the Parzen estimate of classification loss (18) as
\[ R_N = \frac{1}{N} \sum_{j=1}^{M} \sum_{i=1}^{N_j} \ell_1(d_j(x_{ij}, \mathbf{A})). \] (21)
This is identical to the empirical MCE loss defined earlier in this article (Equ. (2)), assuming the piece-wise linear MCE loss used in several MCE-based studies [7]. As the training set grows larger, the kernel width \( h \) should be decreased as per Parzen estimation theory, resulting in an increasingly steep choice of the piece-wise linear function \( \ell_1() \). As the training set size approaches infinity, \( h \) approaches zero, and \( \ell_1() \) approaches the binary loss function.

#### 3.5.2. Sigmoidal MCE loss function corresponds to simple Parzen kernel

We can now go the other way: start with an MCE loss function and ask what Parzen kernel it corresponds to. The most frequently used MCE loss function is the sigmoid function
\[ p(x|\alpha) = \frac{1}{1 + e^{-\alpha x}}. \]
Considering the sigmoid function, we have the following: \( p(x|\alpha) \) decreases as \( \alpha \) increases. We can thus adjust the system parameters so as to minimize an estimate of Equ. (14). The following sections outline this new approach.
(1), it will clearly have the form of the derivative of Equ. (1). One can verify that using the kernel
\[ \phi(u) = \frac{e^u}{(1 + e^u)^2} \]  
(22)
in the integral in Equ. (18) yields back the sigmoid loss. As above, the kernel is centered on the given data points \( d_j(x_{ij}, \Lambda) \) and its width and height are controlled by a scalar \( h \). Again abbreviating \( d = d_j(x_{ij}, \Lambda) \), the integral for a single data point \( x_{ij} \) is
\[ \frac{1}{h} \int_0^\infty \phi\left(\frac{m_j - d}{h}\right)dm_j = \int_0^\infty \frac{1}{h} \frac{e^{\frac{m_j}{h}-d}}{(1 + e^{\frac{m_j}{h}-d})^2} dm_j, \]
which yields, with \( h = 1/\alpha \), the MCE sigmoid loss function Equ. (1).

The corresponding overall estimate of risk (Equ. (18)) then becomes identical to the empirical MCE loss assuming, this time, the sigmoid loss function that has also been used in previous MCE-based studies. Figure 1 illustrates the relation between the sigmoidal loss function and the corresponding Parzen kernel.

3.6. Minimizing the risk estimate

Strikingly, by using Parzen estimates of overall risk, and kernel types that allow us to integrate directly the general expression for risk estimate given by Equ. (18), we have arrived precisely at expressions of empirical MCE cost that we are already familiar with, and that can be optimized in exactly the same way as the MCE cost.

From the point of view of the Parzen analysis provided here, the task of any optimization procedure applied to the risk estimate (Equ. (18)) is to choose the system parameters \( \Lambda \) so as to position the kernel centers in such a way as to minimize the estimated risk. Whether the resulting minimization corresponds to the theoretically optimal Bayes risk rests on the effectiveness of the optimization procedure, and of course on the model structure.

4. SUMMARY

We have shown that typical MCE loss functions can be interpreted as the integral of a Parzen kernel of height \( h \) centered on \( d_j(x_{ij}, \Lambda) \):
\[ \ell(d_j(x_{ij}, \Lambda)) = \frac{1}{h} \int_0^\infty \phi\left(\frac{m_j - d_j(x_{ij}, \Lambda)}{h}\right)dm_j. \]  
(24)
The specific form of the kernel determines the form of the MCE loss function, and vice versa. This approach to risk estimation provides a direct link between empirical error and theoretical error, and furthermore results in expressions of loss that are differentiable and hence amenable to gradient-based optimization methods.

5. REFERENCES