

On a Relation Between Maximum-Likelihood Classification and Minimum-Cross-Entropy Classification

JOHN E. SHORE

*Computer Science and Systems Branch
Information Technology Division*



July 14, 1983



NAVAL RESEARCH LABORATORY
Washington, D.C.

Approved for public release; distribution unlimited.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NRL Report 8707	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ON A RELATION BETWEEN MAXIMUM-LIKELIHOOD CLASSIFICATION AND MINIMUM-CROSS-ENTROPY CLASSIFICATION		5. TYPE OF REPORT & PERIOD COVERED Interim report on a continuing NRL problem
		6. PERFORMING ORG. REPORT NUMBER 7590-037; JES:pkt
7. AUTHOR(s) John E. Shore		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Research Laboratory Washington, DC 20375		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS RR014-09-41 NRL Problem 75-0102-0-3
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217		12. REPORT DATE July 14, 1983
		13. NUMBER OF PAGES 7
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Information theory Estimation Maximum likelihood Spectrum analysis		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) The report considers maximum-likelihood (ML) and minimum-cross-entropy (MCE) classification of samples from an unknown probability density when the hypotheses comprise an exponential family. It is shown that ML and MCE lead to the same classification rule, and the result is illustrated in terms of a method for estimating covariance matrices recently developed by Burg, Luenberger, and Wenger. MCE classification applies to the general case in which it cannot be assumed that the samples were generated by one of the hypothesis densities. The common use of ML in this case is technically incorrect, but the equivalence of MCE and ML provides a theoretical justification.		

DD FORM 1473
1 JAN 73EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ON A RELATION BETWEEN MAXIMUM-LIKELIHOOD CLASSIFICATION AND MINIMUM-CROSS-ENTROPY CLASSIFICATION

INTRODUCTION

Maximum likelihood (ML) and related classification methods are often used to choose from a set of hypotheses based on known data. The theoretical justification for these methods depends on the assumption that one of the hypotheses is true, but they are used even when it is known that this assumption is false. This practice can be justified on the practical grounds that it works, but there is no compelling theoretical justification. In minimum-cross-entropy (MCE) classification, one classifies data in terms of estimated underlying probability densities using a nearest-neighbor rule and an information-theoretic distortion measure [1]. Speech coding by vector quantization [2,3] can be derived as a special case of MCE classification [1].

In this report I consider the relation between ML classification and MCE classification of samples from an unknown probability density when the hypotheses comprise an exponential family. I show that ML and MCE lead to the same classification rule, but that MCE applies in the general case when one cannot assume that one of the hypotheses is true and thereby provides a theoretical foundation for the technically incorrect use of ML. I illustrate the results in terms of a recently developed method of estimating covariance matrices [4].

STATEMENT OF THE CLASSIFICATION PROBLEM

Let $\{\hat{q}_s(\mathbf{x}):s\in\Lambda\}$ be a finite or infinite set of probability densities on some vector space. Let $q^\dagger(\mathbf{x})$ be the probability density for vector-valued samples from some unknown process, and let $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M$ be a sequence of M vector-valued samples from q^\dagger . Let $\{H_s:s\in\Lambda\}$ be the set of mutually exclusive hypotheses

$$H_s \equiv \mathbf{X} \text{ is a sequence of independent samples from } \hat{q}_s. \quad (1)$$

The problem is to classify \mathbf{X} by choosing one of the densities \hat{q}_s . There are really two problems here, depending on whether or not one can assume a priori that one of the H_s is true. If so, then our problem is to find t such that $q^\dagger(\mathbf{x}) = \hat{q}_t(\mathbf{x})$. If not, then the problem is to find t such that $\hat{q}_t(\mathbf{x})$ is "closest to" $q^\dagger(\mathbf{x})$ in some well-defined, acceptable sense. Most of the time, the latter case applies—one cannot assume that $q^\dagger = \hat{q}_t$ for any t . Speech-processing applications are good examples—speech is dealt with in terms of Gaussian models even though it is well known that speech is not Gaussian. We restrict consideration to classification densities \hat{q}_s that comprise an exponential family,

$$\hat{q}_s(\mathbf{x}) = p(\mathbf{x}) \exp \left[-\hat{\lambda}^{(s)} - \sum_{k=1}^n \hat{\beta}_k^{(s)} f_k(\mathbf{x}) \right], \quad (2)$$

where $p(\mathbf{x})$ and $f_k(\mathbf{x})$ are fixed functions and $\hat{\lambda}^{(s)}$ and $\hat{\beta}_k^{(s)}$ are constants. A set of Gaussian densities is one example of such an exponential family. We place no restrictions on the unknown process q^\dagger .

Exponential families can always be expressed as the result of a minimum cross-entropy problem [5-7]. In particular, the \hat{q}_s satisfy

$$H(\hat{q}_s, p) = \min_{q'} H(q', p), \quad (3)$$

where H is the *cross-entropy* (discrimination information, directed divergence, I-divergence, Kullback-Liebler number, etc.),

$$H(q,p) = \int q(\mathbf{x}) \log \left[\frac{q(\mathbf{x})}{p(\mathbf{x})} \right] d\mathbf{x}, \quad (4)$$

and where q' varies over the set of densities that satisfy the *constraints*

$$\int \hat{q}_s(\mathbf{x}) f_k(\mathbf{x}) d\mathbf{x} = \hat{F}_k^{(s)} \quad (5)$$

for known numbers $\hat{F}_k^{(s)}$. In the solution (2) the constants $\hat{\beta}_k^{(s)}$ and $\hat{\lambda}^{(s)}$ are Lagrangian multipliers chosen to satisfy the constraints (5) and

$$\int \hat{q}_s(\mathbf{x}) d\mathbf{x} = 1.$$

In the notation of Refs. 7 and 8, one can express (3) as $\hat{q}_s = p \circ \hat{I}_s$, where \hat{I}_s represents the information given by the constraints (5). The density p is called the *prior*, and the densities \hat{q}_s are called *posteriors*.

REVIEW OF THE TWO CLASSIFICATION METHODS

Maximum-Likelihood Classification

In the maximum-likelihood (ML) approach one classifies \mathbf{X} by

$$\max_s p(\mathbf{X} | H_s), \quad (6)$$

where $p(\mathbf{X} | H_s)$ is the probability that \mathbf{X} is the result of n independent samples from $\hat{q}_s(\mathbf{x})$. Bayes's law yields

$$p(H_s | \mathbf{X}) = p(\mathbf{X} | H_s) \frac{p(H_s)}{p(\mathbf{X})},$$

so that ML classification is equivalent to maximum-a-posteriori (MAP) classification,

$$\max_s p(H_s | \mathbf{X}),$$

when the hypotheses H_s have equal prior probabilities. ML classification is used in a variety of applications, even when clearly one cannot assume a priori that one of the hypotheses is true. This practice can be justified on practical grounds—it works—but it has not been justified on compelling theoretical grounds.

Minimum-Cross-Entropy Classification

Minimum-cross-entropy (MCE) classification of information from the unknown process q^\dagger proceeds from knowledge of the expectations

$$\int q^\dagger(\mathbf{x}) f_k(\mathbf{x}) d\mathbf{x} = F_k, \quad (7)$$

that is, expectations of the same constraint functions $f_k(\mathbf{x})$ as in (2). The quantity $\mathbf{F} \equiv F_1, \dots, F_n$ is called a *feature vector*—its elements are the data to be classified. Let I represent the constraints (7), and let the density p in (2) be considered as a prior estimate of q^\dagger . Then a method of classifying \mathbf{F} using MCE consists of the following two-step procedure [1]:

1. Compute $q = p \circ I$, the minimum-cross-entropy estimate of q^\dagger based on the information (7).

2. Choose one of the classification densities by the MCE rule

$$\min_{s \in \Lambda} H(q, \hat{q}_s). \quad (8)$$

In Ref. 1 it is shown that

$$H(q^\dagger, \hat{q}_s) = H(q^\dagger, q) + H(q, \hat{q}_s) \quad (9)$$

holds. Now, the MCE estimate $q = p \circ I$ minimizes the term $H(q^\dagger, q)$ in the following sense: Of all densities having the general form (2), q is the closest possible density to q^\dagger . (This property is known as expectation-value matching [7].) Since the second term on the right-hand side of (9) is minimized by (8), it follows that MCE classification is optimal in the sense of minimizing the total distortion $H(q^\dagger, \hat{q}_s)$. An alternative MCE method of classifying \mathbf{F} is to use the rule

$$\min_{s \in \Lambda} H(\hat{q}_s \circ I, \hat{q}_s). \quad (10)$$

In words, each of the classification densities \hat{q}_s is in turn considered as a prior estimate of q^\dagger ; when the information \mathbf{F} is taken into account, the resulting posterior estimate of q^\dagger is $\hat{q}_s \circ I$. The rule (10) chooses the classification density \hat{q}_s that, when considered as a prior estimate of q^\dagger , is changed the least by taking \mathbf{F} into account.

Both of the MCE rules (8) and (10) have compelling intuitive and information-theoretic justifications. Fortunately one does not have to choose between them. Because the constraints (5) and (7) involve the same constraint functions $f_k(\mathbf{x})$, it follows [7, Property 14] that

$$\hat{q}_s \circ I = (p \circ \hat{I}_s) \circ I = p \circ I = q \quad (11)$$

holds, which in turn means that (8) and (10) are equivalent.

Computationally, it turns out that one need not compute $q = p \circ I = \hat{q}_s \circ I$, as the rules (8) and (10) are equivalent to

$$\min_{s \in \Lambda} \left\{ \hat{\lambda}^{(s)} + \sum_{k=1}^n \hat{\beta}_k^{(s)} F_k \right\}, \quad (12)$$

where the $\hat{\lambda}^{(s)}$ and $\hat{\beta}_k^{(s)}$ are the Lagrangian multipliers from the classification densities (2) [1].

For the application being considered here, the expectations F_k are estimated from \mathbf{X} by

$$F_k = \frac{1}{M} \sum_{i=1}^M f_k(\mathbf{x}_i). \quad (13)$$

COMPARISON OF THE CLASSIFICATION METHODS

I begin the comparison by computing the consequences of the ML rule (6) given the form (2) for the classification densities. One has

$$\begin{aligned} p(\mathbf{X} | H_s) &= \prod_{i=1}^M \hat{q}_s(\mathbf{x}_i) \\ &= \exp \left[-M \hat{\lambda}^{(s)} - \sum_{i=1}^M \sum_{k=1}^n \hat{\beta}_k^{(s)} f_k(\mathbf{x}_i) \right] \prod_{i=1}^M p(\mathbf{x}_i), \end{aligned} \quad (14)$$

bearing in mind that this is valid only if one knows that \mathbf{X} came from one of the $\hat{q}_s(\mathbf{x})$.

The ML rule (6) is equivalent to the rule

$$\min_s \{ -\log p(\mathbf{X} | H_s) \}.$$

Substitution of (14) yields

$$\min_s \left\{ M\hat{\lambda}^{(s)} + \sum_{i=1}^M \sum_{k=1}^n \hat{\beta}_k^{(s)} f_k(\mathbf{x}_i) - \sum_{i=1}^M \log p(\mathbf{x}_i) \right\}. \quad (15)$$

The last sum in (15) involves terms independent of s and can therefore be dropped. Also, dividing by the constant M has no effect. Hence (15) is equivalent to

$$\min_s \left\{ \hat{\lambda}^{(s)} + \sum_{k=1}^n \hat{\beta}_k^{(s)} \frac{1}{M} \sum_{i=1}^M f_k(\mathbf{x}_i) \right\}.$$

Substitution of (13) yields

$$\min_s \left\{ \hat{\lambda}^{(s)} + \sum_{k=1}^n \hat{\beta}_k^{(s)} F_k \right\},$$

which is the same as the MCE rule (12).

I have just shown that ML classification is equivalent to MCE classification when one can assume that \mathbf{X} comes from one of the classification densities \hat{q}_s . This fact was shown previously by Kupperman [9] and Kullback [5], although the derivation here is carried out more directly and in terms of the computational MCE classification rule (12) that was derived in Ref. 1. Recently, Csiszár and Tusnády have considered the connection between ML and MCE when \mathbf{X} results from a mapping of samples from one of the \hat{q}_s [10].

What about the case when one cannot assume that \mathbf{X} comes from one of the classification densities \hat{q}_s ? In this case it is common to use the ML rule (6) anyway, without good theoretical justification. But the case *is* covered by MCE classification, because rule (12) was derived out in Ref. 1 without assuming that the feature vector \mathbf{F} is the same as any of the $\hat{\mathbf{F}}^{(s)}$ that determine the classification densities by (3), (4), and (5) or that estimates of \mathbf{F} are obtained by sampling one of the \hat{q}_s . It was assumed only that the goal is to find the $\hat{\mathbf{F}}^{(s)}$ that "best resembles" \mathbf{F} and that the MCE criterion (8) is reasonable. When \mathbf{X} cannot be assumed to come from one of the \hat{q}_s , it turns out that those who apply ML anyway are doing MCE classification.

DISCUSSION

MCE classification provides a general method for taking a sequence of independent vector-valued samples \mathbf{x}_i from an unknown process q^\dagger and classifying that sequence by identifying a member of a set of exponential-class densities $\{\hat{q}_s(\mathbf{x}):s \in \Lambda\}$. The classification rule (12) combines the results of a two-step procedure: The first step obtains from \mathbf{X} a minimum-cross-entropy estimate q of q^\dagger . The second step identifies the density \hat{q}_s that is closest to q in the cross-entropy sense. With the assumption that the \mathbf{x}_i come from one of the \hat{q}_s , MCE classification reduces to ML classification. Without this assumption MCE classification applies anyway and thereby provides a theoretical justification for the technically incorrect use of ML.

Furthermore, the \hat{q}_s may themselves be approximations if the constraints $\hat{\mathbf{F}}^{(s)}$ in (5) are approximations based on training data in the same sense as (13). That is, the \hat{q}_s may be approximations based on samples from "true densities" \hat{q}_s^\dagger . Then, even if one can assume that the classification-data vector \mathbf{X} comes from one of the \hat{q}_s^\dagger , one cannot assume that \mathbf{X} comes from one of the classification densities \hat{q}_s ; again, ML cannot be applied in principle.

AN EXAMPLE—ESTIMATION OF STRUCTURED COVARIANCE MATRICES

Recently, Burg, Luenberger, and Wenger [4] have generalized the popular Burg technique [11] for estimating the autocorrelation function of a random process from time-domain samples. The new

method estimates covariance matrix of specified structure from vector-valued samples of a random process. Written in terms of the notation here, Burg et al. consider the set of classification densities

$$\hat{q}_s(\mathbf{x}) = (2\pi)^{-1/2 N} |\mathbf{R}_s|^{-1/2} \exp(-1/2 \mathbf{x}^t \cdot \mathbf{R}_s^{-1} \cdot \mathbf{x}). \quad (16)$$

This is Eq. (1) in Ref. 4. The superscript t indicates a transpose, the raised dot (\bullet) indicates a vector or matrix product, and $\{\mathbf{R}_s; s \in \Lambda\}$ is a finite or infinite set of feasible covariance matrices. Given a data vector \mathbf{X} consisting of M vector-valued samples from an unknown density $q^\dagger(\mathbf{x})$, the sample covariance matrix \mathbf{R} is defined as

$$\mathbf{R} = \frac{1}{M} \sum_{i=1}^M \mathbf{x}_i \mathbf{x}_i^t. \quad (17)$$

Burg et al. assume that \mathbf{X} came from one of the \hat{q}_s (that is, $q^\dagger = \hat{q}_s$ holds for some s), and they classify \mathbf{X} by the ML rule (6). The result is the classification rule

$$\max_s \left\{ -\log |\mathbf{R}_s| - \text{Tr}(\mathbf{R}_s^{-1} \cdot \mathbf{R}) \right\}, \quad (18)$$

where Tr indicates a trace operation. This is Eq. (4) in Ref. 4, except that \mathbf{R} and \mathbf{S} are replaced respectively by \mathbf{R}_s and \mathbf{R} .

Since (16) belongs to the class of generalized exponentials, the results of the section beginning on page 2 apply—(18) must be equivalent to MCE classification, and (18) must also apply in the more realistic case where one cannot assume that \mathbf{X} comes from one of the \hat{q}_s . For completeness, one can demonstrate the connection explicitly by showing that (18) is a special case of the MCE rule (12).

One needs to express (16) as minimum-cross-entropy posteriors $\hat{q}_s = p \circ \hat{I}_s$. That is, one needs to express (16) in the form (2). As a prior, one can use

$$p(\mathbf{x}) = (2\pi)^{-1/2 N} \exp(-1/2 \mathbf{x}^t \cdot \mathbf{I} \cdot \mathbf{x}), \quad (19)$$

where \mathbf{I} is the identity matrix. Using (19), one rewrites (16) as

$$\hat{q}_s(\mathbf{x}) = p(\mathbf{x}) |\mathbf{R}_s|^{-1/2} \exp(-1/2 \mathbf{x}^t \cdot (\mathbf{R}_s^{-1} - \mathbf{I}) \cdot \mathbf{x}). \quad (20)$$

Defining

$$\hat{\lambda}^{(s)} = -\log |\mathbf{R}_s|^{-1/2} \quad (21)$$

and

$$\hat{\beta}_{ij}^{(s)} = 1/2 \{ \mathbf{R}_s^{-1} - \mathbf{I} \}_{ij} \quad (22)$$

permits one to rewrite (20) as

$$\hat{q}_s(\mathbf{x}) = p(\mathbf{x}) \exp \left[-\hat{\lambda}^{(s)} - \sum_{ij} \hat{\beta}_{ij}^{(s)} x_i x_j \right], \quad (23)$$

which is just the desired form (2). The constraint functions in this case are $f_{ij}(\mathbf{x}) = x_i x_j$. The expectations (5) are just the covariances

$$\int \hat{q}_s(\mathbf{x}) x_i x_j d\mathbf{x} = \{ \mathbf{R}_s \}_{ij}.$$

Given the data vector \mathbf{X} , elements $\{ \mathbf{R} \}_{ij}$ of the sample covariance matrix (17) are just estimates of the expectations $\int d\mathbf{x} q^\dagger(\mathbf{x}) x_i x_j$. Hence, using (17), (21), and (22), one can write the MCE classification rule (12) as

$$\min_s \left\{ \frac{1}{2} \log |\mathbf{R}_s| + \frac{1}{2} \sum_{ij} \{\mathbf{R}_s^{-1} - \mathbf{I}\}_{ij} \{\mathbf{R}\}_{ij} \right\}. \quad (24)$$

The term involving the identity matrix \mathbf{I} does not depend on s . It follows that (24) is equivalent to

$$\min_s \left\{ \log |\mathbf{R}_s| + \sum_{ij} \{\mathbf{R}_s^{-1}\}_{ij} \{\mathbf{R}\}_{ij} \right\}, \quad (25)$$

where the factor $\frac{1}{2}$ has also been dropped. Since \mathbf{R} is symmetric, as can be seen from (17), then

$$\sum_{ij} \{\mathbf{R}_s^{-1}\}_{ij} \{\mathbf{R}\}_{ij} = \text{Tr}(\mathbf{R}_s^{-1} \cdot \mathbf{R}).$$

Eq. (25) then becomes

$$\min_s \left\{ \log |\mathbf{R}_s| + \text{Tr}(\mathbf{R}_s^{-1} \cdot \mathbf{R}) \right\},$$

which is equivalent to (16).

REFERENCES

1. J.E. Shore and R.M. Gray "Minimum-cross-entropy pattern classification and cluster analysis," IEEE Trans. Patt. Anal. and Machine Intell. **PAMI-4**, 11-17 (Jan. 1982).
2. A. Buzo, A.H. Gray, Jr., R.M. Gray, and J.D. Markel, "Speech coding based upon vector quantization," IEEE Trans. Acoust. Speech Signal Processing **ASSP-28**, 562-574 (Oct. 1980).
3. R.M. Gray, A.H. Gray, Jr., G. Rebolledo, and J.E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," IEEE Trans. Inform. Theory **IT-27**, 708-721 (Nov. 1981).
4. J.P. Burg, D.G. Luenberger, and D.L. Wenger, "Estimation of structured covariance matrices," Proc. IEEE **76**, 963-974 (Sept. 1982).
5. S. Kullback, *Information Theory and Statistics*, Dover, New York, 1969, and Wiley, New York, 1959.
6. I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," Ann. Math. Stat. **3**, 146-158 (1975).
7. J.E. Shore and R.W. Johnson, "Properties of cross-entropy minimization," IEEE Trans. Inform. Theory **IT-27**, 472-482 (July 1981).
8. J.E. Shore and R.W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," IEEE Trans. Inform. Theory **IT-26**, 26-37 (Jan. 1980).
9. M. Kupperman, "Probabilities of hypotheses and information-statistics in sampling from exponential class populations," Ann. Math. Stat. **29**, 571-574 (1958).
10. I Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," to be published.
11. J.P. Burg, "Maximum Entropy Spectral Analysis," Ph.D. dissertation, Stanford University, 1975 (University Microfilms 75-25, 499).