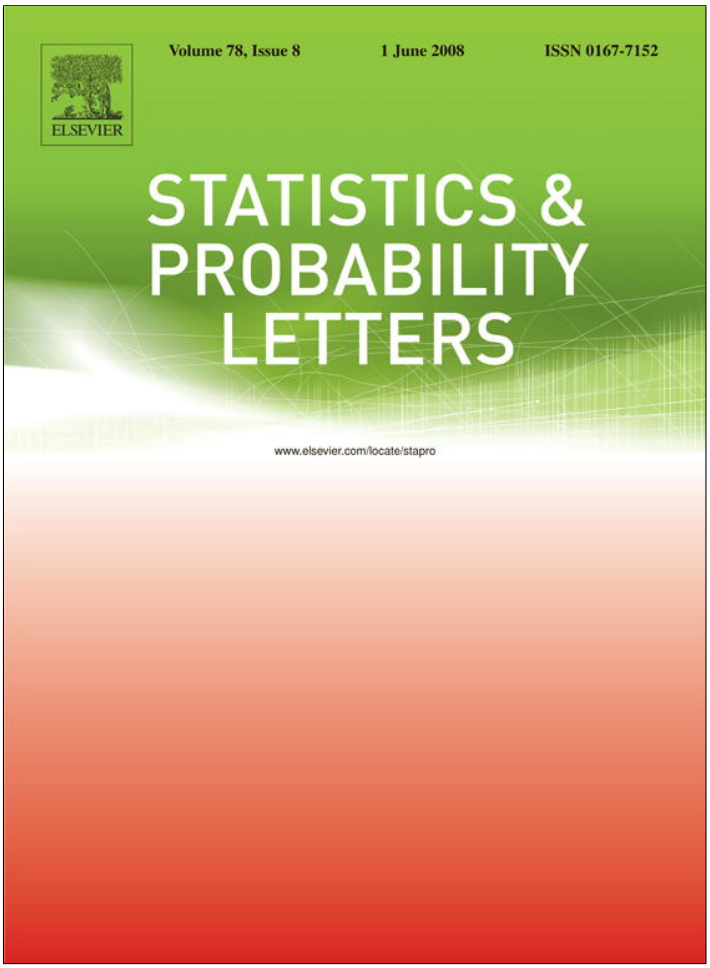

Available online at www.sciencedirect.com
ScienceDirect
Statistics and Probability Letters 78 (2008) 975–984
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

On the minimization of concave information functionals for unsupervised classification via decision trees

Damianos Karakos[a,b], Sanjeev Khudanpur[a,b], David J. Marchette[c], Adrian Papamarcou[d], Carey E. Priebe[e,*]

[a] *Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD 21218, United States*
[b] *Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, United States*
[c] *Naval Surface Warfare Center, Code Q21, Dahlgren, VA 22448-5100, United States*
[d] *Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, United States*
[e] *Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, United States*

Received 13 September 2006; received in revised form 6 September 2007; accepted 25 September 2007
Available online 1 November 2007

Abstract

A popular method for unsupervised classification of high-dimensional data via *decision trees* is characterized as minimizing the empirical estimate of a concave information functional. It is shown that minimization of such functionals under the *true* distributions leads to perfect classification.

1. Introduction

Classification has been a very active area of research across numerous fields. It has been studied under various names: clustering, intrinsic classification, mixture modeling, etc. In a nutshell, the problem is posed as follows. Given a number of objects $x_1, \ldots, x_M$, each having a number of observable *features* or attributes, we would like to group them into *classes* that correspond, somehow, to some meaningful underlying categories. In other words, we would like to assign a *class-label* to each of these objects, so that objects which are similar in some sense are assigned the same label. We are particularly interested in the *unsupervised* classification case, where no statistics of the objects jointly with their class-labels are known, and the goal is to group the objects into *clusters* based only on their observable features, such that each cluster contains objects that share some salient properties. In some cases, there may be a notion of a "true" class-label of each object, which has simply not been provided; it may then be appropriate to view the set of class-labels as given and the class-label of each particular object as a *latent variable*, and to evaluate the performance of a clustering scheme by a *post hoc* assignment of the true class-labels to (a subset of) objects in each resulting cluster. In other cases, there may be no natural notion of "true" class-labels; the efficacy of the clustering

* Corresponding author. Tel.: +1 410 516 7200; fax: +1 410 516 7459.
E-mail addresses: damianos@jhu.edu (D. Karakos), khudanpur@jhu.edu (S. Khudanpur), david.marchette@navy.mil (D.J. Marchette), adrian@eng.umd.edu (A. Papamarcou), cep@jhu.edu (C.E. Priebe).

0167-7152/$ - see front matter
doi:10.1016/j.spl.2007.09.058

scheme is often measured in such cases by the economy in *description length* attained by a two-step description of the objects by first describing the attributes common to the clusters and then describing the differential attributes of each object within the cluster. k-Means Clustering and Mixture Modeling using the Expectation Maximization (EM) Algorithm (Dempster et al., 1977; Jelinek, 1997) are examples of techniques used for unsupervised classification.

In the following, we investigate the problem of unsupervised classification using Rényi's α-divergence (Csiszár, 1995) as the "distance" – or measure of dissimilarity – between objects. We interpret the empirical distribution of the observable features of each object as a probability distribution over possible feature values. The choice of α-divergence is guided by the observation that it arises naturally in many applications, e.g. in bounds on the probability of error in hypothesis testing (Cover and Thomas, 1991), in channel coding problems (Csiszár, 1995), and in problems that involve the Hellinger distance between distributions (cf e.g. Beran (1977)). Moreover, α-divergences give rise to a family of concave information functionals (Csiszár, 1995), whose minimization (as we prove in Theorem 1) can lead to pairs of distributions with disjoint supports – a useful property that guarantees an accurate classification in a suitable limiting case.

We begin by formalizing our notation and providing some background on α-divergences in Section 2. We show in Section 3 how α-divergences may be used as the criterion for unsupervised classification via Iterative Denoising Trees (Priebe et al., 2004; Karakos et al., 2005), and we present experimental results from document categorization and hyperspectral imaging in Section 4. Finally, in Section 5, we present the main theoretical result of this paper, namely, that the convex combination of concave functions, used in the tree construction, is minimized *only* by a pair of distributions with disjoint supports.

A key insight we provide is that (unsupervised) *clustering driven by the minimization of concave functionals of class-conditional distributions* naturally leads to homogeneous clusters.

2. Mathematical preliminaries

Random variables (r.v.) will be denoted by capital letters, while their realizations will be denoted by the corresponding lowercase letters. Random vectors will be denoted by boldface letters, e.g., $\boldsymbol{X} = [X_1, \ldots, X_n]$, while lowercase boldface will denote their realizations. Unless otherwise noted, the dimension of $\boldsymbol{X}$ will be n. All random variables and vectors will be assumed to lie in discrete finite spaces, and we will use the corresponding calligraphic letters to denote their alphabets. For example, $X \in \mathcal{X}$, while $\boldsymbol{X} \in \mathcal{X}^n$.

The probability mass function (pmf) of a r.v. will be denoted by the appropriate subscript, e.g., P_X for r.v. X. The conditional pmf of X given Y will be denoted by $W_{X|Y}(\cdot|\cdot)$, a $\mathcal{Y} \times \mathcal{X}$ matrix whose rows sum to one. The subscripts X and $X|Y$ on pmfs will often be omitted when obvious from context. The support set of a r.v. X is denoted by $\mathcal{S}(X)$ or $\mathcal{S}(P_X)$, and is the subset of $\mathcal{X}$ such that $P_X(x) > 0$ if and only if (iff) $x \in \mathcal{S}(X)$.

Rényi (1961) introduced the order-α information divergence, abbreviated α-divergence, of a pmf P_X from another pmf Q_X as

$$D_\alpha(P_X \parallel Q_X) = \frac{1}{\alpha - 1} \log \sum_{x \in \mathcal{X}} P_X^\alpha(x) Q_X^{1-\alpha}(x), \tag{1}$$

where $\alpha > 0, \alpha \neq 1$. The limit of (1), as $\alpha \to 1$, is the well-known Kullback–Leibler information divergence, or KL-divergence:

$$D(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)},$$

where, by convention, $0 \log 0 = 0 \log \frac{0}{0} = 0$ and $t \log \frac{t}{0} = +\infty$ for $t > 0$. Finally, when $\alpha = 1/2$,

$$D_{\frac{1}{2}}(P \parallel Q) = \log \left(\sum P^{\frac{1}{2}}(x) Q^{\frac{1}{2}}(x) \right)^{-2} = \log \left(1 - \frac{1}{2} \left\| P^{\frac{1}{2}} - Q^{\frac{1}{2}} \right\|^2 \right)^{-2}.$$

The reader may recognize that the quantity subtracted from unity is the Hellinger distance (Beran, 1977) between P and Q.

Given a pmf P_Y on $\mathcal{Y}$ and a conditional pmf $W_{X|Y}$, the order-α mutual information is defined (cf. Csiszár (1995)) as

$$I_\alpha(P_Y; W_{X|Y}) = \min_{Q_X} \sum_{y\in\mathcal{Y}} P_Y(y) D_\alpha(W_{X|Y}(\cdot|y) \parallel Q_X(\cdot)), \tag{2}$$

where, again by convention, $0\, D_\alpha(\cdot \parallel \cdot) = 0$ regardless of whether the α-divergence is finite. Note that $I_\alpha(P; W)$ is always finite (Csiszár, 1995), even if α-divergence is not.

If one views rows of $W_{X|Y}$ as points in $\mathbb{R}^{\mathcal{X}}$, then the minimizer in the definition of I_α may be interpreted as the *centroid* of these points, with P_Y playing the role of their *mass* and α-divergence playing the role of *distance*. Furthermore, if P_Y is uniform, I_α may be interpreted as the *radius of the smallest ball* covering the conditional pmfs that constitute the rows of $W_{X|Y}$.

One can easily see that this definition of order-α mutual information reduces to the usual (Shannon's) definition of mutual information when $\alpha \to 1$:

$$I(P_Y; W_{X|Y}) = \sum_{y\in\mathcal{Y}} P_Y(y) D(W_{X|Y}(\cdot|y) \parallel (P_Y \circ W)(\cdot)),$$

where $(P_Y \circ W)(\cdot)$ is the induced marginal pmf on $\mathcal{X}$ through P_Y and W:

$$(P_Y \circ W)(x) = \sum_{y\in\mathcal{Y}} P_Y(y) W_{X|Y}(x|y).$$

Some useful properties of D_α and I_α (which continue to hold for the limiting case $\alpha \to 1$) are summarized below (from Csiszár (1995)):

(1) $0 \le D_\alpha(P \parallel Q) \le +\infty$: the first equality holds iff $P = Q$, and the second iff either $\mathcal{S}(P) \cap \mathcal{S}(Q) = \emptyset$, or $\alpha > 1$ and $\mathcal{S}(P) \not\subseteq \mathcal{S}(Q)$.
(2) $I_\alpha(P; W)$ is a continuous, concave function of P.
(3) $D_\alpha(P \parallel Q)$ is convex in Q. For $\alpha < 1$, it is convex in the pair (P, Q).

3. Iterative denoising via α-divergence minimization

Let $\mathcal{C}$ denote a collection of data points. Each data point is assumed to belong to a latent class. There are no prior assumptions about the properties of the classes; that is, it is unknown what their memberships are, and what they represent. Our goal is to partition $\mathcal{C}$ into disjoint sets $A_1, \ldots, A_m$, such that all data points which belong to the same set share some common features, distinct from the features shared by points of other sets. The number of sets m may be specified before the partition is determined, or computed automatically from the data (e.g., the smallest partition-size that satisfies some conditions). Formally, we have the following:

(1) A "vocabulary" $\mathcal{X}$, and a class-label space $\mathcal{Y}$.
(2) A collection $\mathcal{C}$ of data points (sequences) $\boldsymbol{X}(1), \ldots, \boldsymbol{X}(N)$, each in $\mathcal{X}^n$.
(3) A "hidden" class-label $Y(j)$ associated with each sequence $\boldsymbol{X}(j)$.
(4) We assume that the $Y(j)$'s are i.i.d with common pmf P_Y, and the $\boldsymbol{X}(j)$'s are (conditionally) mutually independent given the class-labels $Y(j)$. Moreover, the conditional distribution of each sequence $\boldsymbol{X}(j)$ given $Y(j)$ is stationary and ergodic, and its $\mathcal{X}$-marginal is denoted $W(x|y)$.
(5) Our goal is to find a partition $A_1, \ldots, A_m$ of $\mathcal{C}$, such that for all $i \ne j$,

$$\boldsymbol{X}(i), \boldsymbol{X}(j) \in A_k \text{ for some } k \in \{1, \ldots, m\} \quad \Leftrightarrow \quad Y(i) = Y(j),$$

with high probability. This is true, e.g., for a partition in which each A_k contains sequences with the same label, and $m = |\mathcal{Y}|$.

Let S be any subset of $\mathcal{C}$.

Definition 1. The α-divergence $D_\alpha(S)$ of a set S of data points is defined by

$$D_\alpha(S) \triangleq \min_{Q} \sum_{\boldsymbol{X}\in S} D_\alpha(\hat{P}_{\boldsymbol{X}} \parallel Q), \tag{3}$$

where $\hat{P}_{\boldsymbol{X}}$ is the empirical marginal pmf derived from $\boldsymbol{X}$, and Q a pmf on $\mathcal{X}$.

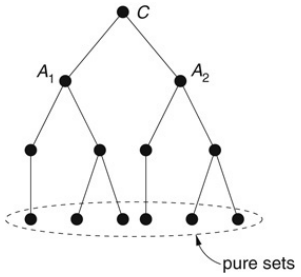

Fig. 1. The classification tree, partitions the data set corresponding to each node into two sets, such that the sum of the divergences of the children is the minimum possible. As the length of the data sequences goes to infinity, this procedure produces pure sets (i.e., belonging only to one class) at the leaves.

$D_\alpha(S)$ provides an indication of the "homogeneity" of the set S. For example, if $D_\alpha(S) = 0$, then all elements of S must have the same empirical distribution—a good indication that they were generated by the same process (i.e., they have the same class-label). This suggests that our classification problem is equivalent to that of finding a partition $A_1, \ldots, A_m$, with the smallest possible m, such that

$$\sum_{j=1}^{m} D_\alpha(A_j) < \epsilon, \tag{4}$$

for some suitable $\epsilon > 0$.

We propose to use sensing and processing trees (Priebe et al., 2004) to find a partition that satisfies (4). At the root of the tree, we start with the whole corpus $\mathcal{C}$. Then, we proceed iteratively, performing the following steps at each tree-node:

- **Stopping criterion:** At a node $S \subseteq \mathcal{C}$, if $D_\alpha(S) < \epsilon$, then declare the set S as "pure", that is, declare that all data points in S have the same class-label, whatever that label may be, and S is one of the sets in the eventual partition of $\mathcal{C}$ (i.e., S is a "leaf").
- **Splitting criterion:** Otherwise, partition the data points S into two subsets A_1, A_2, such that the sum $D_\alpha(A_1) + D_\alpha(A_2)$ is as small as possible, and create two children of S in the tree, each one corresponding to the sets A_1 and A_2.

The iteration continues until all leaf-nodes meet the stopping criterion, or, when a desirable number m of nodes has been reached. Fig. 1 depicts the procedure of growing the tree. As mentioned in Priebe et al. (2004) and Karakos et al. (2005), the data at each node are projected into a low-dimensional space which is, in general, different from the projection used at other nodes. Furthermore, in order to split each node in a computationally tractable way (for nodes with more than 40–50 data points, it is obviously infeasible to perform an exponential number of 2-way splits in order to find the one that minimizes the sum of divergences) some heuristic has to be used; for example, Chou's algorithm (Chou, 1991), which is a variant of K-means, provably converges to a local optimum.

4. Empirical results

To demonstrate the usefulness of the iterative procedure described above, we perform clustering experiments in two areas: text categorization and hyperspectral imaging. We use minimization of KL-divergences ($\alpha \to 1$) as the optimization criterion at each node, and each ISPDT is grown until a specific number of leaves is reached. Furthermore, we compare the performance of ISPDTs with that of K-means clustering (with a random initialization of cluster centroids), and Gaussian mixture modeling (Fraley and Raftery, 1999). Before clustering, the data are projected into a low-dimensional space using principal components analysis, where the number of dimensions is based on the location of the "knee" on the scree plot.

4.1. Text categorization

We are using documents from the Science News corpus, which consists of extended abstracts on various topics within Anthropology, Astronomy, Technology, and Medicine. The number of documents per subject are: 54 in Anthropology, 121 in Astronomy, 60 in Technology, and 280 in Medicine.

Table 1
Error rates in text classification of the Science News corpus

Method	Error (%)
K-means	16.2
Gaussian mixtures	24.2
ISPDT	13.9

Table 2
Error rates in the classification of hyperspectral image pixels

Method	Error (%)
K-means	19.7
Gaussian mixtures	9.6
ISPDT	9.8

This collection $\mathcal{C}$ of documents represents the root of the ISPDT. After performing automatic *stemming* on the words in each document, i.e., transforming each word into its base form, e.g., singular, present tense, etc., and removing stop-words (such as *a, the, is*, etc.), we collect statistics of word occurrences for each document. Thus the observed feature of each document is a fixed-length vector of length equal to the vocabulary size ($|\mathcal{X}| = 10\,000$). Then, we *smooth* (Jelinek, 1997) the probability distribution of each document; that is, we assign non-zero probabilities to unseen words, based on their overall frequency in $\mathcal{C}$. This way, we overcome the problem of sparseness, which is commonplace in statistical text processing.

The error rates of K-means, Gaussian mixture modeling, and KL-divergence based ISPDT are shown in Table 1. Based on the scree plot, the number of PCA dimensions for the K-means and the Gaussian mixture modeling is chosen to be equal to 5. The K-means algorithm was executed 10 times with different random initializations (cluster centroids); the error rates shown are averages over these 10 experiments. For the ISPDT, the data at each node are projected into a two-dimensional simplex, using a pair of principal components chosen among the first 5; the projection finally chosen is the one which yields the split with the lowest sum $D(A_1) + D(A_2)$.

As we can see from Table 1, ISPDTs have performance which is superior to that of K-means and Gaussian mixture modeling; this is noteworthy, especially since K-means and Gaussian mixture modeling are usually considered as two of the most effective approaches to unsupervised classification.

4.2. Hyperspectral imaging

We performed experiments with hyperspectral satellite images, where each data point corresponds to a multi-dimensional pixel — each dimension represents a particular frequency band. Furthermore, the spectrum of each pixel is actually a *distribution* of energy over frequencies. Hence, with the appropriate normalization, the spectrum of a data point plays the role of its "empirical distribution". This allows us to skip the step of computing a high-dimensional distribution for each data point; its distribution is already supplied.

The class-labels of the pixels correspond to different types of vegetation: runway (144 pixels), pine (177 pixels), scrub (200 pixels) and swamp (79 pixels). Table 2 shows the classification error rates for K-means, Gaussian mixture modeling, and ISPDT; the first 3 principal components were chosen for dimensionality reduction, based on the scree plot. As we can see, the performance of ISPDTs exceeds that of K-means, and is almost identical to the performance of the Gaussian mixture modeling. The resulting ISPDT is shown in Fig. 2.

5. Theoretical justification of the splitting criterion

In this section, we give a theoretical justification of the appropriateness of the splitting criterion, by considering the case $n \to \infty$. Assuming that each data sequence X^n is generated by a stationary and ergodic process, it follows that $\hat{P}_{\boldsymbol{X}} \to W(\cdot|y)$, the $\mathcal{X}$-marginal of the process distribution, where y is the true label of $\boldsymbol{X}$. Hence, in this limiting case,

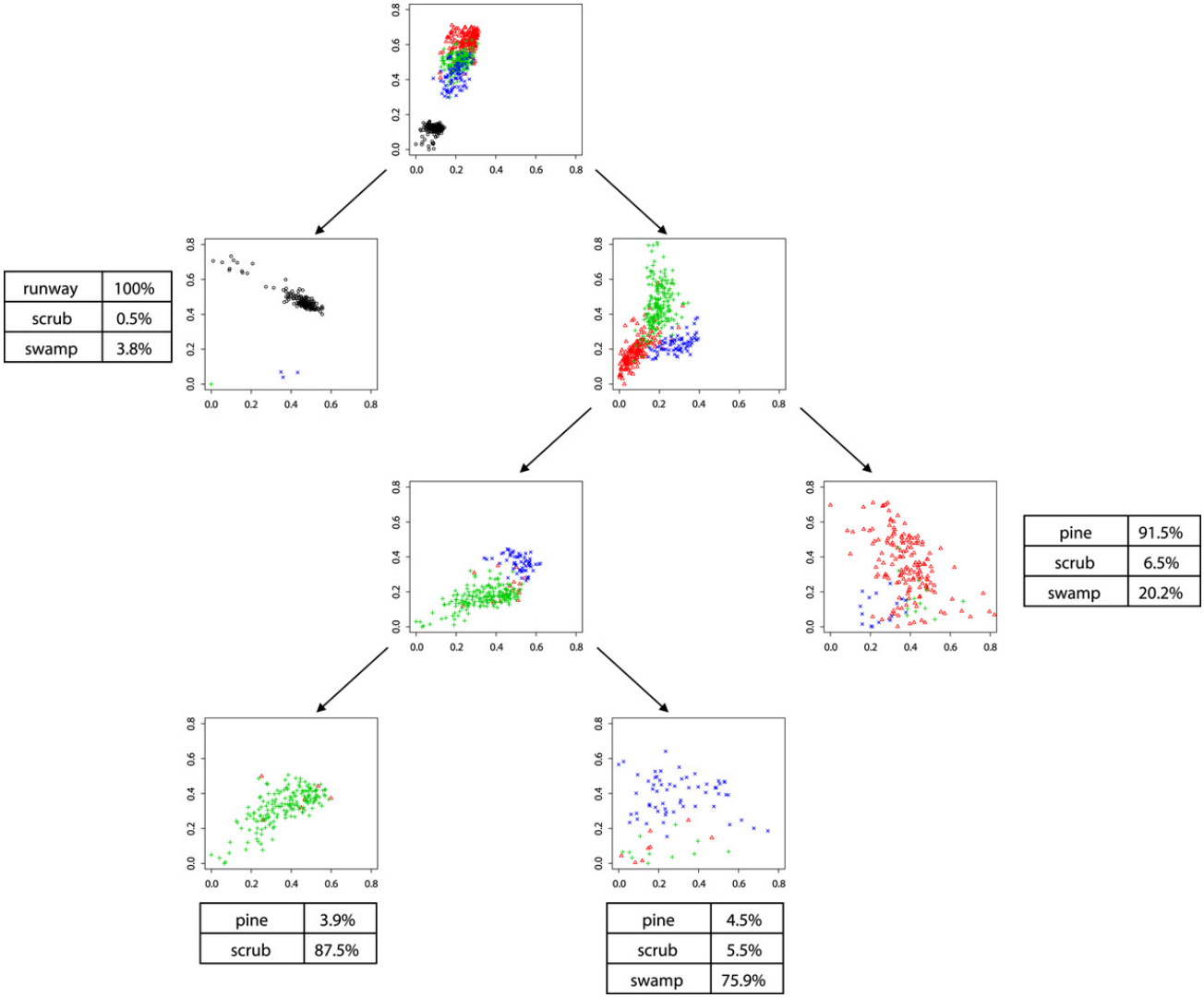

Fig. 2. The ISPDT in the hyperspectral data case. Depiction of labels is as follows: circles correspond to runway, triangles to pine, crosses to scrub and x's to swamp. The total misclassification rate is 9.8%.

the minimization of $D_\alpha(A_1) + D_\alpha(A_2)$, with respect to a partition (A_1, A_2) of S, is equivalent to the minimization of

$$\frac{1}{|S|}(D_\alpha(A_1) + D_\alpha(A_2)) = \lambda \min_{U_1} \sum_{y\in\mathcal{Y}} P_1(y) D_\alpha(W(\cdot|y) \parallel U_1(\cdot)) + (1-\lambda) \min_{U_2} \sum_{y\in\mathcal{Y}} P_2(y) D_\alpha(W(\cdot|y) \parallel U_2(\cdot)), \tag{5}$$

where

$$P_1(y) = \frac{1}{|A_1|} \sum_{j:\boldsymbol{X}(j)\in A_1} \mathbf{1}[Y(j) = y], \qquad P_2(y) = \frac{1}{|A_2|} \sum_{j:\boldsymbol{X}(j)\in A_2} \mathbf{1}[Y(j) = y],$$

$$P_Y(y) = \frac{1}{|S|} \sum_{j:\boldsymbol{X}(j)\in A_1\cup A_2} \mathbf{1}[Y(j) = y] = \lambda P_1(y) + (1-\lambda) P_2(y), \quad \text{and} \quad \lambda = \frac{|A_1|}{|S|}.$$

But by the definition of the order-α mutual information, the right-hand side of (5) is equal to

$$\lambda I_\alpha(P_1; W) + (1-\lambda) I_\alpha(P_2; W). \tag{6}$$

As is proved in Theorem 1, (6) is minimized, over the choices of P_1, P_2 (or, equivalently, of the subsets A_1, A_2) only if P_1, P_2 have disjoint supports (equivalently, A_1, A_2 do not contain any overlapping class-labels). This establishes the appropriateness of the splitting criterion.

We now state and prove the main technical result of this paper.

Theorem 1. *Let P_Y be a pmf on $\mathcal{Y}$ with $|\mathcal{S}(P_Y)| \geq 2$, and $W(x|y)$ a $\mathcal{Y} \times \mathcal{X}$ conditional probability matrix, such that $J(Q; W)$ is a strictly concave function of Q. Let*

$$\mathcal{L}(P_Y) = \{(Q_1, Q_2) : Q_1 \neq Q_2 \text{ and } \theta Q_1 + (1-\theta)Q_2 = P_Y \text{ for some } \theta \in (0,1)\}$$

denote the set of pairs of probability measures whose linear interpolation equals P_Y. Finally, let

$$\mathcal{M}(P_Y, W) = \arg \min_{(Q_1,Q_2)\in\mathcal{L}(P_Y)} \{\theta^* J(Q_1; W) + (1-\theta^*)J(Q_2; W)\}, \tag{7}$$

denote pairs of probability measures that minimize the weighted sum in (7), where $\theta^* = \theta^*(P_Y, Q_1, Q_2)$ uniquely satisfies $\theta^* Q_1 + (1-\theta^*)Q_2 = P_Y$. Then,

$$\mathcal{S}(P_1) \cap \mathcal{S}(P_2) = \emptyset \quad \forall (P_1, P_2) \in \mathcal{M}(P_Y, W).$$

More informally, the minimum of the weighted sum of concave functions $J(\cdot;\cdot)$ in (7), over pairs of pmfs in $\mathcal{L}$, is attained by a pair of pmfs with disjoint supports.

Remark 1. The theorem holds for the family of strictly concave functions of pairs of distributions, of which the order-α mutual information is a member.

Remark 2. The minimization of mutual information is usually encountered in rate-distortion problems (cf Cover and Thomas (1991), page 342), where the minimization is over $W_{X|Y}$. In the usual channel coding scenario on the other hand (Cover and Thomas (1991), page 184), $W_{X|Y}$ is given, and one needs to *maximize* mutual information by choosing a suitable P_Y. Our result, in contrast, addresses the constrained *minimization* of mutual information, given $W_{X|Y}$.

Remark 3. We assume without loss of generality that $\mathcal{S}(P_Y) = \mathcal{Y}$; i.e. $P_Y(y) > 0 \; \forall \, y \in \mathcal{Y}$. Indeed, if $P_Y(y_0) = 0$ for some $y_0 \in \mathcal{Y}$, then one may exclude y_0 from the definition of $\mathcal{Y}$. This does not affect any members of $\mathcal{L}$, since $P_Y(y_0) = 0$ implies that $Q_1(y_0) = Q_2(y_0) = 0$ as well. Furthermore, excluding y_0 from $\mathcal{Y}$ does not affect the value of θ^* or the J's in (7). Finally, since $y_0 \notin \mathcal{S}(P_1)$ and $y_0 \notin \mathcal{S}(P_2)$ for every $(P_1, P_2) \in \mathcal{M} \subset \mathcal{L}$, the removal of y_0 from $\mathcal{Y}$ does not affect the claimed result.

Consider the objective function $I(Q_1, Q_2) : \mathcal{L}(P_Y) \to \mathbb{R}_+$ whose minimizers constitute the set $\mathcal{M}$ of (7):

$$I(Q_1, Q_2) = \theta^* J(Q_1; W) + (1-\theta^*)J(Q_2; W),$$

where the dependence of θ^* on Q_1, Q_2 and P_Y is suppressed for brevity of notation. Note that $\mathcal{L}(P_Y)$ *does* contain pairs of pmfs with disjoint supports. This is guaranteed by $|\mathcal{S}(P_Y)| \geq 2$, and construction of such a pair of distributions will be demonstrated shortly. On the other hand, it should be obvious that $\mathcal{L}(P_Y)$ also contains pairs of pmfs without disjoint supports.

Proof. Consider the set of pairs of distributions

$$\mathcal{L}(P_Y, \lambda) = \{(Q_1, Q_2) : \lambda Q_1 + (1-\lambda)Q_2 = P_Y\},$$

where $\lambda \in (0,1)$. It can be easily established that $\mathcal{L}(P_Y, \lambda)$ is a closed, convex set. For any $\lambda \in (0,1)$, $I(Q_1, Q_2)$ is strictly concave on $\mathcal{L}(P_Y, \lambda)$, since it is the sum of two strictly concave functions. Hence, the minimum of $I(Q_1, Q_2)$ over $\mathcal{L}(P_Y, \lambda)$ is attained at an extremal point of $\mathcal{L}(P_Y, \lambda)$. (Note that an extremal point of a closed, convex set S is any point which cannot be written as a convex combination of points in S.) On the other hand, the set

$$\mathcal{L}(P_Y) = \bigcup_{\lambda\in(0,1)} \mathcal{L}(P_Y, \lambda)$$

is not necessarily convex,[1] hence, we cannot immediately say that the minimum of $I(Q_1, Q_2)$ over $\mathcal{L}(P_Y)$ is achieved at an extremal point of $\mathcal{L}(P_Y)$. We now prove the following lemma:

Lemma 1. *For any $\lambda \in (0,1)$, if (P, Q) is an extremal point of $\mathcal{L}(P_Y, \lambda)$, then P, Q have disjoint supports except for possibly one coordinate, i.e., they are of the form (modulo a permutation of the indices)*

$$P = \begin{bmatrix} p & \boldsymbol{u}_1 & \mathbf{0} \end{bmatrix}$$
$$Q = \begin{bmatrix} q & \mathbf{0} & \boldsymbol{v}_2 \end{bmatrix}$$

for some scalars $p, q \geq 0$ and subvectors $\boldsymbol{u}_1$ and $\boldsymbol{v}_2$, where $\mathbf{0}$ is the zero vector of appropriate length.

Proof. Assume to the contrary that $\mathcal{S}(P) \cap \mathcal{S}(Q)$ contains more than one point. Without loss of generality, let

$$P = \begin{bmatrix} p, & x, & \boldsymbol{v_1} \end{bmatrix}, \qquad Q = \begin{bmatrix} q, & y, & \boldsymbol{v_2} \end{bmatrix},$$

where $x, y > 0$ and $\boldsymbol{v_1}, \boldsymbol{v_2}$ are vectors of equal length (i.e., of length $|\mathcal{Y}| - 2$). Then, for a sufficiently small $\epsilon > 0$, we have

$$(P, Q) = \frac{1}{2}(P_1, Q_1) + \frac{1}{2}(P_2, Q_2),$$

where $(P_1, Q_1), (P_2, Q_2) \in \mathcal{L}(P_Y, \lambda)$ and

$$P_1 = \left[p - \frac{\epsilon}{\lambda}, \quad x + \frac{\epsilon}{\lambda}, \quad \boldsymbol{v_1} \right], \qquad Q_1 = \left[q + \frac{\epsilon}{1-\lambda}, \quad y - \frac{\epsilon}{1-\lambda}, \quad \boldsymbol{v_2} \right],$$
$$P_2 = \left[p + \frac{\epsilon}{\lambda}, \quad x - \frac{\epsilon}{\lambda}, \quad \boldsymbol{v_1} \right], \qquad Q_2 = \left[q - \frac{\epsilon}{1-\lambda}, \quad y + \frac{\epsilon}{1-\lambda}, \quad \boldsymbol{v_2} \right].$$

Hence, (P, Q) cannot be an extremal point of $\mathcal{L}(P_Y, \lambda)$ (contradiction). ■

We will prove the theorem by contradiction. Without loss of generality, P_Y can be written as

$$P_Y = \begin{bmatrix} p, & \boldsymbol{u}, & \boldsymbol{v} \end{bmatrix},$$

where $p > 0$ is a scalar, and $\boldsymbol{u} = [u_1, \ldots, u_{|\boldsymbol{u}|}], \boldsymbol{v} = [v_1, \ldots, v_{|\boldsymbol{v}|}]$ are appropriate vectors whose lengths are upper bounded by $|\mathcal{Y}|$. Obviously, $\sum_j u_j + \sum_k v_k = 1 - p$.

Let (Q^*, R^*) be a member of $\mathcal{M}(P_Y, W)$ (i.e., it is a minimizer of $I(\cdot,\cdot)$), such that Q^*, R^* do *not* have disjoint supports. Then, $(Q^*, R^*) \in \mathcal{L}(P_Y, \theta^*)$ has to be an extremal point of $\mathcal{L}(P_Y, \theta^*)$ (otherwise, it would not be a minimizer). Then, by virtue of Lemma 1, we know that Q^*, R^* have one common non-zero coordinate. Without loss of generality, we have

$$Q^* = \left[q, \quad \frac{1-q}{\sum_j u_j} \boldsymbol{u}, \quad \mathbf{0} \right],$$
$$R^* = \left[r, \quad \mathbf{0}, \quad \frac{1-r}{\sum_k v_k} \boldsymbol{v} \right],$$

for some scalars $q, r > 0$, where

$$\theta^* = \frac{\sum_j u_j}{1-q}, \quad \text{and} \quad 1 - \theta^* = \frac{\sum_k v_k}{1-r}, \tag{8}$$

where $\theta^* \in (0,1)$ by the assumption that $(Q^*, R^*) \in \mathcal{M}(P_Y, W)$. We now consider the following pairs of distributions with disjoint supports:

$$Q_0 = \left[0, \quad \frac{1}{\sum_j u_j} \boldsymbol{u}, \quad \mathbf{0} \right],$$

[1] Consider, for example, the case where $|\mathcal{Y}| = 2$, with $P_Y = (1/3, 2/3)$. Then, $(P_1, Q_1) = ((1/2, 1/2), (0, 1)) \in \mathcal{L}(P_Y)$ and $(P_2, Q_2) = ((1/4, 3/4), (1, 0)) \in \mathcal{L}(P_Y)$. But $1/2(P_1, Q_1) + 1/2(P_2, Q_2) \notin \mathcal{L}(P_Y)$.

$$R_0 = \left[\frac{p}{1-\sum_j u_j}, \quad \mathbf{0} \quad \frac{1-\frac{p}{1-\sum_j u_j}}{\sum_k v_k}\boldsymbol{v}, \right],$$

and

$$Q_1 = \left[\frac{p}{1-\sum_k v_k}, \quad \frac{1-\frac{p}{1-\sum_k v_k}}{\sum_j u_j}\boldsymbol{u}, \quad \mathbf{0} \right],$$

$$R_1 = \left[0, \quad \mathbf{0}, \quad \frac{1}{\sum_k v_k}\boldsymbol{v} \right].$$

It can be easily established that

$$(Q_0, R_0) \in \mathcal{L}(P_Y, \theta_0), \qquad \theta_0 = \sum_j u_j, \tag{9}$$

$$(Q_1, R_1) \in \mathcal{L}(P_Y, \theta_1), \qquad \theta_1 = 1 - \sum_k v_k. \tag{10}$$

Furthermore,

$$Q^* = \beta Q_0 + (1-\beta)Q_1, \qquad \beta = 1 - \frac{q\left(1-\sum_k v_k\right)}{p}$$

$$R^* = \gamma R_0 + (1-\gamma)R_1, \qquad \gamma = \frac{r\left(1-\sum_j u_j\right)}{p}.$$

Now, because of the strict concavity of $J(Q; W)$, we have

$$\begin{aligned} I(Q^*, R^*) &= \theta^* J(Q^*; W) + (1-\theta^*)J(R^*; W) \\ &> \theta^*(\beta J(Q_0; W) + (1-\beta)J(Q_1; W)) + (1-\theta^*)(\gamma J(R_0; W) + (1-\gamma)J(R_1; W)). \end{aligned} \tag{11}$$

Now, using the fact that $p + \sum_j u_j + \sum_k v_k = 1$, it can be easily shown that

$$\begin{aligned} \theta^*\beta = \lambda\theta_0, \qquad & (1-\theta^*)\gamma = \lambda(1-\theta_0) \\ \theta^*(1-\beta) = (1-\lambda)\theta_1, \qquad & (1-\theta^*)(1-\gamma) = (1-\lambda)(1-\theta_1), \end{aligned}$$

where

$$\lambda = \frac{p - q + q\sum_j v_j}{p(1-q)} = \frac{r\sum_j v_j}{(1-r)p}. \tag{12}$$

Hence, the right-hand side of (11) is equal to

$$\begin{aligned} &\lambda(\theta_0 J(Q_0; W) + (1-\theta_0)J(R_0; W)) + (1-\lambda)(\theta_1 J(Q_1; W) + (1-\theta_1)J(R_1; W)) \\ &\quad = \lambda I(Q_0, R_0) + (1-\lambda)I(Q_1, R_1). \end{aligned} \tag{13}$$

But (9)–(11) and (13) imply that there are pairs $(Q_0, R_0), (Q_1, R_1) \in \mathcal{L}(P_Y)$ with disjoint supports, for which $I(Q^*, R^*) > \lambda I(Q_0, R_0) + (1-\lambda)I(Q_1, R_1)$, for some $\lambda \in (0, 1)$. Therefore, $I(Q^*, R^*) > \min\{I(Q_0, R_0), I(Q_1, R_1)\}$ (contradiction, because $I(Q^*, R^*)$ is minimum in $\mathcal{L}(P_Y)$).

Hence, $\mathcal{M}(P_Y, W)$ contains only pairs of distributions with disjoint supports. ■

Acknowledgment

The authors would like to thank Prof. Prakash Narayan for his very useful comments.

References

Beran, R., 1977. Minimum Hellinger distance estimates for parametric models. Ann. Statist. (5), 445–465.

Chou, P.A., 1991. Optimal partitioning for classification and regression trees. IEEE Trans. Pattern Anal. Mach. Intell. 13 (4), 340–354.

Cover, T., Thomas, J., 1991. Elements of Information Theory. John Wiley and Sons.

Csiszár, I., 1995. Generalized cutoff rates and Rényi's information measures. IEEE Trans. Inform. Theory 41 (1), 26–34.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B (Methodological) 39, 1–38.

Fraley, C., Raftery, A., 1999. Mclust: Software for model-based cluster analysis. J. Classification 16, 297–306.

Jelinek, F., 1997. Statistical Methods for Speech Recognition. MIT Press.

Karakos, D., Khudanpur, S., Eisner, J., Priebe, C.E., 2005. Unsupervised classification via decision trees: An information-theoretic perspective. In: Proc. 2005 International Conference on Acoustics, Speech and Signal Processing. ICASSP 2005.

Priebe, C.E., Marchette, D., Healy, D., 2004. Integrated sensing and processing decision trees. IEEE Trans. Pattern Anal. Mach. Intell. 26 (6), 699–708.

Rényi, A., 1961. On measures of entropy and information. In: Proc. 4th Berkeley Symposium on Math. Statist. Probability. vol. 1. pp. 547–561.