Semi–Supervised Probabilistic Distance Clustering and the Uncertainty of Classification

Cem Iyigun[1,⋆] and Adi Ben-Israel[1]

[1]RUTCOR–Rutgers Center for Operations Research, Rutgers University, 640 Bartholomew Rd., Piscataway, NJ 08854-8003, USA. {ciyigun,adi.benisrael}@gmail.com

Summary. Semi–supervised clustering is an attempt to reconcile **clustering** (unsupervised learning) and **classification** (supervised learning, using prior information on the data.) These two modes of data analysis are combined in a parameterized model, the parameter $\theta \in [0, 1]$ is the weight attributed to the prior information, $\theta = 0$ corresponding to clustering, and $\theta = 1$ to classification. The results (cluster centers, classification rule) depend on the parameter θ, an insensitivity to θ indicates that the prior information is in agreement with the intrinsic cluster structure, and is otherwise redundant. This explains why some data sets (such as the Wisconsin breast cancer data [23]) give good results for all reasonable classification methods. The **uncertainty of classification** is represented here by the geometric mean of the membership probabilities, shown to be an entropic distance related to the Kullback–Leibler divergence.

Key words: Clustering, classification, probabilistic clustering, semi–supervised learning, contour approximation of data, home range, classification uncertainty, entropy, Kullback–Leibler divergence, breast cancer data, diabetes data

1 Introduction

1.1 Clustering

A **cluster** is a set of elements that are similar in some sense. In what follows, data points are considered as points in a metric space with a distance function d, and "similar" is taken as "close", the data points $\mathbf{x}, \mathbf{y}$ are similar if $d(\mathbf{x}, \mathbf{y})$ is small. **Clustering**, the process of identifying clusters with dissimilar elements in different clusters, is modeled here as an optimization problem

$$\min_{\mathbf{p},\,\mathbf{c}} F(\mathbf{p}, \mathbf{c}), \tag{P.0}$$

with an objective function F, and variables $\mathbf{p}$ (**probabilities**) and $\mathbf{c}$ (**centers**) that are explained below.

1.2 Classification

Here a population, or a data set, $\mathcal{D}$ is partitioned into several disjoint classes, but the class to which an element belongs may be unknown, and needs to be determined. Data points are of the form $(\mathbf{x}, y)$, where $\mathbf{x}$ is a vector of **observations**, and the **label** y is the (usually unknown) **class** where $\mathbf{x}$ belongs. A **classification rule** is a function that assigns class values y to observations $\mathbf{x}$ from $\mathcal{D}$. It is learned from sample data for which the class labels are known (the **prior information**.) **Classification** is the process of deriving such a classification rule.

A common protocol is to learn the classification rule from a randomly selected subset $\mathcal{T}$ of $\mathcal{D}$ (the **training set**), then test it on the remaining data $\mathcal{D} \setminus \mathcal{T}$ (the **testing set**), recording the percentage of correct classifications as a performance criterion of the rule.

We model classification as an optimization problem

$$\min_{\mathbf{p},\,\mathbf{c}} G(\mathbf{p}, \mathbf{c}), \tag{P.1}$$

where the prior information on the data is incorporated in the objective function G.

⋆ Research by the first author was partially supported by DIMACS Summer Grant-2007.

Example 1. **Medical diagnostics**. Here $\mathcal{D}$ is a medical data set with data points $(\mathbf{x}, y)$, $\mathbf{x}$ the vector of test results for a given patient (abbreviated the **patient**), and y the **medical status**, say $y = 1$ if disease is present, $y = 0$ otherwise (two classes, regardless of the intrinsic structure of the data.) A classification rule $\eta(\cdot)$ may result in error, say $\eta(\mathbf{x}) = 0$ for a patient $\mathbf{x}$ with disease (**false negative**), or $\eta(\mathbf{x}) = 1$ for a patient $\mathbf{x}$ that is disease free (**false positive**). In general, these two classification errors have different consequences, and one may want to reduce one at the cost of increasing the other.

A well known medical data set, the **Wisconsin breast cancer data set** [23], has attracted much research, see, e.g., [22], [26]. Interestingly, all classification methods tried on this data set gave good results, see, e.g., [20]. This is explained in Example 3 below.

1.3 Learning

Classification is also called **supervised learning** to indicate that prior information is available. In contrast, cluster analysis relies only on the intrinsic structure and geometry of the data, and is called **unsupervised learning**.

Another difference is that the number of classes is given in classification problems, while in clustering the "right" number of clusters for the data in question may not be given, and has to be determined.

1.4 Semi–supervised clustering

Semi–supervised clustering is an attempt to reconcile clustering and classification, two contrasting modes of data analysis. The semi–supervised clustering problem is modeled here as a parametric family of optimization problems, using a parameter $\theta \in [0, 1]$ that expresses the weight attributed to the prior information,

$$\min_{\mathbf{p},\,\mathbf{c}} \{(1-\theta)\, F(\mathbf{p}, \mathbf{c}) + \theta\, G(\mathbf{p}, \mathbf{c})\}. \tag{P.θ}$$

The clustering problem (P.0) and the classification problem (P.1) are special cases, for $\theta = 0$ and $\theta = 1$, respectively.

For other approaches to semi–supervised clustering see, e.g., [8], [11] and [15].

1.5 Matching labels

The optimal solutions of (P.θ) depend on θ, and are denoted $\{\mathbf{p}^*(\theta), \mathbf{c}^*(\theta)\}$. A data set with prior information is said to have **well–matching labels** if $\{\mathbf{p}^*(\theta), \mathbf{c}^*(\theta)\}$ are insensitive to θ. In this case, the prior information is in agreement with the intrinsic structure of the data set, and the clusters of the problem (P.0) may be used to derive the classification rule for the problem (P.1). In the opposite case, where $\{\mathbf{p}^*(\theta), \mathbf{c}^*(\theta)\}$ are sensitive to θ, the labels are said to be **ill–matching**.

1.6 Plan of this paper

Probabilistic distance clustering and **classification** (using prior information) are outlined in Sections 2 and 3, respectively. **Semi–supervised clustering** is introduced in § 4 as a parametric family of convex combinations of the clustering and classification problems, with a parameter θ indicating the importance placed on the prior information. The algorithm proposed in § 4.4 updates the cluster centers as convex combinations of the data points. Section 5 illustrates the dependence of the results on the parameter θ, for a synthetic example with ill–matching labels, Example 2, and two medical data sets, Example 3. The cluster probabilities are studied in Appendix A as functions of distances, justifying the probabilistic model of § 2.3. The **classification uncertainty function** of § 2.5 is shown in Appendix B to be an entropic distance, associated with the Kullback–Leibler relative entropy.

2 Probabilistic distance clustering

2.1 Notation

Let $\overline{1,n}$ denote the **index set** $\{1, 2, \cdots, n\}$, and $\forall$ the qualifier **for all**.

Consider a data set $\mathcal{D}$ with N points $\mathbf{x}_i$, $i \in \overline{1,N}$, and K clusters,

$$\mathcal{D} = \mathcal{C}_1 \cup \mathcal{C}_2 \cdots \cup \mathcal{C}_K, \quad \mathcal{C}_k \cap \mathcal{C}_\ell = \emptyset \text{ if } k \neq \ell\ .$$

The data points have n components (**attributes**), and are formally considered as elements of an n–dimensional real space $\mathbb{R}^n$ (although the vector sum of two data points is not necessarily a data point.)

The k^{th}–cluster $\mathcal{C}_k$ has a **center** $\mathbf{c}_k$ (to be computed), and a **distance function** $d_k(\,\cdot\,, \mathbf{c}_k)$, in particular the elliptic distance

$$d_k(\mathbf{x}, \mathbf{c}_k) := \langle (\mathbf{x} - \mathbf{c}_k), Q_k(\mathbf{x} - \mathbf{c}_k) \rangle^{1/2} \,, \tag{1}$$

where $\langle \mathbf{x}, \mathbf{y} \rangle$ is the standard inner product of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and the geometry of the cluster is modeled by the positive definite matrix matrix Q_k. We often use the **Mahalanobis distance**,

$$d_k(\mathbf{x}, \mathbf{c}_k) := \langle (\mathbf{x} - \mathbf{c}_k), \Sigma^{-1}(\mathbf{x} - \mathbf{c}_k) \rangle^{1/2} \,, \tag{2}$$

where Σ_k is the covariance matrix of $\mathcal{C}_k$, see, e.g., [3], [27].

The space $\mathbb{R}^n$ is thus endowed with K metrics. Both centers and distance functions are updated by the clustering algorithm, see § 2.8 below. We abbreviate $d_k(\mathbf{x}, \mathbf{c}_k)$ by $d_k(\mathbf{x})$.

2.2 Probabilistic clustering

In **probabilistic** (**fuzzy** or **soft**) **clustering** the assignment of points to clusters is not deterministic, and is given as probability, see, e.g., [7], [12]. Let $p_k(\mathbf{x})$ denote the **probability** that the point $\mathbf{x}$ belongs to the cluster $\mathcal{C}_k$. This notation allows for deterministic membership, expressed by $p_k(\mathbf{x}) = 1$. The function $p_k(\cdot)$ is also called the **membership function** of $\mathcal{C}_k$.

2.3 Probabilistic distance clustering

In **probabilistic distance clustering** the membership probabilities $\{p_k(\mathbf{x}) : k \in \overline{1,K}\}$ depend on the distances $d_k(\mathbf{x})$ to the clusters. A reasonable assumption is

membership in a cluster is more likely the closer it is (A)

see Appendix A for details. A simple way to model this assumption is

$$p_k(\mathbf{x})\, d_k(\mathbf{x}) = D(\mathbf{x}), \ \forall\ k \in \overline{1,K}, \tag{3}$$

for any $\mathbf{x}$, where the function $D(\cdot)$ is independent of the cluster.

There are other ways to model Assumption (A), but the simple model (3) works well enough for our purposes.

2.4 Probabilities and the joint distance function

From (3), and the fact that the probabilities $\{p_k(\mathbf{x})\}$ add to one, it follows that

$$p_k(\mathbf{x}) = \frac{\prod\limits_{j \neq k} d_j(\mathbf{x})}{\sum\limits_{i=1}^{K} \prod\limits_{j \neq i} d_j(\mathbf{x})}, \quad k \in \overline{1,K}, \quad \text{and} \quad D(\mathbf{x}) = \frac{\prod\limits_{j=1}^{K} d_j(\mathbf{x})}{\sum\limits_{i=1}^{K} \prod\limits_{j \neq i} d_j(\mathbf{x})} \,. \tag{4}$$

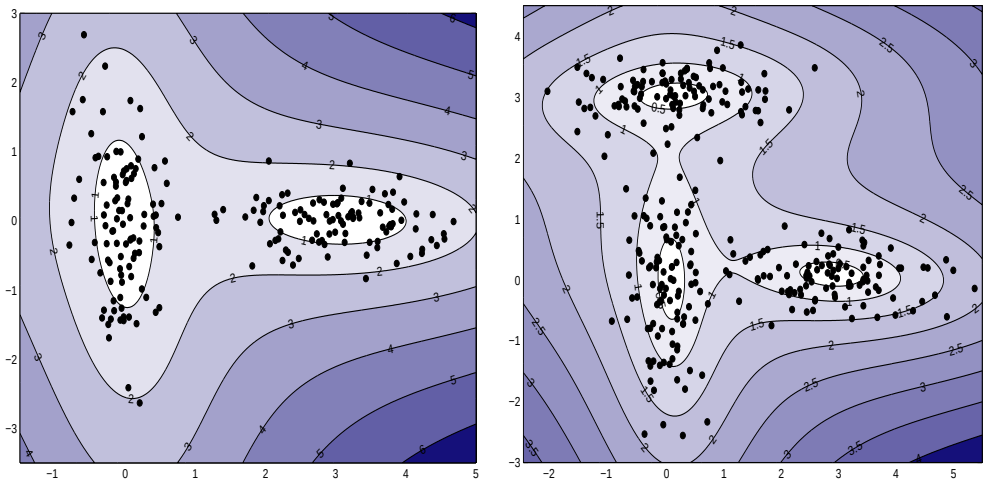
In particular, for $K = 2$,

$$p_1(\mathbf{x}) = \frac{d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})},\ p_2(\mathbf{x}) = \frac{d_1(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})}, \quad \text{and} \quad D(\mathbf{x}) = \frac{d_1(\mathbf{x}) d_2(\mathbf{x})}{d_1(\mathbf{x}) + d_2(\mathbf{x})} \,. \tag{5}$$

The function $D(\mathbf{x})$, called the **joint distance function** (abbreviated **JDF**) at $\mathbf{x}$, is (up to a constant) the harmonic mean of the distances $\{d_1(\mathbf{x}), \cdots, d_K(\mathbf{x})\}$. The JDF is a continuous function that captures the data points in its lower level sets, a property called **contour approximation**, see [2], [14]. Indeed, the geometry of each cluster is represented by its distance function (2), and the overall shape of the data set is given by the harmonic mean of these distances. The contour approximation of data by the JDF is illustrated in Fig. 1, for data sets with 2 and 3 clusters.

The JDF also gives a compact representation of the data in question: To represent a data set with N data points in $\mathbb{R}^n$, arranged in K clusters, the JDF requires K centers and K covariance matrices, a total of $K\, \frac{n(n+3)}{2}$ parameters, a considerable saving if $K \ll N$.

An ecological forerunner of the JDF and contour approximation is the **home range**, the territory of a species, given in [10] in terms of the harmonic mean of area moments, a finding confirmed since then for hundreds of species.



(a) A data set with 2 clusters

(b) A data set with 3 clusters

Fig. 1. The lower level sets of the JDF capture the data points

2.5 The classification uncertainty function

The JDF has the dimension of distance. Normalizing it, we get the dimensionless function

$$E(\mathbf{x}) = K\, D(\mathbf{x}) \,/\, \left(\prod_{j=1}^{K} d_j(\mathbf{x}) \right)^{1/K} , \tag{6}$$

with $0/0$ interpreted as zero. $E(\mathbf{x})$ is the harmonic mean of the distances $\{d_j(\mathbf{x})\}$ divided by their geometric mean. It follows that $0 \le E(\mathbf{x}) \le 1$, with $E(\mathbf{x}) = 0$ if any $d_j(\mathbf{x}) = 0$, i.e. if $\mathbf{x}$ is a cluster center, and $E(\mathbf{x}) = 1$ if and only if the probabilities $p_j(\mathbf{x})$ are all equal.

$E(\mathbf{x})$ can be written, using (4), as the geometric mean of the probabilities (up to a constant),

$$E(\mathbf{x}) = K \left(\prod_{j=1}^{K} p_j(\mathbf{x}) \right)^{1/K} . \tag{7}$$

In particular, for $K = 2$,

$$E(\mathbf{x}) = 2\, \frac{\sqrt{d_1(\mathbf{x}) d_2(\mathbf{x})}}{d_1(\mathbf{x}) + d_2(\mathbf{x})} = 2\, \sqrt{p_1(\mathbf{x}) p_2(\mathbf{x})} \, . \tag{8}$$

In the case $K = 1$, where the whole data set is taken as one cluster, we get formally from (7),

$$E(\mathbf{x}) = 1 \, . \tag{9}$$

The function $E(\mathbf{x})$ represents the **uncertainty** of classifying the point $\mathbf{x}$, see Appendix B. We call $E(\mathbf{x})$ the **classification uncertainty function**, abbreviated **CUF**, at $\mathbf{x}$.

The CUF of the data set $\mathcal{D} = \{\mathbf{x}_i : i \in \overline{1,N}\}$ is defined as

$$E(\mathcal{D}) := \frac{1}{N} \sum_{i=1}^{N} E(\mathbf{x}_i) \, . \tag{10}$$

$E(\mathcal{D})$ is a monotone decreasing function of K, the number of clusters, decreasing from $E(\mathcal{D}) = 1$ (for $K = 1$, see (9)), to $E(\mathcal{D}) = 0$ (for $K = N$, the trivial case where every data point is a separate cluster.) The rate of decrease of $E(\mathcal{D})$ is a natural criterion for determining the "right" number of clusters, if it is not given.

2.6 An extremum problem for the cluster probabilities at a point

Given the distances $\{d_k(\mathbf{x})\}$, and considering the probabilities $\{p_k\}$ as variables (abbreviating $p_k(\mathbf{x})$ by p_k), we note that (3) is the optimality condition of the **extremum problem**

$$\begin{aligned} &\text{Minimize} \quad \tfrac{1}{2}\sum_{k=1}^{K} d_k(\mathbf{x})\, p_k^2 \\ &\text{subject to} \quad \sum_{k=1}^{K} p_k = 1\ , \ \text{and } p_k \geq 0,\ k \in \overline{1,K}. \end{aligned} \tag{11}$$

Indeed, the **Lagrangian** of this problem is

$$L(p_1, \cdots, p_K, \lambda) = \tfrac{1}{2}\sum_{k=1}^{K} d_k(\mathbf{x})\, p_k^2 - \lambda\left(\sum_{k=1}^{K} p_k - 1\right)$$

and zeroing the partial derivatives (with respect to p_k) gives $p_k\, d_k(\mathbf{x}) = \lambda$, which is (3).

The squares of probabilities in (11) serve to smooth the underlying optimization problem which is nonsmooth, see the seminal paper [24] for other smoothing schemes, and a modern optimization framework for clustering.

2.7 An extremum problem for clustering the data set

The optimization problem (P.0) of § 1.1, for clustering a data set $\mathcal{D} = \{\mathbf{x}_i : i \in \overline{1,N}\}$ into K clusters, is written in detail as

$$\begin{aligned} &\text{Minimize} \quad \tfrac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K} d_k(\mathbf{x}_i, \mathbf{c}_k)\, p_k(\mathbf{x}_i)^2 \\ &\text{subject to} \quad \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1\ ,\ p_k(\mathbf{x}_i) \geq 0\ , \quad \forall\ i \in \overline{1,N},\ k \in \overline{1,K}, \end{aligned} \tag{P.0}$$

with the probabilities $\mathbf{p}$ and centers $\mathbf{c}$ as variables.

2.8 An outline of the probabilistic distance clustering algorithm of [4]

The algorithm of [4] solves the above problem (P.0) by iteratively updating the probabilities and centers.

Given the data set $\mathcal{D}$ and the number K of clusters, the algorithm begins with K arbitrary centers $\mathbf{c}_k$. In each iteration, the probabilities are computed by (4), for the given centers and distances. Fixing these probabilities, the centers $\mathbf{c}_k$ are updated as convex combinations of the data points $\mathbf{x}_i$,

$$\mathbf{c}_k = \sum_{i=1}^{N} \lambda_{ki}\, \mathbf{x}_i, \ \text{with weights } \lambda_{ki} = \frac{\dfrac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i)}}{\sum\limits_{j=1}^{N} \dfrac{p_k(\mathbf{x}_j)^2}{d_k(\mathbf{x}_j)}}\ , \quad i \in \overline{1,N}. \tag{12}$$

The update (12) is obtained by differentiating the objective function in (P.0), and zeroing the gradient.

The iterations stop when the centers "stop moving".

If the Mahalanobis distance is used, the covariance matrices taken initially as $\Sigma_k = I$, and are recomputed using the current centers $\mathbf{c}_k$ and probabilities.

This algorithm of [4] was adapted to account for the cluster sizes in [13].

Notes

(a) Iteration (12) is a generalization to several centers of the Weiszfeld method [25] for solving the Fermat–Weber location problem, and can be used for solving multi–facility location problems.

(b) A theoretical issue is that the gradient of the objective function of (P.0) is undefined if one of the data points $\{\mathbf{x}_i\}$ coincides with one of the current centers $\{\mathbf{c}_k\}$. In this case the gradient can be modified, as in [16]–[17], to guarantee that the method converges for all but a denumerable set of initial centers.

(c) The algorithm of [4] is robust: cluster centers are insensitive to outliers, that are discounted because the weights in (12) are inversely proportional to the distances.

3 Prior information and classification

3.1 Probabilistic labels

Let $\mathcal{D}$ be a data set with N points $\{\mathbf{x}_i : i \in \overline{1,N}\}$, and K clusters $\{\mathcal{C}_k : k \in \overline{1,K}\}$.

We assume that prior information is given for each point $\mathbf{x}_i$, as **probabilities** $r_k(\mathbf{x}_i)$ that $\mathbf{x}_i$ belongs to $\mathcal{C}_k$, $k \in \overline{1,K}$. These allow for rigid constraints such as $r_2(\mathbf{x}_3) = 1 = r_2(\mathbf{x}_4)$, saying that $\mathbf{x}_3$ and $\mathbf{x}_4$ both belong to $\mathcal{C}_2$.

If the story ends here, the membership probabilities $p_k(\mathbf{x}_i)$ are taken equal to the probabilistic labels,

$$p_k(\mathbf{x}_i) = r_k(\mathbf{x}_i) \ , \ \forall \ k, i. \tag{13}$$

3.2 An extremum problem for classification

A (trivial) extremum problem resulting in (13) is

$$\begin{aligned} \text{Minimize} \quad & \tfrac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} d_k(\mathbf{x}_i, \mathbf{c}_k) \, (p_k(\mathbf{x}_i) - r_k(\mathbf{x}_i))^2 \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1 \ , \ p_k(\mathbf{x}_i) \geq 0 \ , \quad \forall \ i \in \overline{1,N}, \ k \in \overline{1,K}, \end{aligned} \tag{P.1}$$

which is taken as the problem (P.1) of § 1.2. The distances $d_k(\mathbf{x}_i, \mathbf{c}_k)$ in the objective function serve to give it a dimension of distance, which allows combining (P.1) and (P.0).

4 Semi–supervised distance clustering

4.1 An extremum problem for semi-supervised clustering

We propose combining the clustering and classification problems in a parametric model, using a parameter $\theta \in [0, 1]$ for the **weight** given to the prior information. The model uses an optimization problem that is a convex combination of (P.0) and (P.1),

$$\begin{aligned} \text{Minimize} \quad & \tfrac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} d_k(\mathbf{x}_i, \mathbf{c}_k) \left[(1-\theta) \, p_k(\mathbf{x}_i)^2 + \theta \, (p_k(\mathbf{x}_i) - r_k(\mathbf{x}_i))^2 \right] \\ \text{subject to} \quad & \sum_{k=1}^{K} p_k(\mathbf{x}_i) = 1 \ , \ p_k(\mathbf{x}_i) \geq 0 \ , \quad \forall \ i \in \overline{1,N}, \ k \in \overline{1,K}. \end{aligned} \tag{P.θ}$$

This formulation gives a continuum of problems, with the clustering problem (P.0) and the classification problem (P.1) as special cases.

For fixed centers $\{\mathbf{c}_k\}$ and distances $\{d_k(\mathbf{x}_i, \mathbf{c}_k)\}$, the problem (P.$\theta$) is separable, reducing to N problems, one for each point $\mathbf{x}_i, \ i \in \overline{1,N}$.

4.2 Probabilities

To simplify notation, consider the case of 2 clusters, and a single data point $\mathbf{x}$ (since the problem (P.θ) is separable.) The distances $d_k(\mathbf{x})$, probabilities $p_k(\mathbf{x})$ and labels $r_k(\mathbf{x})$ are abbreviated below by d_k, p_k and r_k respectively.

Given d_1, d_2 and r_1, r_2, the problem (P.θ) becomes,

$$\begin{aligned} \min_{p_1,p_2} \quad & \tfrac{1}{2} \left[(1-\theta) \left(d_1 \, p_1^2 + d_2 \, p_2^2 \right) + \theta \left(d_1 \, (p_1 - r_1)^2 + d_2 \, (p_2 - r_2)^2 \right) \right] \\ \text{s.t.} \quad & p_1 + p_2 = 1, \\ & p_1, \, p_2 \geq 0. \end{aligned} \tag{14}$$

The Lagrangian of this problem is

$$L(p_1, p_2, \lambda) = \tfrac{1}{2} \left[(1-\theta) \left(d_1 \, p_1^2 + d_2 \, p_2^2 \right) + \theta \left(d_1 \, (p_1 - r_1)^2 + d_2 \, (p_2 - r_2)^2 \right) \right] - \lambda \, (p_1 + p_2 - 1).$$

Zeroing the gradient (with respect to p_1, p_2), and using the fact that the probabilities add to one, we get

$$p_1 = (1-\theta) \frac{d_2}{d_1 + d_2} + \theta \, r_1 \ , \quad p_2 = (1-\theta) \frac{d_1}{d_1 + d_2} + \theta \, r_2 \ , \tag{15}$$

giving the probabilities as convex combinations of the clustering probabilities (5) and the labels r_1, r_2.

4.3 Cluster centers

Given a data set $\mathcal{D} = \{\mathbf{x}_i : i \in \overline{1,N}\}$, identified for simplicity with the training set, and fixing the probabilities $(p_1(\mathbf{x}_i), p_2(\mathbf{x}_i)$ as in (15)), the extremal problem (P.θ) is

$$\min_{\mathbf{c}_1,\mathbf{c}_2} \tfrac{1}{2}\Big[(1-\theta)\sum_{i=1}^{N}\Big(d_1(\mathbf{x}_i,\mathbf{c}_1)\,p_1(\mathbf{x}_i)^2 + d_2(\mathbf{x}_i,\mathbf{c}_2)\,p_2(\mathbf{x}_i)^2\Big) + \\ + \theta\sum_{i=1}^{N}\Big(d_1(\mathbf{x}_i,\mathbf{c}_1)\,(p_1(\mathbf{x}_i)-r_1(\mathbf{x}_i))^2 + d_2(\mathbf{x}_i,\mathbf{c}_2)\,(p_2(\mathbf{x}_i)-r_2(\mathbf{x}_i))^2\Big)\Big] \quad (16)$$

For the elliptic distance (1), the gradient of the objective function in (16) w.r.t. $\mathbf{c}_1$ is

$$-\nabla_{\mathbf{c}_1} = \tfrac{1}{2}\Big[(1-\theta)\sum_{i=1}^{N} p_1(\mathbf{x}_i)^2\,\frac{Q_1(\mathbf{x}_i-\mathbf{c}_1)}{d_1(\mathbf{x}_i,\mathbf{c}_1)} + \theta\sum_{i=1}^{N}(p_1(\mathbf{x}_i)-r_1(\mathbf{x}_i))^2\,\frac{Q_1(\mathbf{x}_i-\mathbf{c}_1)}{d_1(\mathbf{x}_i,\mathbf{c}_1)}\Big]$$

Zeroing the gradient, and canceling the nonsingular matrix Q_1, we can express the center $\mathbf{c}_1$ as a convex combination of the data points $\{\mathbf{x}_i : i \in \overline{1,N}\}$. Repeating for the center $\mathbf{c}_2$, we can summarize

$$\mathbf{c}_k = \sum_{i=1}^{N}\lambda_{ki}\,\mathbf{x}_i\ ,\ k=1,2\ , \quad (17)$$

where the weights λ_{ki} are given by,

$$\lambda_{ki} = \frac{u_k(\mathbf{x}_i)}{\sum\limits_{j=1}^{N} u_k(\mathbf{x}_j)}\ ,\ \text{with } u_k(\mathbf{x}_i) = (1-\theta)\frac{p_k(\mathbf{x}_i)^2}{d_k(\mathbf{x}_i,\mathbf{c}_k)} + \theta\,\frac{(p_k(\mathbf{x}_i)-r_k(\mathbf{x}_i))^2}{d_k(\mathbf{x}_i,\mathbf{c}_k)}\ ,\ k=1,2\ . \quad (18)$$

The coefficients $u_k(\mathbf{x}_i)$ in (18) depend on the parameter θ. The limits of the coefficient $u_1(\mathbf{x}_i)$ in the extreme cases $\theta = 0$ and 1 are

$$u_1(\mathbf{x}_i) = \begin{cases} \dfrac{1}{d_1}\left(\dfrac{d_2}{d_1+d_2}\right)^2\ ,\ \theta = 0\ , \\ \left(\dfrac{r_1^2}{d_1}\right)\qquad\quad\ ,\ \theta \to 1\ . \end{cases}$$

Analogous results apply to the coefficient $u_2(\mathbf{x}_i)$.

4.4 Algorithm

The above ideas are implemented in an algorithm for semi–supervised distance clustering of data. A schematic description, presented – for simplicity – for the case of 2 clusters, follows.

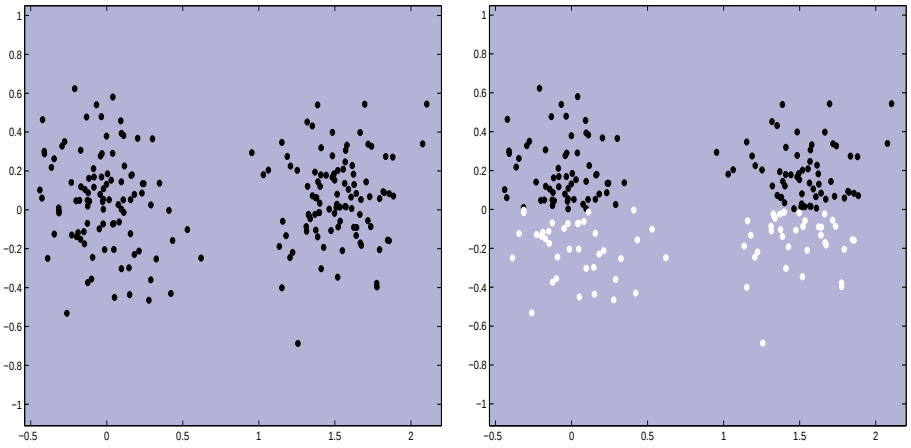
Algorithm 1 ***Semi-supervised distance clustering***

Initialization*: given data $\mathcal{D}$, any two points $\mathbf{c}_1, \mathbf{c}_2$, covariances $\Sigma_1 = \Sigma_2 = I$, a value θ, and $\epsilon > 0$*	
Iteration*:*	
Step 1	***compute*** *distances $d_1(\mathbf{x})$, $d_2(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{D}$*
Step 2	***compute*** *probabilities $p_1(\mathbf{x})$, $p_2(\mathbf{x})$, using (15) for all $\mathbf{x} \in \mathcal{D}$*
Step 3	***update*** *the centers $\mathbf{c}_1^+, \mathbf{c}_2^+$, using (17)–(18)*
Step 4	***compute*** *the cluster covariances Σ_1, Σ_2 using the current centers and probabilities*
Step 5	***if*** $\|\mathbf{c}_1^+ - \mathbf{c}_1\| + \|\mathbf{c}_2^+ - \mathbf{c}_2\| < \epsilon$ ***stop*** ***return*** *to step 1*

The algorithm solves the problem (P.θ) and reduces for $\theta = 0$ to the probabilistic distance clustering algorithm of [4]. Step 4 is needed if Mahalanobis distances are used, and is absent otherwise.

5 Examples

Recall that a data set has well–matching labels if the results of clustering are insensitive to the parameter θ, and ill–matching labels otherwise. We illustrate this for a synthetic data set, Example 2, with ill–matching labels, and 2 medical data sets in Example 3.



(a) The data points (b) The data points and their labels

Fig. 2. Illustration of Example 2

Example 2. Figure 2(a) shows a data set $\mathcal{D}$ in $\mathbb{R}^2$ with $N = 200$ data points in two equal clusters. The labels of these points are shown in different colors in Figure 2(b). These labels are clearly in conflict with the intrinsic clusters.

Figures 3(a)–(c) show level sets of the CUF $E(\mathbf{x})$ of (8) for different values of θ. These were computed using results obtained by Algorithm 1. Darker colors indicate higher values of $E(\mathbf{x})$, and greater uncertainty in classification.

For $\theta = 0$, see Figure 3(a), the labels are ignored, and the data set is partitioned following its intrinsic clusters as in Figure 2(a). The vertical white line in the center is the locus of equal probabilities $p_1(\mathbf{x}) = p_2(\mathbf{x}) = 0.5$. This line, that coincides here with the Fisher linear discriminant, can serve as a classification rule for assigning points to the two intrinsic clusters.

For $\theta = 0.25$, see Figure 3(b), the level sets of $E(\mathbf{x})$ are evolved to take account of the prior information. The locus of equal probabilities $p_1(\mathbf{x}) = p_2(\mathbf{x})$, which can serve as a classification rule, is again shown in white. Figure 3(c) shows the level sets of $E(\mathbf{x})$ for $\theta = 1$, i.e. where only the prior information is considered [2]. The equiprobability locus is here the horizontal white line, contrasting with the vertical line in Figure 3(a).

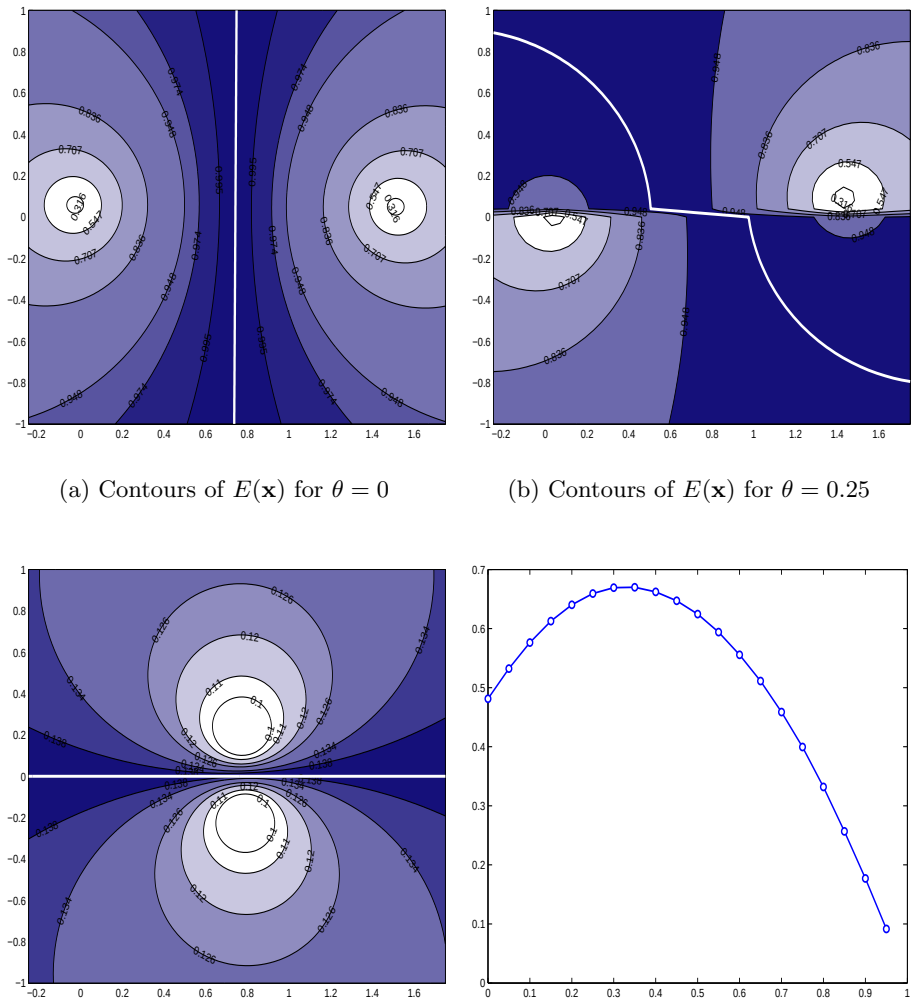
Figure 3(d) displays $E(\mathcal{D})$, the CUF of the data set $\mathcal{D} = \{\mathbf{x}_i : i \in \overline{1,N}\}$, see (10), for different values of θ. For $\theta = 1$ the uncertainty is zero, since the probabilities are given by the binary labels, see (15). That $E(\mathcal{D})$ does not decrease monotonically as θ increases, is due to the conflict between the intrinsic clusters in Fig. 2(a), and the prior information in Fig. 2(b). A mixture of these two models may have greater uncertainty than the "pure" models (P.0) and (P.1).

Example 3. We consider 2 well known data sets, given in [23]. For each data set, the cluster centers and classification rules were computed for different values of θ, and the percentages of correct classifications are plotted in Figures 4(a)–(b) (the thick curves.) The thin curves are the graphs of the CUF $E(\mathcal{D})$, which decreases to zero as $\theta \to 1$.

Figure 4(a) concerns the Wisconsin breast cancer data set, shown to have well–matching labels (the percentage of correct classifications is insensitive to θ.) This set would be clustered correctly even without the prior information, that is needed only to put the right labels on the clusters. This explains why all 33 methods reported in [20] give excellent results for this set. The CUF $E(\mathcal{D})$ is monotone decreasing since there is no conflict between the intrinsic clusters and the labels.

Figure 4(b) illustrates the diabetes data set, [23], shown to have ill–matching labels. The percentages of correct classifications are sensitive to the parameter θ, and the CUF $E(\mathcal{D})$ is non–monotonic.

[2] The level sets shown have low values of $E(\mathbf{x})$, reflecting no uncertainty of classification, and the colors would all be white or near white if the color scale was the same as in the previous figures.



(a) Contours of $E(\mathbf{x})$ for $\theta = 0$

(b) Contours of $E(\mathbf{x})$ for $\theta = 0.25$

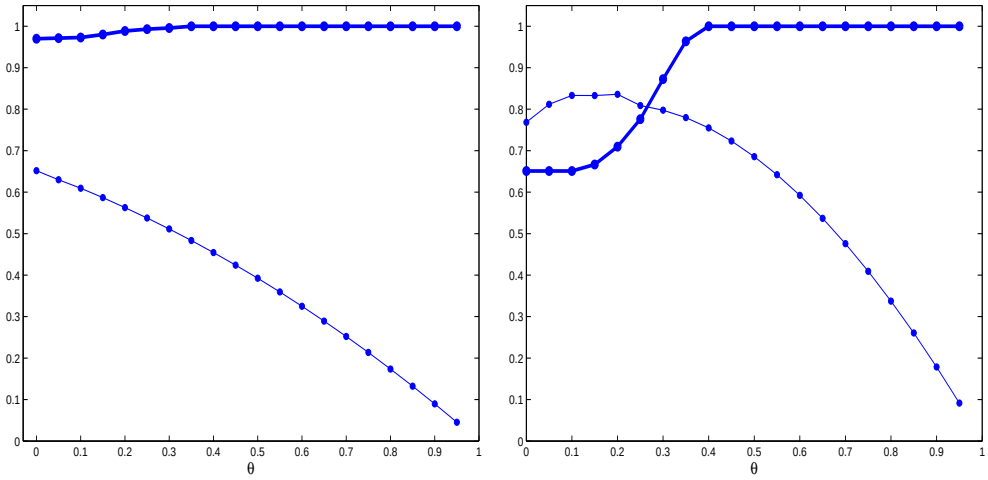
(c) Contours of $E(\mathbf{x})$ for $\theta = 1.00$

(d) $E(\mathcal{D})$ as a function of θ

Fig. 3. The CUF of Example 2 for different θ values

References

1. J. Aczél. Measuring information beyond communication theory - why some generalized information measures may be useful, others not. *Aequationes Mathematicae*, 27:1–19, 1984.
2. M. Arav. Contour approximation of data and the harmonic mean. *J. of Mathematical Inequalities*, 2:161–167, 2008.
3. A. Bar–Hillel, T. Hertz, N. Shental and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *J. of Machine Learning Research*, 6:937–965, 2005.
4. A. Ben–Israel and C. Iyigun. Probabilistic distance clustering. *J. of Classification*, 25:5–26, 2008.
5. A. Ben–Tal and M. Teboulle. Penalty functions and duality in stochastic programming via ϕ-divergence functionals. *Math. Oper. Res.*, 12:224–240, 1987.
6. A. Ben–Tal, A. Ben–Israel and M. Teboulle. Certainty equivalents and information measures: Duality and extremal principles. *J. Math. Anal. Appl.*, 157:211–236, 1991.
7. J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, 1981.
8. O. Chapelle, B. Schölkopf and A. Zien, editors. *Semi–Supervised Learning*. MIT Press, Cambridge MA, 2006.
9. I. Csiszár. Information measures: A critical survey. *Trans. 7th Prague Conf. on Info. Th., Statist., Decis. Funct., Random Processes and 8th European Meeting of Statist.* volume B, pp. 73–86. Academia, Prague, 1978.
10. K.R. Dixon and J.A. Chapman. Harmonic mean measure of animal activity areas. *Ecology*, 61:1040–1044, 1980.
11. N. Grira, M. Crucianu, and N. Boujemaa. Unsupervised and semi–supervised clustering: A brief survey. In *A Review of Machine Learning Techniques for Processing Multimedia Content*. Report of the MUSCLE European Network of Excellence, 2005.
12. F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis*. Wiley, New York, 1999.



(a) Wisconsin breast cancer data (b) Diabetes data

Fig. 4. Examples of well–matching, and ill–matching, labels

13. C. Iyigun and A. Ben–Israel. Probabilistic distance clustering adjusted for cluster size. *Probability in the Engineering and Informational Sciences*, 22:1–19, 2008.
14. C. Iyigun and A. Ben–Israel. Contour approximation of data: The dual problem. *Linear Algebra and its Applications*, (to appear).
15. A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31:264-323, 1999.
16. H.W. Kuhn. On a pair of dual nonlinear programs. In J. Abadie, editor. *Methods of Nonlinear Programming*, pp. 38–54. North-Holland, Amsterdam, 1967.
17. H.W. Kuhn. A note on Fermat's problem. *Math. Programming*, 4:98–107, 1973.
18. S. Kullback. *Information Theory and Statistics*. J. Wiley, New York, 1959.
19. S. Kullback and R.A. Leibler. On information and sufficiency. *Annals Math. Statist.*, 22:79–86, 1951.
20. T-S. Lim, W-Y. Loh, and Y-S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms. *Machine Learning*, 40: 203–228, 2000.
21. R.D. Luce. *Individual Choice Behavior*. Wiley, New York, 1959.
22. O.L. Mangasarian, R. Setiono, and W.H. Wolberg. Pattern recognition via linear programming: theory and application to medical diagnosis. In T. Coleman and Y. Li, editors. *Large–Scale Numerical Optimization*, pages 22–30. SIAM Publications, Philadelphia, 1999.
23. C. Merz and P. Murphy. UCI Repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA., 1996. http://www.ics.uci.edu/ mlearn/MLRepository.html
24. M. Teboulle. A unified continuous optimization framework for center–based clustering methods. *J. of Machine Learning Research*, 8:65–102, 2007.
25. E. Weiszfeld. Sur le point par lequel la somme des distances de n points donnés est minimum. *Tohoku Math. J.*, 43:355–386, 1937.
26. W.H. Wolberg and O.L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences of U.S.A.* 87:9193–9196, 1990.
27. E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metrice learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003.
28. J.I. Yellott, Jr. Luce's Choice Axiom. In N.J. Smelser and P.B. Baltes, editors. *International Encyclopedia of the Social & Behavioral Sciences*, pp. 9094–9097. ISBN 0-08-043076-7, 2001.

Appendix A: The membership probabilities

In this appendix, $d_k(\mathbf{x})$ stands for $d_k(\mathbf{x}, \mathbf{c}_k)$, the distance of $\mathbf{x}$ to the center $\mathbf{c}_k$ of the k^{th}–cluster, $k \in \overline{1,K}$.

The **cluster membership probabilities** $\{p_k(\mathbf{x}) : k \in \overline{1,K}\}$ of a point $\mathbf{x}$ depend only on the **distances** $\{d_k(\mathbf{x}) : k \in \overline{1,K}\}$,

$$\mathbf{p}(\mathbf{x}) = \mathbf{f}(\mathbf{d}(\mathbf{x})) \tag{A-1}$$

where $\mathbf{p}(\mathbf{x}) \in \mathbb{R}^K$ is the vector of probabilities $(p_k(\mathbf{x}))$, and $\mathbf{d}(\mathbf{x})$ is the vector of distances $(d_k(\mathbf{x}))$. Natural assumptions for the relation (A-1) include

$$d_i(\mathbf{x}) < d_j(\mathbf{x}) \implies p_i(\mathbf{x}) > p_j(\mathbf{x}), \text{ for all } i,j \in \overline{1,K} \tag{A-2a}$$

$$\mathbf{f}(\lambda\,\mathbf{d}(\mathbf{x})) = \mathbf{f}(\mathbf{d}(\mathbf{x})), \text{ for any } \lambda > 0 \tag{A-2b}$$

$$Q\,\mathbf{p}(\mathbf{x}) = \mathbf{f}(Q\,\mathbf{d}(\mathbf{x})), \text{ for any permutation matrices } Q \tag{A-2c}$$

Condition (A-2a) states that membership in a cluster is more probable the closer it is, which is Assumption (A) of § 2.3. The meaning of (A-2b) is that the probabilities $p_k(\mathbf{x})$ do not depend on the scale of measurement, i.e., the function $\mathbf{f}$ is homogeneous of degree 0. It follows that the probabilities $p_k(\mathbf{x})$ depend only on the ratios of the distances $\{d_k(\mathbf{x}) : k \in \overline{1,K}\}$.

The symmetry of $\mathbf{f}$, expressed by (A-2c), guarantees for each $k \in \overline{1,K}$, that the probability $p_k(\mathbf{x})$ does not depend on the numbering of the other clusters.

Assuming continuity of $\mathbf{f}$ it follows from (A-2a) that

$$d_i(\mathbf{x}) = d_j(\mathbf{x}) \implies p_i(\mathbf{x}) = p_j(\mathbf{x}),$$

for any $i,j \in \overline{1,K}$. In particular, the probabilities $p_k(\mathbf{x})$ are all equal only if so are the distances $d_k(\mathbf{x})$.

For any nonempty subset $\mathcal{S} \subset \overline{1,K}$, let

$$p_{\mathcal{S}}(\mathbf{x}) = \sum_{s\in\mathcal{S}} p_s(\mathbf{x}),$$

the probability that $\mathbf{x}$ belongs to one of the clusters $\{\mathcal{C}_s : s \in \mathcal{S}\}$, and let $p_k(\mathbf{x}|\mathcal{S})$ denote the **conditional probability** that $\mathbf{x}$ belongs to the cluster $\mathcal{C}_k$, given that it belongs to one of the clusters $\{\mathcal{C}_s : s \in \mathcal{S}\}$.

Since the probabilities $p_k(\mathbf{x})$ depend only on the ratios of the distances $\{d_k(\mathbf{x}) : k \in \overline{1,K}\}$, and these ratios are unchanged in subsets $\mathcal{S}$ of the index set $\overline{1,K}$, it follows that for all $k \in \overline{1,K}$, $\emptyset \neq \mathcal{S} \subset \overline{1,K}$,

$$p_k(\mathbf{x}) = p_k(\mathbf{x}|\mathcal{S})\,p_{\mathcal{S}}(\mathbf{x}) \tag{A-3}$$

which is the **choice axiom** of Luce, [21, Axiom 1], and therefore, [28],

$$p_k(\mathbf{x}|\mathcal{S}) = \frac{v_k(\mathbf{x})}{\sum\limits_{s\in\mathcal{S}} v_s(\mathbf{x})} \tag{A-4}$$

where $v_k(\mathbf{x})$ is a scale function, in particular,

$$p_k(\mathbf{x}) = \frac{v_k(\mathbf{x})}{\sum\limits_{s\in\overline{1,K}} v_s(\mathbf{x})}\ . \tag{A-5}$$

Assuming $v_k(\mathbf{x}) \neq 0$ for all k, it follows that

$$p_k(\mathbf{x})v_k(\mathbf{x})^{-1} = \frac{1}{\sum\limits_{s\in\overline{1,K}} v_s(\mathbf{x})}, \tag{A-6}$$

where the right hand side is a function of $\mathbf{x}$, and does not depend on k.

Property (A-2a) implies that the function $v_k(\cdot)$ is monotone decreasing. A simple choice is

$$v_k(\mathbf{x}) = \frac{1}{d_k(\mathbf{x})}\ , \tag{A-7}$$

for which (A-6) gives

$$p_k(\mathbf{x})\,d_k(\mathbf{x}) = \frac{1}{\sum\limits_{s\in\overline{1,K}} \left(\dfrac{1}{d_s(\mathbf{x})}\right)} = D(\mathbf{x}), \tag{A-8}$$

in agreement with (3)–(4).

Appendix B: The classification uncertainty function

Let $\mathbb{P}^K$ be the set of K–dimensional probability vectors, denoted $\mathbf{p} = (p_i)$, $\mathbf{q} = (q_i)$. Given a convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$, the **Csiszár ϕ–divergence**, [9], defined by

$$I_\phi(\mathbf{p},\mathbf{q}) := \sum_{i=1}^{K} q_i\, \phi\left(\frac{p_i}{q_i}\right), \quad \text{with} \quad 0\,\phi\left(\frac{0}{0}\right) := 0\,, \tag{B-1}$$

is a distance function on $\mathbb{P}^K$, a generalized measure of entropy, [1], whose distance-like properties make it useful in stochastic optimization, [5], [6]. For the special case

$$\phi_{\mathrm{KL}}(t) := t\,\log t, \quad t > 0,$$

(B-1) gives

$$I_{\phi_{\mathrm{KL}}}(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{K} p_i\,\log\left(\frac{p_i}{q_i}\right),$$

the **Kullback–Leibler distance**, [18], [19]. Rewriting the CUF (7) as

$$E(\mathbf{x}) = \left(\prod_{j=1}^{K}\left(\frac{p_j(\mathbf{x})}{1/K}\right)\right)^{1/K}$$

and taking logarithms, we get

$$-\log E(\mathbf{x}) = \sum_{i\in\overline{1,K}} (1/K)\,\log\left(\frac{1/K}{p_i(x)}\right) = I_{\phi_{\mathrm{KL}}}\left(\mathbf{p}(\mathbf{x}),\tfrac{1}{K}\,\mathbf{1}\right), \tag{B-2}$$

the Kullback–Leibler distance between the distributions

$$\mathbf{p}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), \cdots, p_K(\mathbf{x})) \text{ and } \tfrac{1}{K}\,\mathbf{1} = (\tfrac{1}{K}, \tfrac{1}{K}, \cdots, \tfrac{1}{K}).$$

The latter distribution, $\frac{1}{K}\,\mathbf{1}$, is of maximal uncertainty in $\mathbb{P}^K$, and consequently the divergence $I_{\phi_{\mathrm{KL}}}\left(\mathbf{p}(\mathbf{x}),\frac{1}{K}\,\mathbf{1}\right)$ is a measure of the uncertainty of the distribution $\mathbf{p}(\mathbf{x})$, with smaller values corresponding to greater uncertainty.

Writing (B-2) as

$$E(\mathbf{x}) = \exp\left\{-I_{\phi_{\mathrm{KL}}}\left(\mathbf{p}(\mathbf{x}),\tfrac{1}{K}\,\mathbf{1}\right)\right\} \tag{B-3}$$

it follows that $E(\mathbf{x})$ is an entropic measure of the uncertainty of classification, a monotone increasing function of the uncertainty.