Journal of Machine Learning Research 16 (2015) 775-785 Submitted 12/13; Revised 10/14; Published 4/15

Strong Consistency of the Prototype Based Clustering in Probabilistic Space

Vladimir Nikulin VNIKULIN.UQ@GMAIL.COM
Department of Mathematical Methods in Economy
Vyatka State University, Kirov, 610000, Russia

Editor: Inderjit Dhillon

Abstract

In this paper we formulate in general terms an approach to prove strong consistency of the Empirical Risk Minimisation inductive principle applied to the prototype or distance based clustering. This approach was motivated by the Divisive Information-Theoretic Feature Clustering model in probabilistic space with Kullback-Leibler divergence, which may be regarded as a special case within the Clustering Minimisation framework.

Keywords: clustering, probabilistic space, consistency

1. Introduction

Clustering algorithms group objects into subsets (clusters) of similar items according to the given criteria. For example, it may be Spectral Clustering (Ng et al., 2001) or Prototype Based model (Hinneburg and Keim, 2003). Clustering has application in various areas of computer science such as machine learning, data compression, data mining or patterns recognition. Depending on the area of application, there are many different formulations of the clustering problem (Ackerman et al., 2008). For example, we can consider text document as an object with words as features, and the task is to cluster text documents into subsets, corresponding to a few given topics. This problem maybe effectively approximated by the clustering model in probabilistic space with Kullback Leibler (KL) divergence (Dhillon et al., 2003) which arises as a natural measure of the dissimilarity between two distributions in numerical way. Further related results are presented by Chaudhuri and McGregor (2008), where authors provide algorithms for clustering using the KL-divergence measure with an objective to achieve guaranteed approximation in the worst case.

In this paper we consider a prototype based approach which may be described as follows. Initially, we have to choose k prototypes. The corresponding empirical clusters will be defined in accordance to the criteria of the nearest prototype measured by the distance Φ. Respectively, we will generate initial k clusters. As a second Minimisation step, we shall recompute cluster centers or Φ-means (Cuesta-Albertos et al., 1997), using data strictly from the corresponding clusters. Then, we can repeat Clustering step, using new prototypes, obtained from the previous step as cluster centers. Above algorithm has descending property. Respectively, it will reach local minimum in a finite number of steps.

Stability is a common tool to verify the validity of sample based algorithms. Clustering is one of the most widely used techniques for exploratory data analysis. Across all disciplines, from social sciences over biology to computer science, people try to get a first intuition about their data by identifying meaningful groups among the data points. Despite this popularity of clustering, distressingly little is known about theoretical properties of clustering (Ben-David et al., 2006).

1.1 Related Work

One formulation of stability is: if parameters are learned over two different samples from the same distribution, how close they are? The statistical stability for clustering have been extensively studied (Rakhlin and Caponnetto, 2006; Shamir and Tishby, 2008).

Pollard (1981) demonstrated that the classical K-means algorithm in $\mathbb{R}^m$ with squared loss function satisfies the Key Theorem of Learning Theory (Vapnik, 1995), p.36, *"the minimal empirical risk must converge to the minimal actual risk"*. Note, also study (Biau et al., 2008), where the number of theorems that establish the universal consistency of averaging rules are given.

Telgarsky and Dasgupta (2013) constructed an explicit moment-based uniform deviation bounds for the convergence of the soft clustering processes in Euclidean space. The results of Telgarsky and Dasgupta (2013) are general and significant: assuming that some probability moments are limited, an explicit convergence bounds are constructed. Compared to the model of Pollard (1981), the framework presented by Telgarsky and Dasgupta (2013) represents a novel direction, and we are considering to extend it to the probabilistic space in our future work.

Note that both functions 1) squared loss and 2) KL divergence are covered by the general structural definition of *Bregman* divergence. Bregman divergences give us a lot of freedom in fitting the performance measure of our algorithm to the nature of the data, and, as a consequence, this will lead to qualitatively better clustering (Banerjee et al., 2005), (Nock et al., 2008). Bregman divergences have found many applications in the fields of machine learning and computational geometry.

A new clustering algorithm in probabilistic space $\mathcal{P}^m$ was proposed by Dhillon et al. (2003). It provides an attractive approach based on the Kullback-Leibler divergence. The above methodology requires a general formulation and framework which we present in the following Section 2.

There are many useful and popular models and algorithms in the field of machine learning in addition to clustering. Consistency of those models represents a very essential property which should be investigated. For example, the subject of the papers (Glasmachers, 2010), (McAllester and Keshet, 2011) is consistency of support vector classifiers. Also, it is very interesting to identify those models, which are not consistent (Long and Servedio, 2013), and explain the reasons for not consistency. Particularly interesting is to find general conditions under which the common approaches, with various algorithmic variations, are consistent (Kpotufe et al., 2014).

1.2 Structure of the Paper

The paper is organised as follows. Section 3 extends the methodology of Pollard (1981) in order to cover the case of $\mathcal{P}^m$ with Kullback-Leibler divergence. With the aim to highlight the most essential properties, we formulate the model in general terms, where the probabilistic space is considered as an important example. We do believe that the structural approach, which is formulated in Section 3, maybe useful for the consideration of other cases (different space or the same space with different loss function). In this sense our work is similar to Kpotufe et al. (2014). Using the results and definitions of the Section 3, we investigate relevant properties of $\mathcal{P}^m$ in the final Section 4 and prove a strong consistency of the Empirical Risk Minimisation inductive principle.

2. Prototype Based Approach

In this paper we consider a sample of i.i.d. observations $\mathbf{X} := \{x_1, \dots, x_n\}$ drawn from probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ where probability measure $\mathbb{P}$ is assumed to be unknown.

Key in this scenario is an encoding problem. Assuming that we have a codebook $\mathcal{Q} \in \mathcal{X}^k$ with *prototypes* $q(c)$ indexed by the code $c = 1, \dots, k$, the aim is to encode any $x \in \mathcal{X}$ by some $q(c(x))$ such that the distortion between x and $q(c(x))$ is minimized:

$$c(x) := \underset{c}{\operatorname{argmin}} \, \mathcal{L}(x, q(c)), \tag{1}$$

where $\mathcal{L}(\cdot, \cdot)$ is a loss function.

Using criterion (1) we split empirical data into k clusters. As a next step we compute the *cluster center* specifically for any particular cluster in order to minimise overall distortion error.

We estimate actual distortion error

$$\Re^{(k)}[\mathcal{Q}] := \mathbf{E} \;\; \mathcal{L}(x, \mathcal{Q}) \tag{2}$$

by the empirical error

$$\Re^{(k)}_{\text{emp}}[\mathcal{Q}] := \frac{1}{n} \sum_{t=1}^{n} \mathcal{L}(x_t, \mathcal{Q}), \tag{3}$$

where $\mathcal{L}(x, \mathcal{Q}) := \mathcal{L}(x, q(c(x)))$.

The following Theorem, which may be proved similarly to the Theorems 4 and 5 (Dhillon et al., 2003), formulates the most important descending and convergence properties within the *Clustering Minimisation* (CM) framework:

Theorem 1 *The CM-algorithm includes 2 steps:*

***Clustering Step**: recompute $c(x)$ according to (1) for a fixed prototypes from the given codebook $\mathcal{Q}$, which will be updated as a cluster centers from the next step,*

***Minimisation Step**: recompute cluster centers for a fixed mapping $c(x)$ or minimize the objective function (3) over $\mathcal{Q}$, and*

1) monotonically decreases the value of the objective function (3);

2) converges to a local minimum in a finite number of steps if Minimisation Step has exact solution.

We define an optimal actual codebook $\overline{\mathcal{Q}}$ by the following condition:

$$\Re^{(k)}(\overline{\mathcal{Q}}) := \underset{\mathcal{Q} \in \mathcal{X}^k}{inf} \, \Re^{(k)}(\mathcal{Q}). \tag{4}$$

The following relations are valid

$$\Re^{(k)}_{\text{emp}}[\mathcal{Q}_n] \leq \Re^{(k)}_{\text{emp}}[\overline{\mathcal{Q}}]; \;\; \Re^{(k)}_{\text{emp}}[\overline{\mathcal{Q}}] \Rightarrow \Re^{(k)}[\overline{\mathcal{Q}}] \;\; a.s., \tag{5}$$

where $\mathcal{Q}_n$ is an optimal empirical codebook:

$$\Re^{(k)}_{\text{emp}}(\mathcal{Q}_n) := \underset{\mathcal{Q}\in\mathcal{X}^k}{inf} \{\Re^{(k)}_{\text{emp}}(\mathcal{Q})\}. \tag{6}$$

The main target is to demonstrate asymptotic (*almost sure*) convergence

$$\Re^{(k)}_{\text{emp}}(\mathcal{Q}_n) \Rightarrow \Re^{(k)}[\overline{\mathcal{Q}}] \;\; a.s. \;\; (n \to \infty) \,. \tag{7}$$

In order to prove (7) we define in Section 3 general model which has direct relation to the model in probabilistic space $\mathcal{P}^m$ with with KL divergence (Dhillon et al., 2003).

2.1 Plan of the Proof

The general strategy is to split consideration into outer deviations, and local deviations (Telgarsky and Dasgupta, 2013). Note that the significance of outer deviations is declining as we extend local deviation. The local deviations maybe be controlled by the technique as described below.

The proof of the main result which is formulated in the Theorem 18 includes two steps:

(1) by Lemma 10 we prove existence of n_0 such that $\mathcal{Q}_n \subset \Gamma$ for all $n \geq n_0$, where subset $\Gamma \subset \mathcal{X}$ (local deviation) satisfies condition: $\mathcal{L}(x, q) < \infty$ for all $x \in \mathcal{X}, q \in \Gamma$; and

(2) by Lemma 11 we prove (under some additional constraints of general nature)

$$\underset{\mathcal{Q}\in\Gamma^k}{sup} |\Re^{(k)}_{\text{emp}}[\mathcal{Q}] - \Re^{(k)}[\mathcal{Q}]| \Rightarrow 0 \;\; a.s. \tag{8}$$

3. General Theory and Definitions

In this section we employ some ideas and methods proposed by Pollard (1981) which cover the case of $\mathbb{R}^m$ with loss function $\mathcal{L}(x, q) := \varphi(\|x - q\|)$, where φ is a strictly increasing function.

Let us assume that the following structural representation with $\mathbb{P}$-integrable vector-functions ξ and η is valid

$$\mathcal{L}(x, q) := \sum_{i=0}^{m} \xi_i(x) \cdot \eta_i(q) = \langle \xi(x), \eta(q) \rangle \geq 0 \;\; \forall x, q \in \mathcal{X}. \tag{9}$$

Remark 2 *Above definition (9) was motivated by the structure of KL-divergence, see (29a) and (29b).*

Let us define subsets of $\mathcal{X}$ as extensions of the empirical clusters:

$$\mathcal{X}_c(\mathcal{Q}) := \{x \in \mathcal{X} : \;\; c = \text{argmin}_i \, \mathcal{L}(x, q(i))\} \,,$$

$$\mathcal{X} = \cup_{c=1}^{k} \mathcal{X}_c(\mathcal{Q}), \mathcal{X}_i(\mathcal{Q}) \cap \mathcal{X}_c(\mathcal{Q}) = \emptyset, i \neq c.$$

Then, we re-write (2) as follows

$$\Re^{(k)}[\mathcal{Q}] := \sum_c \langle \xi(\mathcal{X}_c), \eta(q(c)) \rangle, \tag{10}$$

where $\xi(A) := \int_A \xi(x)\mathbb{P}(dx), A \in \mathcal{A}$.

Definition 3 *We define a ball with radius r and a corresponding remainder in $\mathcal{X}$*

$$B(r) = \{q \in \mathcal{X} : \mathcal{L}(x, q) \leq r, \quad \forall x \in \mathcal{X}\}, \tag{11a}$$

$$T(r) = \mathcal{X} \setminus B(r), \quad r \geq \mathbf{r}_0, \tag{11b}$$

$$\mathbf{r}_0 = \inf\{r \geq 0 : B(r) \neq \emptyset\}. \tag{11c}$$

Remark 4 *By the following Lemma 10 we prove that all components of the codebook will be within ball $B(Z), 0 < Z < \infty$, if sample size is large enough. Further, we shall assume that η-transformation of the ball $B(Z)$ represents a compact set (26), and, consequently, we shall be able to prove strong consistency (8) by Lemma 11.*

The following properties are valid

$$\langle \xi(A_1) - \xi(A_2), \eta(q) \rangle \geq 0 \tag{12}$$

for all $q \in \mathcal{X}$ and any $A_1, A_2 \in \mathcal{A} : A_2 \subset A_1$;

$$\langle \xi(\mathcal{X}), \eta(q) \rangle \leq r \quad \forall q \in B(r). \tag{13}$$

Suppose, that

$$\mathbb{P}(T(U)) \underset{U \to \infty}{\longrightarrow} 0. \tag{14}$$

Remark 5 *Condition (14) is the only one requirement which is necessary to prove the main result of this paper: Theorem 18, see, also, Remark 19.*

Definition 6 *The following distances will be used below:*

$$\rho(A_1, A_2) := \underset{a_1 \in A_1}{inf} \underset{a_2 \in A_2}{inf} \mathcal{L}(a_1, a_2), A_1, A_2 \in \mathcal{A}; \tag{15a}$$

$$\mu(A_1, A_2) := \underset{a_1 \in A_1}{inf} \underset{a_2 \in A_2}{sup} \mathcal{L}(a_1, a_2), A_1, A_2 \in \mathcal{A}. \tag{15b}$$

Remark 7 *Above distances ρ and μ have very simple interpretation: ρ - absolutely minimal distance between elements of the subsets A_1 and A_2 (it is symmetrical); μ - uniformly minimal distance between elements of the subset A_1 (approximator) and elements of another subset A_2 (it is not symmetrical).*

Suppose, that

$$\rho(B(r), T(U)) \underset{U \to \infty}{\longrightarrow} \infty \tag{16}$$

for any fixed $\mathbf{r}_0 \leq r < \infty$.

Remark 8 *Above condition (16) is always valid for KL-divergence, see Corollary 17.*

Remark 9 *We assume that*

$$T(U) \neq \emptyset \tag{17}$$

for any fixed $U : \mathbf{r}_0 \leq U < \infty$*, alternatively, the following below Lemma 10 becomes trivial.*

Lemma 10 *Suppose, that the structure of the loss function* $\mathcal{L}$ *is defined in (9) under condition (16). Probability distribution* $\mathbb{P}$ *satisfies condition (14) and the number of clusters* $k \geq 1$ *is fixed. Then, we can select large enough radius* $Z : 0 < Z < \infty$ *and* $n_0 \geq 1$ *such that all components of the optimal empirical codebook* $\mathcal{Q}_n$ *defined in (6) will be within the ball* $B(Z)$*:* $\mathcal{Q}_n \subset B(Z)$ *if sample size is large enough:* $\forall n \geq n_0$*.*

Proof. Existence of the element $\mathbf{a} \in \mathcal{X}$ such that

$$D_{\mathbf{a}} = \Re^{(1)}(\{\mathbf{a}\}) = \langle \xi(\mathcal{X}), \eta(\mathbf{a}) \rangle < \infty \tag{18}$$

follows from (13) and (14).

Suppose that

$$\mathbb{P}(B(r)) = P_0 > 0, \;\; r \geq \mathbf{r}_0. \tag{19}$$

We construct $B(V)$ in accordance with conditions (16) and (17):

$$V = \inf\{v > r : \rho(B(r), T(v)) \geq \frac{D_{\mathbf{a}} + \epsilon}{P_0}\}, \;\; \epsilon > 0. \tag{20}$$

Suppose, there are no empirical prototypes within $B(V)$. Then, in accordance with definition (19)

$$\Re^{(k)}_{\text{emp}}[\mathcal{Q}_n] \geq D_{\mathbf{a}} + \epsilon > D_{\mathbf{a}} \;\; \forall n > 0.$$

Above contradicts to (18) and (5). Therefore, at least one prototype from $\mathcal{Q}_n$ must be within $B(V)$ if n is large enough (this fact is valid for $\overline{\mathcal{Q}}$ as well). Without loss of generality we assume that

$$q(1) \in B(V). \tag{21}$$

The proof of the Lemma has been completed in the case if $k = 1$.

Assumption. *Following the method of mathematical induction, suppose, that* $k \geq 2$ *and*

$$\Re^{(k-1)}(\overline{\mathcal{Q}}) - \Re^{(k)}(\overline{\mathcal{Q}}) \geq \varepsilon > 0. \tag{22}$$

Then, we define a ball $B(U)$ by the following conditions

$$U = \inf \{u > V : \sup_{q \in B(V)} \langle \xi(T(u)), \eta(q) \rangle < \varepsilon\}. \tag{23}$$

Existence of the $U : V < U < \infty$ in (23) follows from (13) and (14).

By definition of the distance μ and the ball $B(V)$

$$0 < \mathcal{D}(U, V) = \mu(T(U), B(V)) \leq V < \infty. \tag{24}$$

Now, we define reminder $T(Z) \neq \emptyset$ in accordance with condition (16):

$$Z = \inf \{z > U : \rho(B(U), T(z)) \geq \mathcal{D}(U, V)\}. \tag{25}$$

Suppose, that there is at least one prototype within $T(Z)$, for example, $q(2) \in T(Z)$. On the other hand, we know about (21). Let us consider what will happen if we remove $q(2)$ from the optimal empirical codebook $\mathcal{Q}_n$ (the case of optimal actual risk $\overline{\mathcal{Q}}$ may be considered similarly), and replace it by $q(1)$:

(1) as a consequence of (24) and (25) all empirical data within $B(U)$ are closer to $q(1)$ anyway, means the data from $B(U)$ will not increase empirical (or actual) risk (3);

(2) by definition, $\mathcal{X} = B(U) \cup T(U), B(U) \cap T(U) = \emptyset$ and in accordance with the condition (23) an empirical risk increases because of the data within $T(U)$ must be strictly less compared with ε for all large enough $n \geq n_0$ (actual risk increase will be strictly less compared with ε for all $n \geq 1$).

Above *contradicts* to the condition (22) and (5). Therefore, all prototypes from $\overline{\mathcal{Q}}$ must be within $\Gamma = B(Z)$ for all $n \geq 1$, and $\mathcal{Q}_n \subset \Gamma$ if n is large enough. ■

3.1 Uniform Strong Law of Large Numbers (SLLN)

Let $\mathcal{F}$ denote the family of $\mathbb{P}$-integrable functions on $\mathcal{X}$.

A sufficient condition for *uniform SLLN* (8) is: for each $\delta > 0$ there exists a *finite* class $\mathcal{F}_\delta \in \mathcal{F}$ such that for each $\mathcal{L} \in \mathcal{F}$ there are functions $\underline{\mathcal{L}}$ and $\overline{\mathcal{L}} \in \mathcal{F}_\delta$ with the following 2 properties:

$$\underline{\mathcal{L}}(x) \leq \mathcal{L}(x) \leq \overline{\mathcal{L}}(x) \text{ for all } x \in \mathcal{X}; \quad \int_{\mathcal{X}} \left(\overline{\mathcal{L}}(x) - \underline{\mathcal{L}}(x)\right) \mathbb{P}(dx) \leq \delta.$$

We assume here existence of the function φ such that

$$\|\eta(q)\| \leq \varphi(Z) < \infty \tag{26}$$

for all $q \in B(Z)$, where $\mathbf{r}_0 \leq Z < \infty$.

Lemma 11 *Suppose that the number of clusters k is fixed, and the loss function $\mathcal{L}$ is defined by (9) under condition (26) and*

$$\|\xi(x)\| \leq \mathbf{R} < \infty \ \ \forall x \in \mathcal{X}. \tag{27}$$

Then, the asymptotic relation (8) is valid for any $\Gamma = B(Z), \mathbf{r}_0 \leq Z \leq \infty$.

Proof. Let us consider the definition of Hausdorff metric $\mathcal{H}$ in $\mathbb{R}^{m+1}$:

$$\mathcal{H}(A_1, A_2) = \underset{a_1 \in A_1}{sup} \underset{a_2 \in A_2}{inf} \|a_1 - a_2\|,$$

and denote by $\mathcal{G}$ a subset in $\mathbb{R}^{m+1}$ which was obtained from Γ as a result of η-transformation. According to the condition (26), $\mathcal{G}$ represents a compact set. It means, existence of a finite subset $\mathcal{G}_\delta$ for any $\delta > 0$ such that $\mathcal{H}(\mathcal{G}, \mathcal{G}_\delta) \leq \frac{\delta}{2\mathbf{R}}$, where $\mathbf{R}$ is defined in (27). We denote by $\Gamma_\delta \subset \Gamma$ subset which corresponds to $\mathcal{G}_\delta \subset \mathcal{G}$ according to the η-transformation. Respectively, we can define transformation (according to the principle of the nearest point) f_δ from Γ to Γ_δ, and $\mathcal{Q}_\delta = f_\delta(\mathcal{Q})$, where closeness may be tested independently for any particular component of $\mathcal{Q}$, that means absolute closeness.

In accordance with the Cauchy-Schwartz inequality, the following relations take place

$$\underline{\mathcal{L}} = \mathcal{L}(x, \mathcal{Q}_\delta) - \frac{\delta}{2} \leq \mathcal{L}(x, \mathcal{Q}) \leq \mathcal{L}(x, \mathcal{Q}_\delta) + \frac{\delta}{2} = \overline{\mathcal{L}} \ \forall x \in \mathcal{X}.$$

Finally, $\int_{\mathcal{X}} \left(\overline{\mathcal{L}}(x, \mathcal{Q}_\delta) - \underline{\mathcal{L}}(x, \mathcal{Q}_\delta)\right) \mathbb{P}(dx) \leq \delta$, where $\mathcal{Q}_\delta \in \Gamma_\delta^k$ is the absolutely closest codebook for the arbitrary $\mathcal{Q} \in \Gamma^k$. ■

4. A Probabilistic Framework

Following Dhillon et al. (2003), we assume that the probabilities $p_{\ell t} = P(\ell | x_t), \sum_{\ell=1}^m p_{\ell t} = 1, t = 1, \ldots, n$, represent relations between observations x_t and attributes or classes $\ell = 1, \ldots, m, m \geq 2$.

Accordingly, we define probabilistic space $\mathcal{P}^m$ of all m-dimensional probability vectors with *Kullback-Leibler* (KL) divergence:

$$KL(v, u) := \sum_\ell v_\ell \cdot \log \frac{v_\ell}{u_\ell} = \langle v, \log \frac{v}{u} \rangle, \ v, u \in \mathcal{P}^m.$$

Remark 12 *As it was demonstrated by Dhillon et al. (2003), cluster centers q_c in the space $\mathcal{P}^m$ with KL-divergence must be computed using K-means:*

$$q_c = \frac{1}{n_c} \sum_{x_t \in \mathbf{X}_c} p_t, \tag{28}$$

where $c(x_t) = c$ if $x_t \in \mathbf{X}_c$ and $n_c = \#\mathbf{X}_c$ is the number of observations in the cluster $\mathbf{X}_c, c = 1, \ldots, k$, $p_t = \{p_{1t}, \ldots, p_{mt}\}, q_c = \{q_{1t}, \ldots, q_{mt}\}$.

In difference to the model of Pollard (1981) in $\mathbb{R}^m$, the structure (9) covers an important case of $\mathcal{P}^m$ with KL-divergence:

$$\xi_0(v) = \sum_{\ell=1}^m v_\ell \log v_\ell; \quad \xi_\ell(v) = v_\ell; \tag{29a}$$

$$\eta_0(u) = 1; \quad \eta_\ell(u) = -\log u_\ell, \ell = 1, \ldots, m. \tag{29b}$$

Definition 13 *We call element $v \in \mathcal{P}^m$ as 1) uniform center if $v_\ell = \frac{1}{m}, \ell = 1, \dots, m$; as 2) absolute margin if* $\min_\ell v_\ell = 0$.

Proposition 14 *The ball $B(Z) \subset \mathcal{P}^m$ contains only one element named as uniform center in the case if $Z = \mathbf{r}_0 = \log(m)$, and $B(Z) = \emptyset$ if $Z < \mathbf{r}_0$.*

Proof. Suppose, that u is a uniform center. Then, $KL(v, u) = \sum_{i=1}^m v_i \log v_i + \log m \leq \log m$ for all $v \in \mathcal{P}^m$. In any other case, one of the components of u must be less than $\frac{1}{m}$. Respectively, we can select the corresponding component of probability vector v as 1. Therefore, $KL(v, u) > \log(m)$ and $\mathbf{r}_0 = \log(m)$. ■

Lemma 15 *The KL divergence in probabilistic space $\mathcal{P}^m$ always satisfies condition (27), where vector-function ξ is expressed by (29a)with the following upper bounds:*

$$|\xi_0(v)| \leq \log(m); \quad |\xi_\ell(v)| \leq 1, \ell = 1, \dots, m, \qquad \forall v \in \mathcal{P}^m.$$

Lemma 16 *The following relations are valid in $\mathcal{P}^m$*

(1) $min_\ell\{u_\ell\} < e^{-r}$ for all $u \in T(r)$ $\forall r \geq \mathbf{r}_0$;

(2) $u_\ell \geq e^{-r}$ for all $\ell = 1, \dots, m$, and any $u \in B(r)$ $\forall r \geq \mathbf{r}_0$.

Proof. As far as $\mathcal{P}^m = B(r) \cup T(r), B(r) \cap T(r) = \emptyset$, the first statement may be regarded as a consequence of the second statement. Suppose, that $u \in B(r)$ and $u_1 = e^{-r-\varepsilon}, \varepsilon > 0$. Then, we can select $v_1 = 1$, and $KL(v, u) = r + \varepsilon > r$ - *contradiction.* ■

Corollary 17 *The KL divergence in $\mathcal{P}^m$ always satisfies conditions (16), and*

$$-\log(m) + Z \cdot e^{-r} < \rho(B(r), T(Z)) \leq e^{-r} \cdot (Z - r) + \left(1 - e^{-r}\right) \log \frac{1 - e^{-r}}{1 - e^{-Z}}$$

for all $\mathbf{r}_0 \leq r < Z$, *where the distance ρ is defined in (15a).*

Proof. Suppose, that $v \in B(r)$ and $u \in T(Z)$. Then, $-\sum_{i=1}^m v_i \log(u_i) > Z \cdot e^{-r}$ for all $r : \mathbf{r}_0 \leq r < Z$. On the other hand, the entropy $H(v) = -\sum_{i=1}^m v_i \log(v_i)$ may not be smaller compared to $\log(m)$. The low bound is *proved.* In order to prove the upper bound we suppose without loss of generality that $v_1 = e^{-r}, u_1 = e^{-Z}$, and all the other components are proportional. ■

Theorem 18 *Suppose that probability measure $\mathbb{P}$ satisfies condition (14) in probabilistic space $\mathcal{P}^m$ with KL divergence and the number of clusters k is fixed. Then, the minimal empirical error (6) converges to the minimal actual error (4) with probability 1 or a.s.*

Proof. Follows directly from the Lemmas 10, 11, 15 and 16.

Remark 19 *Condition (14) is not valid if and only if the probability of the subset of all absolute margins is strictly positive. Note that in order to avoid any problems with consistency we can generalise definition of KL-divergence using special smoothing parameter $0 \leq \theta \leq 1$:*

$$KL_\theta(v, u) = KL(v_\theta, u_\theta),$$

where $v_\theta = \theta v + (1 - \theta) v_0$, and $u_\theta = \theta u + (1 - \theta) v_0$, v_0 is uniform center.

5. Concluding Remarks

Consistency is a key property of all statistical procedures analyzing randomly sampled data. Surprisingly, despite decades of work, little is known about consistency of most clustering algorithms (von Luxburg et al., 2008). In this paper we developed a general framework to investigate and to prove consistency of the popular family of prototype based clustering algorithms. As an illustration, we considered probabilistic space with Kullback-Leibler divergence.

Acknowledgment

The author would like to thank two reviewers for very helpful comments and advice.

References

M. Ackerman, J. Blomer, and C. Sohler. Clustering with metric and non-metric distance measures. In *SODA*, 2008.

A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

S. Ben-David, U. Von Luxburg, and D. Pal. A sober look at clustering stability. In *COLT*, 2006.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

K. Chaudhuri and A. McGregor. Finding metric structure in information theoretic clustering. In *COLT*, 2008.

J. Cuesta-Albertos, A. Gordaliza, and C. Matran. Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, 25(2):553–576, 1997.

I. Dhillon, S. Mallela, and R. Kumar. Divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, 2003.

T. Glasmachers. Universal consistency of multi-class support vector classification. In *NIPS*, 2010.

A. Hinneburg and D. Keim. A general approach to clustering in large databases with noise. *Knowledge and Information Systems*, 4:387–415, 2003.

S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Scholkopf. Consistency of causal inference under the additive noise model. In *ICML*, 2014.

P. Long and R. Servedio. Consistency versus realizable H-consistency for multiclass classification. In *ICML*, pages 801–809, 2013.

D. McAllester and J. Keshet. Generalization bounds and consistency for latent structural probit and rump loss. In *NIPS*, 2011.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In *NIPS*, 2001.

R. Nock, P. Luosto, and J. Kivinen. Mixed Bregman clustering with approximation guarantees. *Machine Learning and Knowledge Discovery in Databases*, pages 154–169, 2008.

D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 10(1): 135–140, 1981.

A. Rakhlin and A. Caponnetto. Stability of k-means clustering. In *NIPS*, 2006.

O. Shamir and N. Tishby. Model selection and stability of k-means clustering. In *COLT*, 2008.

M. Telgarsky and S. Dasgupta. Moment-based uniform deviation bounds for k-means and friends. In *NIPS*, 2013.

V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.