Consensual Aggregation of Clusters Based on Bregman Divergences to Improve Predictive Models

Sothea Has

LPSM, Université Paris-Diderot
75013 Paris, France
sothea.has@lpsm.paris

Résumé

Dans cet exposé, nous introduisons une nouvelle approche pour construire des modèles prédictifs dans les problèmes d'apprentissage supervisé en prêtant attention à la structure de regroupement des données d'entrée. Cet exposé est basé sur un travail conjoint de Has et al. (2018). Nous nous intéressons aux situations où les données d'entrée sont composées de plusieurs grappes et qu'il existe différents modèles sous-jacents sur ces grappes. Ainsi, au lieu de construire un seul modèle prédictif sur l'ensemble des données, nous proposons dans un premier temps d'utiliser l'algorithme K-means avec différentes options de divergences de Bregman qui sont les membres d'une large classe de mesures de dissimilarité, pour approcher la structure des données d'entrée. Pour chaque divergence, nous construisons un prédicteur local sur chaque cluster observé, ce qui conduira à un modèle global qui est la collection de ces prédicteurs locaux. Dans un deuxième temps, nous proposons de combiner intelligemment tous ces modèles globaux de manière à préserver la qualité de la combinaison, voire à l'améliorer, par rapport au meilleur modèle de la combinaison. Les résultats numériques réalisés sur plusieurs types de données simulées et une donnée réelle de Air Compressor montrent qu'il est très intéressant de prendre en compte la structure de clustering des données d'entrée, ainsi que d'utiliser des méthodes d'estimation combinées pour améliorer les performances de modèles prédictifs.

Mots-clés. Apprentissage supervisé, divergences de Bregman, cluster, modèles prédictifs, méthode d'estimation combinée.

Abstract

In this talk, we introduce a new approach to construct predictive models in supervised learning problems by paying attention to the clustering structure of the input data. This talk is based on a joint work of Has et al. (2018). We are interested in the situations where the input data consists of more than one cluster, and there exist different underlying models on these clusters. Thus, instead of constructing a single predictive model on the whole data, we propose in the first step to use

K-means clustering algorithm with different options of Bregman divergences which are the members of a broad class of dissimilarity measures, to approximate the clustering structure of the input data. For each divergence, we construct a local predictor on each observed cluster, and this will lead to a global model which is the collection of these local predictors. In the second step, we propose to combine all of these global models in a smart way in a sense that the quality of the combination is asymptotically preserved, or even improved compared to the best model of the combination. The numerical results carried out on several kinds of simulated data and a real data set of Air Compressor, show that it is very interesting to take into account the clustering structure of the input data, and also, to use combining estimation methods to improve the performances of predictive models.

Keywords. Supervised learning, Bregman divergences, cluster, predictive models, combined estimation method.

1 Introduction

We assume that the number of clusters K of the input data is available. In the first step of the unsupervised learning part, our goal is trying to identify the unknown clustering structure of the input data using a broad class of dissimilarity measures known as Bregman divergences (See Bregman (1967)). Each member of this class is defined associated to a strictly convex and continuously differentiable function $\phi : \mathcal{C} \to \mathbb{R}$ where $\mathcal{C} \subset \mathbb{R}^d$ is a measurable convex subset of $\mathbb{R}^d$ and its relative interior is denoted by $int(\mathcal{C})$. A Bregman divergence $d_\phi : \mathcal{C} \times int(\mathcal{C}) \to \mathbb{R}$, indexed by ϕ is defined for any pair $(x, y) \in \mathcal{C} \times int(\mathcal{C})$ by

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle \tag{1}$$

where $\langle ., . \rangle$ denotes the standard inner product defined on $\mathbb{R}^d$, and $\nabla\phi(y)$ denotes the gradient of ϕ computed at a point $y \in int(\mathcal{C})$. A Bregman divergence is not necessarily a metric. However, it carries many interesting properties, and the most interesting one is *mean as minimizer* property given in the proposition below. See, for example, Banerjee et al. (2005) and Fischer (2010).

Proposition *(Mean as Minimizer Property) Suppose U is a random variable over an open set $\mathcal{O} \subset \mathbb{R}^d$, then we have,*

$$\mathbb{E}[U] = \underset{x \in \mathcal{O}}{\operatorname{argmin}}\, \mathbb{E}[d_\phi(U, x)]$$

1.1 Relationship between Bregman divergences and Exponential family

A strong motivation of using Bregman divergences in K-means algorithm, to approximate the clustering structure of the input data, is its relationship with a broad class of probability distributions known as Exponential family. This relationship is given in the following theorem.

Theorem *(Banerjee et al. (2005)) Each member of a regular exponential family corresponds to a unique regular Bregman divergence. Mathematically, if the distribution of a random variable X is a member of an exponential family $\mathcal{E}_\psi$ and if ϕ is a convex conjugate of ψ defined by*

$$\phi(x) = \max_y \{\langle x, y\rangle - \psi(y)\}$$

then there exists a Bregman divergence d_ϕ such that the following representation holds,

$$f_\theta(x) = \exp(\langle \theta, T(x)\rangle - \psi(\theta)) = \exp(-d_\phi(T(x), \mathbb{E}[T(X)]) + \phi(T(x))) \tag{2}$$

K-mean clustering with a Bregman divergence d_ϕ is provided in the following algorithm.

Algorithm

1. *Randomly initialize the centroids*
   $C = \{c_1, c_2, ..., c_K\}$ *among the data points.*

2. ***for*** $i = 1, 2, ..., n$, *assign x_i to kth cluster if*

$$d_\phi(x_i, c_k) = \min_{1 \le j \le K} d_\phi(x_i, c_j)$$

3. *Denote C_k the set of points contained in the kth cluster then,*
   for $k = 1, 2, ..., K$, *recompute the new centroid by,*

$$c_k^{new} = \frac{1}{|C_k|} \sum_{x \in C_k} x$$

Repeat step 2 and 3 until convergence.

1.2 Candidate estimator construction

We suppose that each cluster contains enough data points so that a separate local predictor could be constructed. The choices of these local predictors should not be too complicated as the input-output relation within each cluster should not be too complicated either. Take, for example, linear regression for regression and logistic regression for classification. Thus, a Bregman divergence would lead to a candidate estimator which is the collection of all the constructed local predictors.

2 Consensual aggregation and applications

2.1 Consensual aggregation methods

Choosing the "best" one among all the candidate estimators is not a simple problem as neither the distribution nor the clustering structure of the input data is available. Therefore, we propose in the second step to combine all of these candidate estimators in a smart way so that the quality of the best model is asymptotically preserved or improved. In this part, we consider several consensual aggregation methods in both classification and regression problems which are based on the consensual predictions of the training sample. This idea was first introduced by Mojirsheibani (1999) in a classification setting. Later, this idea was extended by Biau et al. (2016) into a regression framework which is known as COBRA method, and then it was extended into both frameworks of classification and regression by Fischer and Mathilde Mougeot (2019) in a study known as MixCOBRA. We also consider the kernel-based version of consensual aggregation in regression with several options of kernel functions. More precisely, each consensual regression in this study takes the following form,

$$Comb^R(x) = \frac{1}{n}\sum_{i=1}^{n} W_{n,i}(x) y_i \tag{3}$$

where the weight $W_{n,i}(x)$ is defined differently as follow,

- COBRA method (Biau et al. (2016)): for a given threshold $\varepsilon > 0$,

$$W_{n,i}(x) = \frac{\prod_{\ell=1}^{M} \mathbb{1}_{\{|m^\ell(x_i) - m^\ell(x)| < \varepsilon\}}}{\sum_{j=1}^{n} \prod_{\ell=1}^{M} \mathbb{1}_{\{|m^\ell(x_j) - m^\ell(x)| < \varepsilon\}}}$$

- Kernel based COBRA: for a given smoothing parameter $h > 0$,

$$W_{n,i}(x) = \frac{K_h(\mathbf{m}(x_i) - \mathbf{m}(x))}{\sum_{j=1}^{n} K_h(\mathbf{m}(x_j) - \mathbf{m}(x))}$$

  with a kernel function K such that $K_h(x) = K(x/h)$ and a vector of predictions at a point $x \in \mathbb{R}^d, \mathbf{m}(x) = (m^1(x), m^2(x), ..., m^M(x)) \in \mathbb{R}^M$.

- MixCOBRA (Fischer and Mathilde Mougeot (2019)): for a given couple of smoothing parameter $\alpha, \beta > 0$,

$$W_{n,i}(x) = \frac{K(\frac{x_i - x}{\alpha}, \frac{\mathbf{m}(x_i) - \mathbf{m}(x)}{\beta})}{\sum_{j=1}^{n} K(\frac{x_j - x}{\alpha}, \frac{\mathbf{m}(x_j) - \mathbf{m}(x)}{\beta})}$$

  where a $\mathbb{R}^{d+M}$-vector of individual in the equation is composed of the original input $x \in \mathbb{R}^d$ and the vector of predictions $\mathbf{m}(x) \in \mathbb{R}^M$.

2.2 Applications

At the end of this study, we illustrate the performances and the benefits of the constructed estimators with several experiments carried out on several kinds of simulated data. The numerical results show that the quality of the combination is quite satisfactory and sometimes even better compared to the best candidate estimator in the aggregation. The numerical results of the constructed regression models performed on a real data set of Air Compressor given in Cadet et al. (2005) also show that even without the exact information of the number of clusters of the input data, the constructed models still perform well regardless of the information of K.

References

Banerjee, A., Srujana Merugu, Inderjit S. Dhillon, Joydeep Ghosh, 2005. Clustering with Bregman divergences. Journal Machine Learning Research 6, 1705–1749.

Biau, G., Aurélie Fischer, Benjamin Guedj, James D. Malley, 2016. COBRA: a Combined Regression Strategy. Journal of Multivariate Analysis 146, 18–28.

Bregman, L.M., 1967. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematical and Mathematical Physics 7, 200–217.

Cadet, O., Harper, C., Mougeot, M., 2005. Monitoring energy performance of compressors with an innovative auto-adaptive approach., in: Instrumentation System and Automation -ISA- Chicago.

Fischer, A., 2010. Quantization and clustering with Bregman divergences. Journal of Multivariate Analysis 101, 2207–2221.

Fischer, A., Mathilde Mougeot, 2019. Aggregation using input-output trade-off. Journal of Statistical Planning and Inference 200, 1–19.

Has, S., Fischer, A., Mougeot, M., 2018. Consensual aggregation of clusters based on Bregman divergences to improve predictive models. Technical Report. Université Paris-Diderot.

Mojirsheibani, M., 1999. Combined classifiers via disretization. Journal of the American Statistical Association 94, 600–609.