Notes on Prototype Regression *

G. Biau, Q. Du and Y. Liu

Abstract

Prototype regression is a family of supervised learning algorithm that aim at making a summary of the original dataset, which is assumed to be very large, by a small number of representative samples. In particular, we are interested by a family of prototype methods that are characterized by a selection of quantizer and the associated weights. The idea of this note is to keep a track on the workflow, such as notation changes, minutes, questions, etc.

Contents

*Working document, with no intention to be published

1 General background of prototype regression

1.1 Notation

Due to the remark of Yating 10-09-2020, we decide to apply some minor modifications to the original document so that the notation is more standard to the quantization community. The following notation is mainly from the notes (Yating→Qiming, 10-09-2020 and 14-09-2020).

- (General) The canonical probability space is denoted by $(\Omega, \mathcal{F}, \mathbb{P})$. The canonical covariate X is a random variable on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and the response variable Y is a random variable on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The original dataset is denoted by $\mathcal{D}_n := \{(X_i, Y_i) : 1 \leq i \leq n\}$. The samples (X_i, Y_i) are assumed to be i.i.d. copies of the canonical random variable (X, Y). The Euclidean norm will be denoted by "$|\cdot|$". If not mentioned otherwise, all the measures and random variables considered in this article are Borel measurable w.r.t. the Euclidean norm of $\mathbb{R}^d$.

- (Quantization) We call the K-tuple $\mathbf{c} = (c_1, c_2, \ldots, c_K) \in (\mathbb{R}^d)^K$ a quantizer, where K is the number of centers. Given a quantizer $\mathbf{c}$, the Voronoï partitions are denoted by $(V_k(\mathbf{c}); 1 \leq k \leq K)$. More precisely, $V_k(\mathbf{c})$ is defined by

$$V_k(\mathbf{c}) := \left\{ x \in \mathbb{R}^d : |x - c_k| = \min_{1 \leq k \leq K} |x - c_k| \right\}.$$

  Based on a Voronoï partition $(V_k(\mathbf{c}); 1 \leq k \leq K)$, one can define a projection function $\mathrm{Proj}_{\mathbf{c}} : \mathbb{R}^d \mapsto \{c_1, c_2, \ldots, c_K\}$ by

$$x \mapsto \sum_{k=1}^{K} c_k \mathbf{1}_{V_k(\mathbf{c})}(x).$$

  Thus, for a random variable X on $\mathbb{R}^d$, we denote $\hat{X}^{\mathbf{c}} := \mathrm{Proj}_{\mathbf{c}}(X)$. Intuitively speaking, $\hat{X}^{\mathbf{c}}$ can be regarded as an estimation of X w.r.t. the quantizer $\mathbf{c}$. Similarly, for a probability measure μ on $\mathbb{R}^d$, we define the associated projection $\hat{\mu}^{\mathbf{c}}$ by

$$\hat{\mu}^{\mathbf{c}} := \sum_{k=1}^{K} \mu(V_k(\mathbf{c}))\delta_{c_i}.$$

1.2 Prototype regression based on a quantizer

We provide in the following the construction of our prototype regressor at a high level, by fixing a quantizer $\mathbf{c}$ with K centers:

- The weight vector $\boldsymbol{w} := (w_1, w_2, \ldots, w_k)$ is a K-tuple defined by

$$w_k = \sum_{i=1}^{n} \frac{\mathbf{1}_{V_k(\mathbf{c})}(X_i)}{\sum_{j=1}^{n} \mathbf{1}_{V_k(\mathbf{c})}(X_j)} Y_i. \tag{1}$$

  More concretely, the weight w_k of the k-th cluster is but the average of the Y_i that fall into the Voronoï cell $V_k(\mathbf{c})$.

- The prototype estimator $\hat{m}_n^{\mathbf{c}}(\cdot)$ is defined by

$$\forall x \in \mathbb{R}^d, \quad \hat{m}_n^{\mathbf{c}}(x) := \sum_{k=1}^{K} w_k \mathbf{1}_{V_k(\mathbf{c})}(x).$$

1.3 On the choice of quantizer

We are mainly interested by two types of quantizers:

(i) (Optimal Quantization) Given the number of centers K, the quantizer $\mathring{\mathbf{c}} = (\mathring{c}_1, \mathring{c}_2, \ldots, \mathring{c}_K)$ is chosen to minimize de squared Euclidean error

$$\mathring{\mathbf{c}} = \underset{\mathbf{c}\in(\mathbb{R}^d)^K}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[K]} |X_i - c_k|^2 .$$

(Theoretical version? existence? We only need X is square integrable?)

$$\mathring{\mathbf{c}}_\star = \underset{\mathbf{c}\in(\mathbb{R}^d)^K}{\operatorname{argmin}} \mathbb{E}\left[\min_{k\in[K]} |X - c_k|^2\right] .$$

The associated prototype regressor is therefore denoted by $m_n^{\mathring{\mathbf{c}}}(\cdot)$.

(ii) (Supervised Quantization) Given the number of centers K, the quantizer $\dot{\mathbf{c}}$ is chosen to minimize the empirical L^2-loss, namely,

$$\dot{\mathbf{c}} = \underset{\mathbf{c}\in(\mathbb{R}^d)^K}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{m}_n^{\mathbf{c}}(X_i)|^2 .$$

We denote $\hat{m}_n(\cdot)$ (or $\hat{m}_n^{\dot{\mathbf{c}}}(\cdot)$) is the prototype regressor associated to the quantizer $\dot{\mathbf{c}}$.

2 Asymptotic behaviors

2.1 Universal consistency

2.2 Rate of convergence

3 Stochastic Gradient Decent for prototype regression

Appendices

Appendix A Intuition and ideas

A.1 Optimal quantization and supervised quantization

Our first and fundamental conjecture is that the prototype regressor with optimal quantization (i.e., $\hat{m}_n^{\mathring{\mathbf{c}}}(\cdot)$) and supervised quantization (i.e., $\hat{m}_n(\cdot)$ or $\hat{m}_n^{\dot{\mathbf{c}}}(\cdot)$) have the "same" asymptotic behavior. For example, on the Lipschitz family, they have the same (minimax) convergence rate (at least for the case $d > 1$). At the same time, we also believe that $K \to \infty$ is the only condition needed to establish universal consistency. (To be verified.)

As a consequence, a simple idea to study the asymptotic behavior of supervised quantization is to exploit the nice theoretical properties of optimal quantization, and to prove that the L^2-error of the prototype regressor associated to latter can serve as a natural upper bound in the former case. The difficulty of studying directly the supervised prototype regressor lies in the fact that the construction of $\hat{m}_n^{\dot{\mathbf{c}}}(\cdot)$ depends on the values of Y_i, which involves some classical technical problems when dealing with the associated L^2-error. To make it clearer, by definition of $\dot{\mathbf{c}}$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \hat{m}_n^{\dot{\mathbf{c}}}(X_i)\right|^2 \le \frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \hat{m}_n^{\mathring{\mathbf{c}}_\star}(X_i)\right|^2 , \quad a.s.$$

whence, considering that Y is assumed to be square integrable,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Y_i-\hat{m}_n^{\mathring{\mathbf{c}}_\star}(X_i)\right|^2\right] \leq& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left|Y_i-\hat{m}_n^{\mathring{\mathbf{c}}_\star}(X_i)\right|^2\right] \\ =&\mathbb{E}\left[\left|Y-\hat{m}_n^{\mathring{\mathbf{c}}_\star}(X)\right|^2\right] \\ =& \underbrace{\mathbb{E}\left[\left|m(X)-\hat{m}_n^{\mathring{\mathbf{c}}_\star}(X)\right|^2\right]}_{\text{Optimal qunatization prototype regression}} +\mathbb{E}\left[\left|Y-m(X)\right|^2\right]\end{aligned}.$$

One step further, if, for the left hand side of the inequality above, we are able to prove that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|Y_i-\hat{m}_n^{\dot{\mathbf{c}}}(X_i)\right|^2\right]-\mathbb{E}\left[\left|Y-m_n^{\dot{\mathbf{c}}}(X)\right|^2\right]\xrightarrow[n\to\infty]{}0, \tag{2}$$

we can thus study the convergence of $\mathbb{E}\left[\left|m_n^{\dot{\mathbf{c}}}(X)-m(X)\right|^2\right]$ by analyzing $\mathbb{E}\left[\left|m_n^{\mathring{\mathbf{c}}_\star}(X)-m(X)\right|^2\right]$. The advantage, as mentioned above, is that $m_n^{\mathring{\mathbf{c}}_\star}(\cdot)\perp\!\!\!\perp\sigma(Y_1,Y_2,\ldots,Y_n)$. In addition, the asymptotic behavior of optimal quantization based prototype regression is already an interesting topic. It can also be implemented with various tractable methods (CITE PAPERS, Lloyd/online k-means, etc.)

A.2 Approximation error of supervised quantization

Now, let us take a closer look at (2). In general, it is clear that

$$\mathbb{E}\left[\left|Y_i-m_n^{\dot{\mathbf{c}}}(X_i)\right|^2\right]\neq\mathbb{E}\left[\left|Y-m_n^{\dot{\mathbf{c}}}(X)\right|^2\right].$$

This is a well-known phenomenon due to the complex dependence between $m_n^{\dot{\mathbf{c}}}(X_i)$ and Y_i. (It is also important to note that there is no such difficulty in the case of optimal quantization, where $m_n^{\mathring{\mathbf{c}}}(X_i)\perp\!\!\!\perp Y_i$). Hence, it is not easy to exploit concentration inequality to prove (2). In the following, this type of error is referred to as *approximation error*. A classical way to deal with it is to consider the method of covering number. More precisely, since