An indirect nonparametric regression method for one-dimensional continuous distributions using warping functions

Zhicheng Chen*

ABSTRACT

Distributions play a very important role in many applications. Inspired by the newly developed warping transformation of distributions, an indirect nonparametric distribution to distribution regression method is proposed in this article for predicting correlated one-dimensional continuous probability density functions.

Keywords: distribution regression, warping transformation, nonparametric regression

1. Introduction

In this article, the correlation between two distribution classes means their density functions will change simultaneously to some extent (see Fig. 1). The correlation between two distribution classes is distinct from the correlation between two random variables. Two correlated distribution classes do not guarantee random variables respectively follow these two distribution classes are also correlated. For instance, suppose probability density functions g_t and f_t are correlated, consider two time-varying random variables $X_t \sim g_t$ and $Y_t \sim f_t$, if the joint distribution of (X_t, Y_t) can be factorized as $p_{X_t Y_t}(x_t, y_t) = g_t(x_t) f_t(y_t)$, then X_t and Y_t are independent with each other. Similarly, two correlated random variables also do not guarantee the distributions they follow are also correlated.

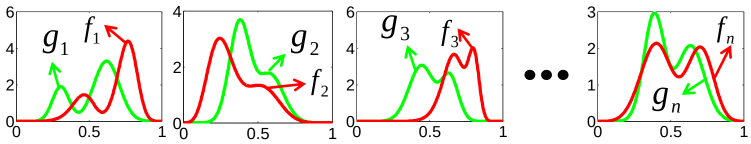

Fig.1 Graphical representation of two correlated distribution classes

Inspired by the seminal work done by Dasgupta *et al.* [1], an indirect nonparametric distribution to distribution regression method is proposed in this article for two correlated one-dimensional continuous distribution classes. Other related work includes the conventional distribution to distribution regression [2] and distribution to real-value or vector-value regression [3-7], etc.

2. Introduction to the warping transformation of distributions

The warping transformation of a distribution is a map that used to transform a distribution to another by deforming the original probability density function with a warping function [1, 8]. The newly reported article by Dasgupta *et al.* [1] has given a very detailed discussion of this transformation.

All distributions in this study are assumed to be continuous with strictly positive support on $[0,1]$, distributions with general finite supports can be easily tackled by the scale transformation introduced in [1]. Given a probability density function $g(x)$ with strictly positive support on $[0,1]$, the warping transformation of $g(x)$ by a warping function $\gamma(x)$ defined on $[0,1]$ is

$$g_{warp}(x) = g(\gamma(x))\dot{\gamma}(x), \quad x \in [0,1] \tag{1}$$

* PhD student; Email: 13B933002@hit.edu.cn

where $\dot{\gamma}(x) = \frac{d}{dx}\gamma(x)$.

3. The related work: conditional density estimation using warping functions

The related work is the estimation of the conditional density function given the observation of a correlated random variable. Follow the description in [1], let X be a d-dimensional random variable, such as $X \sim \mathrm{U}[0,1]^d$, and Y be another random variable that correlated with X.

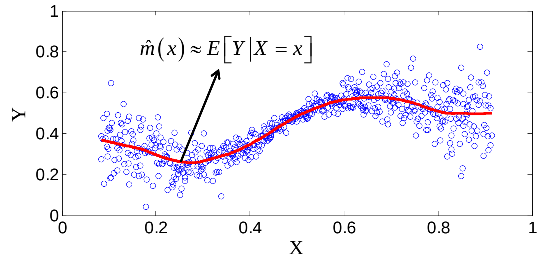

Fig. 2 Observed samples of two correlated random variables X and Y with estimated mean function $\hat{m}(x) \approx E[Y|X=x]$

Suppose we have observed n pairs realizations of X and Y, i.e. $\{(X_i, Y_i)\}_{i=1}^n$. Let $f_{y|X=x_0}$ be the conditional density function of Y given $X = x_0$, one of the tasks in [1] is to obtain a warped estimate for $f_{y|X=x_0}$, i.e. $\hat{f}_{y|X=x_0} = f_{w,x_0}$, from the initial density function $f_p\left(y|\hat{m}(x_0), \hat{\sigma}^2\right)$, where $\hat{m}(x)$ is the estimated mean function of Y obtained by local linear regression (see Fig. 2), $\hat{\sigma}^2$ is the estimated variance of the residuals $\{Y_i - \hat{m}(X_i)\}_{i=1}^n$. The initial density function f_p can be any parametric family with simple estimation procedure. Suppose $\gamma_0(x)$ is the warping function from $f_p\left(y|\hat{m}(x_0), \hat{\sigma}^2\right)$ to f_{w,x_0}, i.e.

$$f_{w,x_0}(y|X=x_0) = f_p\left(\gamma_0(y)|m(x_0), \hat{\sigma}^2\right)\dot{\gamma}_0(y) \tag{2}$$

The optimal estimate for γ_0 in [1] is obtained by the following weighted likelihood method

$$\hat{\gamma}_0 = \arg\max_{\gamma\in\Gamma}\left(\sum_{i=1}^{n}\log\left(f_p\left(\gamma(y_i)|\hat{m}(x_i), \hat{\sigma}^2\right)\dot{\gamma}(y_i)\right)W_{x_0,i}\right) \tag{3}$$

where $f_p\left(\cdot|\hat{m}(x_i), \hat{\sigma}^2\right)$ is the initial density function for $Y|X = x_i$, $W_{x_0,i}$ is the local weight at X_i, calculated by

$$W_{x_0,i} = \frac{N\left(\|X_i - x_0\|_2/h\,|0,1\right)}{\sum_{j=1}^{n} N\left(\|X_j - x_0\|_2/h\,|0,1\right)} \tag{4}$$

where $N(\cdot|\ 0,1)$ is the standard normal probability density function, h is the bandwidth parameter, $\|\cdot\|_2$ is the 2-norm.

This warping transformation-based approach for conditional distribution estimation in [1] can be regarded as real-value to distribution regression (if X is an univariate random variable) or vector-value to distribution regression (if X is a multivariate random variable). It can predict the dynamic change of the density function of Y only when random variables X and Y are correlated with each other.

4. Extend to distribution to distribution regression

In this section, the warping transformation-based method is extended to distribution to distribution regression. For convenience, let $\boldsymbol{\pi}_g$ and $\boldsymbol{\pi}_f$ respectively be two correlated one-dimensional continuous distribution classes with strictly positive support on $[0,1]$. Suppose we have obtained n pairs of probability density functions respectively from $\boldsymbol{\pi}_g$ and $\boldsymbol{\pi}_f$, i.e. $\{g_i,\ f_i\}_{i=1}^n$, given a new density function g_0 from $\boldsymbol{\pi}_g$, the task in this section is to develop a nonparametric regression model to predict the corresponding density function f_0 from $\boldsymbol{\pi}_f$, i.e. use $\{g_i,\ f_i\}_{i=1}^n \bigcup g_0$ to predict f_0. For this purpose, a nonparametric distribution to warping function regression is first used to predict the warping function γ_0, i.e. the mapping relationship from g_0 to f_0, then use the predicted warping function to transform g_0 to obtain a prediction for f_0, i.e.

$$\hat{\gamma}_0 = \sum_{k=1}^{n} \hat{\gamma}_k \frac{K\left(\delta\left(g_0, g_k\right)/h\right)}{\sum_{j=1}^{n} K\left(\delta\left(g_0, g_j\right)/h\right)} \tag{5a}$$

$$\hat{f}_0\left(x\right) = g_0\left(\hat{\gamma}_0\left(x\right)\right)\dot{\hat{\gamma}}_0\left(x\right) \tag{5b}$$

where, $\hat{\gamma}_k$ is the estimated warping function from g_k to f_k, i.e. $f_k\left(x\right) \approx g_k\left(\hat{\gamma}_k\left(x\right)\right)\dot{\hat{\gamma}}_k\left(x\right)$, $K\left(\cdot\right)$ is the kernel function, h is the bandwidth for the kernel regression, $\delta\left(g_0, g_k\right)$ is a metric (such as the L_1 distance: $\delta\left(g_0, g_k\right) = \int\left|g_0\left(\tau\right) - g_k\left(\tau\right)\right| d\tau$) used to measure the similarity between g_0 and g_k, $\hat{f}_0\left(x\right)$ is the prediction of $f_0\left(x\right)$. Note, a convex combination of warping functions is also a warping function [8], thus the regression result in Eq. (5a) being a warping function is guaranteed.

This distribution prediction approach can be regarded as an indirect nonparametric distribution to distribution regression. Unlike the conventional distribution to distribution regression proposed by Oliva *et al.* [2], the proposed regression model in Eq. (5) can reflect the shape mapping relationship between input and output distributions, thus it has more potential in extrapolating prediction.

5. Conclusions

An indirect nonparametric distribution to distribution regression method is proposed in this article, which can reflect the shape mapping relationship between input and output distributions. The correlation between two distribution classes is distinct from the correlation between two random variables. The real-value (or vector-value) to distribution regression and the distribution to distribution regression have different application scopes, the former is suited to correlated random variables, while the latter is suited to correlated distributions.

6. Acknowledgments

This work is financially supported by the National Basic Research Program of China (Grant No. 2013CB036305), the National Natural Science Foundation of China (Grant No. 51638007), the Ministry of Science and Technology of China (Grant No. 2015DFG82080) and the China Scholarship Council (CSC).

References

[1] Dasgupta S, Pati D and Srivastava A. A Geometric Framework For Density Modeling. arXiv preprint arXiv:1701.05656; 2017.

[2] Oliva JB, Póczos B and Schneider JG. Distribution to Distribution Regression. In ICML (3) 2013; pp: 1049-1057.

[3] Póczos B, Singh A, Rinaldo A and Wasserman LA. Distribution-Free Distribution Regression. In AISTATS 2013; pp: 507-515.

[4] Oliva JB, Neiswanger W, Póczos B, Schneider JG and Xing EP. Fast Distribution To Real Regression. In AISTATS (Vol. 3, p. 2) 2014.

[5] Szabó Z, Gretton A, Póczos B and Sriperumbudur B. Consistent, two-stage sampled distribution regression via mean embedding. arXiv preprint arXiv:1402.1754; 2014.

[6] Szabó Z, Gretton A, Póczos B and Sriperumbudur BK. Two-stage sampled learning theory on distributions. In AISTATS 2015.

[7] Szabó Z, Sriperumbudur B, Póczos B and Gretton, A. Learning theory for distribution regression. Journal of Machine Learning Research 2016; 17(152), 1-40.

[8] Srivastava A and Klassen EP. Functional and Shape Data Analysis. Springer: New York; 2016.