Distribution-Free Distribution Regression

Barnabás Póczos, Alessandro Rinaldo, Aarti Singh and Larry Wasserman

AISTATS 2013

Presented by Esther Salazar
Duke University

February 28, 2014

Outline

Regression Models

- **Standar regression model:** predict a real-valued response $Y_i \in \mathbb{R}$ from a covariate vector $X_i \in \mathbb{R}^d$,

$$Y_i = \boldsymbol{\beta} X_i + \mu_i$$

- **An extension: functional regression**. The covariate is a function instead of a finite dimensional vector

$$Y_i = f(X_i) + \mu_i$$

- **Distribution regression:** the covariate is a probability distribution P_i, i.e.

$$Y_i = f(P_i) + \mu_i$$

where f is an unknown function and μ is a random error

Distribution Regression

Model: $Y_i = f(P_i) + \mu_i$

This differs from functional regression in two ways:

1. P is probability measure on $\mathbb{R}^k$ rather than a one-dimensional function
2. *Important:* We do not observe the covariate P directly. We observe a sample from P

Distribution Regression

Model: $Y_i = f(P_i) + \mu_i$

Example: Patient classification

- Suppose we have m patients, the goal is to predict the class label

$$Y_i \in \{\text{“healthy”}, \text{“diseased”}\}, \quad i = 1, \ldots, m$$

- Traditional machine learning based approach: using feature vectors $X_i \in \mathbb{R}^d$ we can apply a standard classifier to predict the class labels of the feature vectors
- Problem: features based on heart rate, blood pressure, and other medical conditions in our body are always changing
- Let the set of these measurements be $\{X_{i,1}, \ldots, X_{i,n_i}\}$ where $X_{i,n_i} \in \mathbb{R}^d$
- We could construct a new feature vector $\tilde{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}$ (but we lose information)
- In distribution regression model we say that each person is presented by an unknown distribution P_i so $X_{i,j} \sim P_i$,

Goal: Classify these unknown P_i distributions

Regression problem

Regression problem with variables $(P_1, Y_1), \ldots, (P_m, Y_m)$ where $Y_i \in \mathbb{R}$ and each P_i is a probability distribution

$$Y_i = f(P_i) + \mu_i, \quad i = 1, \ldots, m$$

- f is a function (defined later)
- μ_i is a noise variable with mean 0
- We do not observe P_i but we observe a sample $X_{i,1}, \ldots, X_{i,n_i} \sim P_i$
- Observed data: $(\mathcal{X}_1, Y_1), \ldots, (\mathcal{X}_m, Y_m)$, where $\mathcal{X}_i = \{X_{i,1}, \ldots, X_{i,n_i}\}$
- Goal is to predict a new Y_{m+1} from a new batch $\mathcal{X}_{m+1}$ drawn from a new distribution P_{m+1}

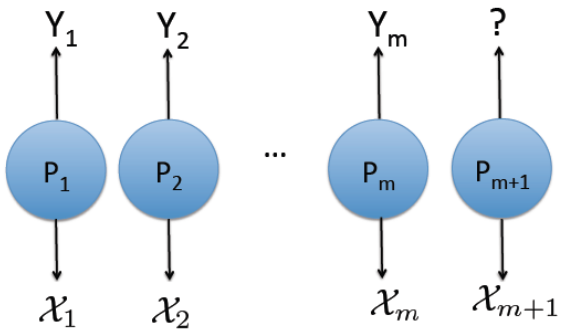


Figure 1: Illustration of the model - distributions $P_1, \ldots, P_m, P_{m+1}$ are unobserved, only the $\mathcal{X}_1, \ldots, \mathcal{X}_m, \mathcal{X}_{m+1}$ sample sets are observable.

The kernel-kernel estimator

The authors assume that the distributions P_i are an i.i.d sample from a measure $\mathcal{P}$

$$P_1, \ldots, P_m, P_{m+1} \overset{i.i.d}{\sim} \mathcal{P}.$$

Estimator $\hat{f}$ for the unknown function f:
Let $\hat{P}_i$ denote an estimator of P_i and let $\mathcal{X}$ be sample from a new distribution $P = P_{m+1}$. Predictor for Y_{m+1} is $\hat{Y}_{m+1} = \hat{f}(\hat{P}_{m+1})$

Given a bandwidth $h > 0$ and a kernel function K (whose properties will be specified later), we define

$$\begin{aligned}\widehat{f}(\widehat{P}) &= \widehat{f}(\widehat{P}; \widehat{P}_1, \ldots, \widehat{P}_m) \\ &= \begin{cases} \dfrac{\sum_i Y_i K\left(\frac{D(\widehat{P}_i, \widehat{P})}{h}\right)}{\sum_i K\left(\frac{D(\widehat{P}_i, \widehat{P})}{h}\right)} & \text{if } \sum_i K\left(\frac{D(\widehat{P}_i, \widehat{P})}{h}\right) > 0 \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

To complete the definition we need to specify $\hat{P}_i$, $\hat{P}$ and D

- The estimate P_i, more precisely, the density p_i of P_i with a kernel density estimator

$$\widehat{p}_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{b_i^k} B\left(\frac{\|x - X_{ij}\|_2}{b_i}\right)$$

  B is an appropriate kernel function (Tsybakov,2010) with bandwidth $b_i > 0$ and $\hat{P}_i$ is defined by

$$\hat{P}_i(A) = \int_A \hat{p}_i(u)du$$

- For D: for any two probabilities in P and Q they take $D(P, Q)$ to be the L_1 distance of their densities $D(P, Q) = \|p - q\| = \int |p(x) - q(x)|dx$

Therefore, the kernel-kernel estimator is

$$\widehat{f}(\widehat{P}) = \widehat{f}(\widehat{P}; \widehat{P}_1, \ldots, \widehat{P}_m) = \frac{\sum_{i=1}^m Y_i K\left(\frac{||\widehat{p} - \widehat{p}_i||}{h}\right)}{\sum_{i=1}^m K\left(\frac{||\widehat{p} - \widehat{p}_i||}{h}\right)}$$

Assumptions

- *(A1) Hölder continuous functional.* The unknown functional f belongs to the class $\mathcal{M} = \mathcal{M}(L, \beta, D)$ of Hölder continuous functionals on $\mathbb{D}$:

$$\mathcal{M} = \left\{ f : \ |f(P_i) - f(P_j)| \leq L\, D(P_i, P_j)^{\beta} \right\},$$

  for some $L > 0$ and $0 < \beta \leq 1$, where D is the above specified L_1 metric on $\mathbb{D}$. In the $\beta = 1$ special case this means that f is Lipschitz continuous.

- *(A2) Asymmetric boxed and Lipschitz kernel.* The kernel K satisfies the following properties: $K : [0, \infty] \to \mathbb{R}$ is non-negative and Lipschitz continuous with Lipschitz constant L_K. In addition, there exist constants $0 < \underline{K} < 1$ and $0 < r < R < \infty$ such that, for all $x > 0$, it holds that

$$\underline{K} I_{\{x \in \mathcal{B}(0,r)\}} \leq K(x) \leq I_{\{x \in \mathcal{B}(0,R)\}}.$$

Upper bound on risk

Concern: upper bounding the risk

$$R(m,n) = \mathbb{E}\left[\left|\widehat{f}(\widehat{P};\widehat{P}_1,\ldots,\widehat{P}_m) - f(P)\right|\right],$$

Note that the absolute prediction risk is $\mathbb{E}|\hat{Y} - Y| \leq R(m,n) + c$ where $c = \mathbb{E}(|\mu|)$ is a constant

Main result:

Theorem 1 *Suppose that the assumptions (A1)-(A7) stated above hold. Then*

$$R(m,n) \leq \frac{1}{h}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right] C_1 n^{-\frac{1}{2+k}} + C_2 h^{\beta}$$
$$+ C_3\sqrt{\frac{1}{m}}\sqrt{\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]} + \frac{C_4}{m}\mathbb{E}\left[\frac{1}{\Phi_P(rh/2)}\right]$$
$$+ (m+1)e^{-\frac{1}{2}n^{\frac{k}{2+k}}},$$

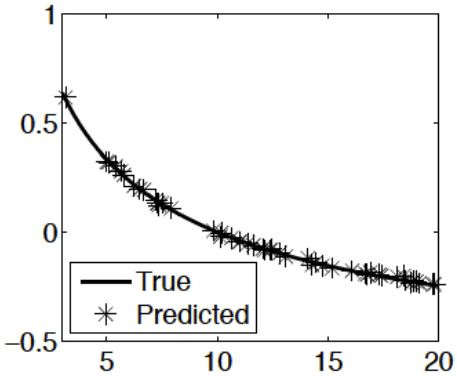
where the constants C_i's are specified in the proof.

Numerical illustration

- They assume $n = n_1 = \ldots = n_m = 500$ (set sizes)
- $m = 250$ sample sets for training, 25 for validation and 50 for testing
- Each sample set contained $n = 500$ Beta$(a, 3)$ distributed i.i.d. points
- Task: learn the skewness of Beta(a, b) distributions, $f = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$

The estimator is not aware of that the sample sets are coming from beta distributions and also it does not know the skewness function.

Kernel used: $k(x) = 1 - |x|$ if $-1 \leq x \leq 1$, and 0 otherwise



(a) Skewness of Beta

Figure 2: (a) Learned skewness of $Beta(a, 3)$ distribution. Axis x: parameter a in $[3, 20]$. Axis y: skewness of $Beta(a, 3)$.