SZTAKI

Semi-Parametric Uncertainty Bounds for Binary Classification

Balázs Csanád Csáji & Ambrus Tamás

SZTAKI: Institute for Computer Science and Control, Budapest, Hungary

58th IEEE Conference on Decision and Control, Nice, France, 2019

Binary Classification

- In binary classification the sample $\{(x_j, y_j)\}_{j=1}^{n}$ consists of inputs, $x_j \in \mathbb{X}$, from a measurable space, and labels, $y_i \in \mathbb{Y} \doteq \{-1, +1\}$.
- The sample is i.i.d. and have (unknown) distribution $\mathbb{P}$ on $\mathbb{X} \times \mathbb{Y}$.
- We call any (measurable) $g : \mathbb{X} \to \{-1, +1\}$ function a classifier.
- A loss function penalizes label mismatch, $\ell : \mathbb{Y} \times \mathbb{Y} \to [0, \infty)$.
- Typical choice: zero-one loss, $\ell(\hat{y}, y) \doteq \mathbb{I}(\,\hat{y} \neq y\,) = (1 - \hat{y}y)/2$.
- The overall (expected) risk of classifier g is (cf. "test error")

$$R(f) \doteq \mathbb{E}\big[\,\ell(g(X), Y)\,\big] = \int_{\mathbb{X}\times\mathbb{Y}} \ell(g(x), y)\, \mathbb{P}(\mathrm{d}x, \mathrm{d}y),$$

  where X and Y are general random elements with $(X, Y) \sim \mathbb{P}$.
- For the zero-one loss, the risk is simply $R(f) = \mathbb{P}(g(X) \neq Y)$.
- In general, we aim at finding a classifier with minimal risk.

Regression Function

- If distribution $\mathbb{P}$ was known, an ideal choice would be

$$g_* \in \arg\min \left\{ R(f) \mid g : \mathbb{X} \to \mathbb{Y} \text{ and } g \text{ is measurable} \right\},$$

  called Bayes optimal or target classifier (not unique in general).
- For the zero-one loss, an optimal classifier is (if $\mathbb{P}(\eta(x) \neq 0) = 1$)

$$g_*(x) = \operatorname{sign}(\, f_*(x)\,), \qquad \text{where} \qquad f_*(x) \doteq \mathbb{E}\big[\, Y \,|\, X = x \,\big].$$

- Function f_* is a key object, it is called the regression function.
- Note that it contains more information than g_*, as for example the probability of misclassification can also be computed from f_*.
- There are many methods that provide point estimates for f_*, but there are much less that can efficiently build region estimates for it.
- Here, we aim at building non-asymptotic region estimates for f_*.

Reproducing Kernel Hilbert Spaces

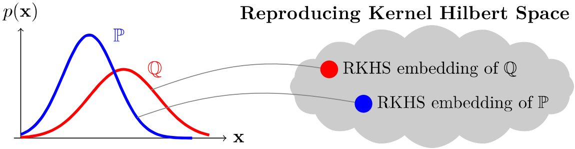
- A Hilbert space, $\mathcal{H}$, of functions $f : \mathcal{X} \to \mathbb{R}$, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, is called a Reproducing Kernel Hilbert Space (RKHS), if $\forall\, z \in \mathcal{X}$ the point evaluation (Dirac) functional $\delta_z : f \to f(z)$ is bounded (that is $\exists\, \kappa > 0$ with $|\delta_z(f)| \leq \kappa \, \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$).
- Then, one can construct a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, having the reproducing property that is for all $z \in \mathcal{X}$ and $f \in \mathcal{H}$, we have

$$\langle\, k(\cdot, z), f \,\rangle_{\mathcal{H}} \,=\, f(z),$$

  which is ensured by the Riesz-Fréchet representation theorem.
- As a special case, the kernel satisfies $k(z, s) = \langle\, k(\cdot, z), k(\cdot, s) \,\rangle_{\mathcal{H}}$.
- A kernel is therefore a symmetric and positive-definite function.
- Conversely, by the Moore-Aronszajn theorem, for every symmetric and positive definite function, there uniquely exists an RKHS.

Kernel Mean Embedding

- Idea: map distributions to elements of an RKHS with the kernel.


$p(\mathbf{x})$
$\mathbb{P}$
$\mathbb{Q}$
$\mathbf{x}$
Reproducing Kernel Hilbert Space
RKHS embedding of $\mathbb{Q}$
RKHS embedding of $\mathbb{P}$

Source: Muandet *et al.*: Kernel Mean Embedding of Distributions: A Review and Beyond, Now Publishers, 2017

- $\mathcal{D}(\mathbb{X})$ is the set of prob. distributions over meas. space $(\mathbb{X}, \Sigma)$.
- The kernel mean embedding of probability measures into an RKHS $\mathcal{H}$ endowed with a reproducing kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is

$$\mu : \mathcal{D}(\mathbb{X}) \;\to\; \mathcal{H},$$

$$\mathbb{P} \;\to\; \int_{\mathbb{X}} k(x, \cdot)\, \mathbb{P}(\mathrm{d}x),$$

if this Bochner integral exists, e.g., if $\mathbb{E}_{X \sim \mathbb{P}}\Big[\sqrt{k(X,X)}\,\Big] < \infty$.

Universal and Characteristic Kernels

- The kernel embedding has many nice properties, e.g., for $f \in \mathcal{H}$,

$$\mathbb{E}_{X \sim P}\big[\, f(X) \,\big] \;=\; \langle f, \mu_P \rangle_{\mathcal{H}}$$

- If $k(x, y) = \exp(\langle x, y \rangle)$, then we recover the moment generating function (with the Fourier kernel we get the characteristic funct.).
- A kernel is called characteristic if the embedding, μ, is injective.
- A characteristic kernel induces a metric on space $\mathcal{D}(\mathbb{X})$, namely, $d(P, Q) \doteq \| \mu_P - \mu_Q \|_{\mathcal{H}}$, with $d(P, Q) = 0$ if and only if $P = Q$.
- $\mathcal{C}(\mathbb{X})$ is the set of continuous fun. on a compact metric space $\mathbb{X}$.
- A kernel is universal if the corresponding $\mathcal{H}$ is dense in $\mathcal{C}(\mathbb{X})$: for all $f \in \mathcal{C}(\mathbb{X})$ and $\varepsilon > 0$ there is $h \in \mathcal{H}$ such that $\| f - h \|_{\infty} < \varepsilon$.
- Let $\mathbb{X}$ be a compact metric space and let k be a universal kernel on $\mathbb{X}$, then one can show that k is also characteristic.

Examples of Kernels

Kernel	$k(x, y)$	Domain	U	C
Gaussian	$\exp\left(\frac{-\lVert x-y \rVert_2^2}{\sigma}\right)$	$\mathbb{R}^d$	✓	✓
Linear	$\langle x, y \rangle$	$\mathbb{R}^d$	×	×
Polynomial	$(\langle x, y \rangle + c)^p$	$\mathbb{R}^d$	×	×
Laplacian	$\exp\left(\frac{-\lVert x-y \rVert_1}{\sigma}\right)$	$\mathbb{R}^d$	✓	✓
Rat. quadratic	$\exp(\lVert x - y \rVert_2^2 + c^2)^{-\beta}$	$\mathbb{R}^d$	✓	✓
Exponential	$\exp(\sigma \langle x, y \rangle)$	compact	×	✓
Poisson	$1/(1 - 2\alpha \cos(x - y) + \alpha^2)$	$[0, 2\pi)$	✓	✓

Figure: typical kernels; *U* means "universal" and *C* means "characteristic" (where the hyper-parameters satisfy $\sigma, \beta, c > 0$, $\alpha \in (0, 1)$ and $p \in \mathbb{N}$).

Resampling Framework

- Let us fix a distribution on $\mathbb{S} \doteq \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X}$ and $\mathbb{Y}$ are the input and output spaces, respectively (in our case $\mathbb{Y} = \{-1, +1\}$).
- The conditional expectation of Y given X can be expressed as

$$f_*(x) \doteq \mathbb{E}\big[\, Y \mid X = x \,\big] = 2 \cdot \mathbb{P}(\, Y = +1 \mid X = x \,) - 1.$$

- We are given an (indexed) family of possible regression functions that also contains f_* (the true system is in the model class), that is

$$f_* \in \mathcal{F} \doteq \big\{\, f_\theta : \mathbb{X} \to [\,-1, +1\,] \;\big|\; \theta \in \Theta \,\big\}.$$

- The true "parameter" is denoted by θ^*, namely, $f_{\theta^*} = f_*$.
- Assume that the parametrization is injective (in the $\mathcal{L}^2(\mathbb{X})$ sense).
- Otherwise, Θ can be an arbitrary set! ($\dim(\Theta) = \infty$ is allowed).

Resampling Labels

- The original i.i.d. input-output dataset is denoted by

$$\mathcal{D}_0 \doteq ((x_1, y_1), \dots, (x_n, y_n)).$$

- Given a θ, we can generate $m-1$ alternative samples by

$$\mathcal{D}_i(\theta) \doteq ((x_1, y_{i,1}(\theta)), \dots, (x_n, y_{i,n}(\theta))),$$

  for $i = 1, \dots, m-1$, where for all (i,j) label $y_{i,j}(\theta)$ is generated randomly according to the conditional distribution:

$$\mathbb{P}_\theta(\, Y = y \mid X = x\,) \doteq {}^1\!/_2 \left(y(f_\theta(x) + 1)\right).$$

Crucial Observations

- $\mathcal{D}_0$ and $\mathcal{D}_i(\theta^*)$ have the same distribution ("Law"), for i.
- If $\theta \neq \theta^*$, $\text{Law}(\mathcal{D}_0)$ is typically different than $\text{Law}(\mathcal{D}_i(\theta))$.

Ranking Functions

- Let $\mathbb{A}$ be a measurable space, a function $\psi : \mathbb{A}^m \to [\, m\,]$ where $[\, m\,] \doteq \{1, \dots, m\}$, is called a ranking function if for all $(a_1, \dots, a_m) \in \mathbb{A}^m$ it satisfies the two properties:

  (P1) For all permutations μ of the set $\{2, \dots, m\}$, we have

  $$\psi(\, a_1, a_2, \dots, a_m\,) \;=\; \psi(\, a_1, a_{\mu(2)}, \dots, a_{\mu(m)}\,),$$

  that is the function is invariant with respect to reordering the last $m-1$ terms of its arguments.

  (P2) For all $i, j \in [\, m\,]$, if $a_i \neq a_j$, then we have

  $$\psi(\, a_i, \{a_k\}_{k \neq i}\,) \;\neq\; \psi(\, a_j, \{a_k\}_{k \neq j}\,).$$

- We can think of ψ as a function which "sorts" the elements and returns the rank of the first element in the order.

Uniform Ordering of Exchangeable Elements

The Main Idea Underlying the Framework

Compare the original dataset with alternative samples randomly generated according to a given hypothesis. Accept the hypothesis if the original dataset behaves "similarly" to the alternative ones and reject otherwise. Measure "similar" behavior with ranking.

- Fundamental quation: how to find a suitable ranking function?

Uniform Ordering Lemma

Let $A_1, \dots, A_m$ be exchangeable, almost surely pairwise different random elements from $\mathbb{A}$. Then, $\psi\big(\, A_1, A_2, \dots, A_m \,\big)$ has discrete uniform distribution: $\forall\, k \in [\, m\,]$, the rank is k with probability $1/m$.

- Pairwise difference is a technical assumptions (cf. tie-breaking).

General Confidence Region Construction

- Given a ranking function ψ (i.e., satisfying P1 and P2).
- User-chosen hyper-parameters $p, q \in [\, m\,]$ with $p \leq q$.
- One can build a confidence region based on ψ by

Confidence Region

$$\Theta_{\varrho}^{\psi} \doteq \left\{\, \theta \in \Theta :\; p \leq \psi\big(\, \mathcal{D}_0, \{\mathcal{D}_k(\theta)\}_{k \neq 0}\,\big) \leq q \,\right\}$$

- $\varrho \doteq (m, p, q)$ denotes the applied hyper-parameters, with $m \geq 1$ being the total number of samples (original & alternative datasets).
- Intuitively: the region contains those models for which the rank of the original dataset compared to the ranks of the alternative ones, generated based on the model, is neither too low nor too high.

Exact Confidence

- The main abstract result of the resampling framework is:

Theorem: Exact Confidence

We have for all ranking function ψ and hyper-parameter $\varrho = (m, p, q)$ with integers $1 \leq p \leq q \leq m$ that

$$\mathbb{P}\big(\, \theta^* \in \Theta_{\varrho}^{\psi} \,\big) \; = \; \frac{q - p + 1}{m}.$$

- Note that ψ is an arbitrary ranking function (satisfying P1 and P2).
- The coverage probabilty is user-chosen (rational), and exact.
- This probability is independent of the underlying probability distribution generating the data, the result is distribution-free.
- Further, the claim is non-asymptotic (holds for finite samples).

Strong Consistency

- Warning: exact confidence in itself could be misleading as, for example, purely randomized methods can have this property.
- We also study other properties of the methods, e.g., consistency.
- Formally, a method is strongly consistent if

$$\mathbb{P}\bigg(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \Big\{ \theta \in \Theta_{\varrho,n}^{\psi} \Big\} \bigg) = 0,$$

  for all parameter $\theta \neq \theta^*$, $\theta \in \Theta$, where $\Theta_{\varrho,n}^{\psi}$ denotes the confidence region constructed based on a sample of size n.
- Informally: eventually, as the sample size tends to infinity, any false parameter will be excluded from the regions with probability one.

Kernel-Based Constructions

- Now, we propose three kernel-based algorithms:

1. A neighborhood based (Algorithm I)
2. An embedding based (Algorithm II)
3. A discrepancy based (Algorithm III)

- Each of these methods builds region-estimates (confidence regions) for the underlying regression function of binary classification.
- They are based on the suggested resampling framework and all of them have exact coverage probabilities and are strongly consistent.

Algorithm I: Neighborhood Based

- If there is a metric on the input space, $\mathbb{X}$, we can estimate f_* based on the original dataset by the kNN (k-nearest neighbors) algorithm.
- Similarly, we can estimate f_* based on the alternative datasets:

$$f_{\theta,n}^{(i)}(x) \doteq \frac{1}{k_n} \sum_{j=1}^{n} y_{i,j}(\theta)\, \mathbb{I}\big(\, x_j \in N(x, k_n)\,\big),$$

  for $i = 0, \dots, m-1$, where $\mathbb{I}$ is an indicator function (its value is 1 if its argument is true, and 0 otherwise), $N(x, k_n)$ denotes the k_n closest neighbors of x from $\{x_j\}_{j=1}^{n}$, and $k_n \leq n$ is a constant (window size), which can depend on the sample size n.
- Idea: we can construct a ranking function by comparing the "distances" of these functions from the model generating the data.

Algorithm I: Neighborhood Based

- The $\mathcal{L}^2(\mathbb{X})$ distance of the i th estimate from the model is

$$Z_n^{(i)}(\theta) \doteq \left\| f_{\theta,n}^{(i)} - f_\theta \right\|_2^2,$$

which can be calculated directly or by Monte Carlo approximations.

- Then, we can define the rank of $Z_n^{(0)}$ among $\{Z_n^{(i)}(\theta)\}$ as

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\left(Z_n^{(0)} \prec_\pi Z_n^{(i)}(\theta) \right),$$

where relation "$\prec_\pi$" is the standard "$<$" with random tie-breaking.

- Finally, the confidence region can be constructed as

$$\Theta_{\varrho,n}^{(1)} \doteq \left\{ \theta \in \Theta : \mathcal{R}_n(\theta) \leq q \right\}.$$

Algorithm I: Neighborhood Based

Theorem: Stochastic Guarantees of Algorithm I

Assume that the following properties hold

1. *The input space is $\mathbb{X} \subseteq \mathbb{R}^d$ and $\mathbb{X}$ is compact.*
2. *The support of the input distribution, $P_{\mathbb{X}}$, is the whole $\mathbb{X}$.*
3. *The input distribution, $P_{\mathbb{X}}$, is absolutely continuous.*

Then, the coverage probability of the constructed region is

$$\mathbb{P}\big(\, \theta^* \in \Theta^{(1)}_{\varrho,n} \,\big) \,=\, q \,/\, m,$$

i.e., it is exact for any sample size n. Moreover, if $\{k_n\}$ are chosen such that $k_n \to \infty$ and $k_n/n \to 0$, as $n \to \infty$, then the confidence regions are strongly consistent (eventually exclude false parameters).

Algorithm II: Embedding Based

- Idea: embed the distribution of the original sample and that of the alternative ones in an RKHS using a characteristic kernel.
- The kernel mean embedding of the true distribution generating the data $(*)$ and the one based on a hypothetical model (θ) are

$$h_*(\cdot) \doteq \mathbb{E}\big[\, k(\cdot, S_*)\,\big], \qquad \text{and} \qquad h_\theta(\cdot) \doteq \mathbb{E}\big[\, k(\cdot, S_\theta)\,\big],$$

where S_* and S_θ are random elements from $\mathbb{S} = \mathbb{R}^d \times \{+1, -1\}$, distributed according to true distribution and the tested one.

- Functions $h_*(\cdot)$ and $h_\theta(\cdot)$ can be estimated from empirical data:

$$h^{(i)}_{\theta,n}(\cdot) \doteq \frac{1}{n}\sum_{j=1}^{n} k(\cdot, s_{i,j}(\theta)),$$

for $i = 0, \dots, m-1$, where $s_{i,j}(\theta) \doteq (x_j, y_{i,j}(\theta))$.

Algorithm II: Embedding Based

- The kernel is characteristic, therefore, $h_\theta = h_* \iff \theta = \theta^*$.
- The construction of the confidence region is as follows

$$Z_n^{(i)}(\theta) \doteq \sum_{j=0}^{m-1} \| \, h_{\theta,n}^{(i)} - h_{\theta,n}^{(j)} \, \|_{\mathcal{H}}^2$$

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\big(\, Z_n^{(0)} \prec_\pi Z_n^{(i)}(\theta) \, \big)$$

$$\Theta_{\varrho,n}^{(2)} \doteq \big\{ \, \theta \in \Theta : \mathcal{R}_n(\theta) \leq q \, \big\}$$

- Note: cumulative distances are used in the definition of $\{Z_n^{(i)}(\theta)\}$.
- The terms $\| \, h_{\theta,n}^{(i)} - h_{\theta,n}^{(j)} \, \|_{\mathcal{H}}^2$ can be easily computed in practice using the Gram matrix (based on the reproducing property of the kernel).

Algorithm II: Embedding Based

Theorem: Stochastic Guarantees of Algorithm II

Assume that the following properties hold

1. *$\mathcal{H}$ is a separable RKHS containing $\mathbb{S} \to \mathbb{R}$ functions.*
2. *The kernel is (measurable) bounded and characteristic.*

Then, the confidence regions of Algorithm II have exact coverage

$$\mathbb{P}\big(\, \theta^* \in \Theta^{(2)}_{\varrho,n} \,\big) \;=\; q \,/\, m,$$

for any sample size n; and they are strongly consistent, if $m \geq 3$.

- One can show that $\mathrm{Var}(k(\cdot, S)) < \infty$, for $S \in \{S_*, S_\theta\}$, therefore a Hilbert space valued strong law of large numbers can be applied.
- Algorithm II is of theoretical interest as it is computationally heavy.

Algorithm III: Discrepancy Based

- In order to formalize the method, let us introduce residuals

$$\varepsilon_{i,j}(\theta) \doteq y_{i,j}(\theta) - f_\theta(x_j)$$

for $i = 0, \dots, m-1$ and $j = 1, \dots, n$. Note that if $i \neq 0$, $\varepsilon_{i,j}(\theta)$ has zero mean for all j, as $f_\theta(x_j) = \mathbb{E}_\theta\big[\, y_{i,j}(\theta) \,|\, x_j \,\big]$.

- Algorithm III constructs the confidence region as

$$Z_n^{(i)}(\theta) \doteq \Bigg\| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_{i,j}(\theta)\, k(\cdot, x_j) \Bigg\|_{\mathcal{H}}^2 = \frac{1}{n^2}\, \varepsilon_i^{\mathrm{T}}(\theta)\, K\, \varepsilon_i(\theta)$$

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\big(\, Z_n^{(0)} \prec_\pi Z_n^{(i)}(\theta) \,\big)$$

$$\Theta_{\varrho,n}^{(3)} \doteq \big\{\, \theta \in \Theta : \mathcal{R}_n(\theta) \leq q \,\big\}$$

Algorithm III: Discrepancy Based

Theorem: Stochastic Guarantees of Algorithm III

Assume that the following properties hold

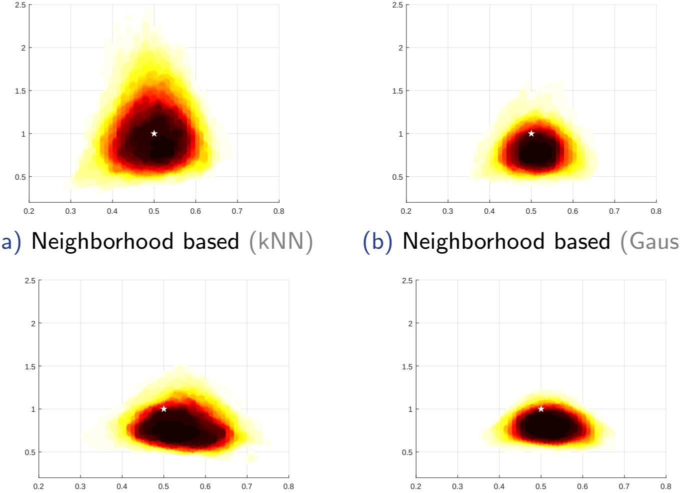
1. *$\mathcal{H}$ is a separable RKHS containing $\mathbb{X} \to \mathbb{R}$ functions.*
2. *The kernel is (measurable) bounded and universal.*
3. *$\mathbb{X}$ is a compact Polish metric space (complete and separable).*
4. *Each potential regression function $f \in \mathcal{F}$ is continuous.*

Then, the confidence regions of Algorithm III have exact coverage

$$\mathbb{P}\big(\,\theta^* \in \Theta^{(3)}_{\varrho,n}\,\big) \,=\, q\,/\,m,$$

for any sample size n; and they are strongly consistent.

Experiments: Ranks in the Parameter Space



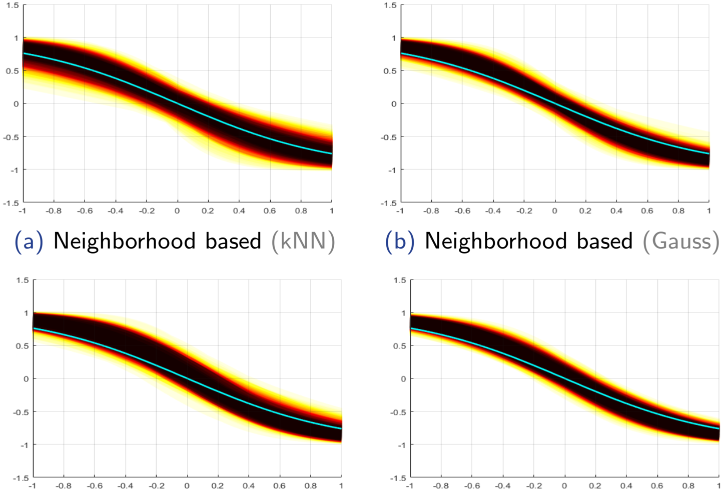
(a) Neighborhood based (kNN)

(b) Neighborhood based (Gauss)

(c) Embedding based (Gauss)

(d) Discrepancy based (Gauss)

Experiments: Ranks in the Model Space



(a) Neighborhood based (kNN)

(b) Neighborhood based (Gauss)

(c) Embedding based (Gauss)

(d) Discrepancy based (Gauss)

Conclusions

- The regression function is a key object of binary classification, as it can provide an optimal classifier and can also evaluate the risk.
- We aimed at designing region estimates for the regression function.
- A general framework based on resampling was presented with which confidence regions with exact coverage can be built.
- A general non-asymptotic theorem ensuring this was provided.
- The approach is semi-parametric as the regression function does not contain information about the marginal distribution of inputs.
- Three particular kernel-based (resampling) methods were suggested based on neighborhoods, (mean) embeddings and discrepancy.
- Besides having exact coverage probabilities, we argued that each of these methods is also strongly consistent (under mild assumptions).
- Finally, numerical experiments were shown supporting the ideas.

Thank you for your attention!

✉ csaji@sztaki.hu