Distribution-Free Prediction Intervals for Kernel Regression

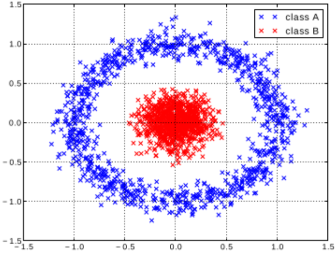
László Keresztes

Applied Mathematics MSc, ELTE-TTK

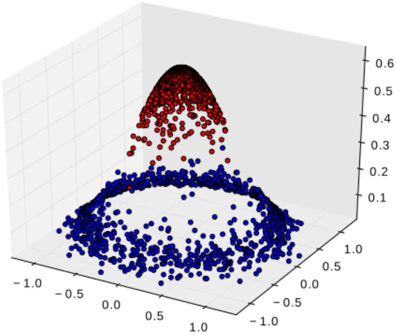
Supervisor: Balázs Csanád Csáji, SZTAKI

December 16, 2020

Kernel methods



(a) A non linearly separable dataset.



(b) Possible feature space representation.

Figure: Example of a feature space in binary classification

Kernel methods

Gram matrix

Given a function $k : \mathbb{X}^2 \to \mathbb{R}$ and patterns $x_1, \dots, x_m \in \mathbb{X}$, the $m \times m$ matrix K with elements $K_{i,j} = k(x_i, x_j)$ is called the Gram matrix (or kernel matrix) of k with respect to the data points $x_1, \dots, x_m \in \mathbb{X}$.

Positive definite kernel

Let $\mathbb{X}$ be a nonempty set. A symmetric $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ function which $\forall m$ $\forall x_1, \dots x_m \in \mathbb{X}$ points gives a positive semi-definite Gram matrix is called a positive definite kernel (or kernel). If for all m and distinct $\{x_i\}$ the Gram matrix is positive definite, the kernel is called strictly positive definite.

Kernel methods

Reproducing Kernel Hilbert Spaces

Let $\mathbb{X}$ be a nonempty set and $\mathbb{H}$ a Hilbert space of functions $f : \mathbb{X} \to \mathbb{R}$ endowed with the dot product $\langle \cdot, \cdot \rangle$. Then $\mathbb{H}$ is called a Reproducing Kernel Hilbert Space if $\exists k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ with the following properties:

1. k has the reproducing property
   $\langle f, k(\cdot, x) \rangle = f(x) \; \forall f \in \mathbb{H} \; \forall x \in \mathbb{X}$
2. k spans $\mathbb{H}$
   $\mathbb{H} = \overline{span\{k(x, \cdot) : x \in \mathbb{X}\}}$

Representer Theorem

Let $\mathbb{H}$ be a Reproducing Kernel Hilbert Space associated to the kernel k. Denote by $\Omega : [0, \infty] \to \mathbb{R}$ a strictly monotonic increasing function, by $\mathbb{X}$ a set, and by $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ an arbitrary loss function. Then each minimizer $\hat{f} \in \mathbb{H}$ of the regularized risk $\frac{1}{m} \sum_{i=1}^{m} \ell(\hat{f}(x_i), y_i) + \Omega(\|\hat{f}\|_{\mathbb{H}})$ admits a representation of the form $\hat{f}(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i)$

Split Conformal Prediction

Algorithm 1: Split Conformal Prediction

Input : Data $(X_i, Y_i), i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm $\mathcal{A}$

Output: Prediction band, over $x \in \mathbb{R}^d$

Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1, \mathcal{I}_2$

$\hat{\mu} = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$

$R_i = |Y_i - \hat{\mu}(X_i)|, i \in \mathcal{I}_2$

$d =$ the kth smallest value in $\{R_i : i \in \mathcal{I}_2\}$, where

$k = \lceil (n/2 + 1)(1 - \alpha) \rceil$

Return $C_{split}(x) = [\hat{\mu}(x) - d, \hat{\mu}(x) + d]$

Split Conformal Prediction

Stochastic guarantees for Split Conformal Prediction

If $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d., then for a new i.i.d. draw (X_{n+1}, Y_{n+1})

$$\mathbb{P}(Y_{n+1} \in C_{split}(X_{n+1})) \geq 1 - \alpha,$$

for the split conformal prediction band C_{split} constructed in Algorithm 1. Moreover, if we assume additionally that the residuals $R_i, i \in \mathcal{I}_2$ have a continuous joint distribution, then

$$\mathbb{P}(Y_{n+1} \in C_{split}(X_{n+1})) \leq 1 - \alpha + \frac{2}{n+2}$$

Experiments

Regression problem: $f(x) = xsin(16x)$, sample: $(X_i, Y_i), i = 1, \ldots, m$ i.i.d, where $X_i \sim U([0,1])$ and $Y_i = f(X_i) + N_i$, $N_i \sim Laplace(0, b)$ i.i.d.