Non-parametric Models for Non-Negative Functions

Ulysse Marteau-Ferey, Francis Bach, Alessandro Rudi

NEURAL INFORMATION PROCESSING SYSTEMS

Models for machine learning tasks

Inputs : space $\mathcal{X}$, observed data points $(x_i)_{1\leq i\leq n} \in \mathcal{X}^n$

Goal : find $f_\star : \mathcal{X} \to \mathbb{R}$ by solving

$$f_\star \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \ell_i(f(x_i)) + \Omega(f)$$

Losses : ℓ_i adapted to the task

Model : $\mathcal{F}$ class of functions and Ω is a penalty on $\mathcal{F}$

Linear Models : a simple model with great properties

Feature map $\varphi : \mathcal{X} \to \mathcal{H}$. It can be :

explicit : $\varphi(x) \in \mathbb{R}^d$ is an engineered feature

implicit : defined by a given positive semi-definite kernel $k(x, y)$: $\varphi(x)^\top\varphi(y) = k(x, y)$. $\mathcal{H}$ can be infinite dimensional

Model : f is parametrised by $\theta \in \mathcal{H}$

$$f_\theta(x) = \langle\theta, \varphi(x)\rangle_{\mathcal{H}} = \theta^\top\varphi(x)$$

Goal : find the best *linear* predictor $f_{\theta_\star}$ by solving

$$\theta_\star \arg\min_{\theta\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell_i(\theta^\top\varphi(x_i)) + \frac{\lambda}{2}\|\theta\|^2$$

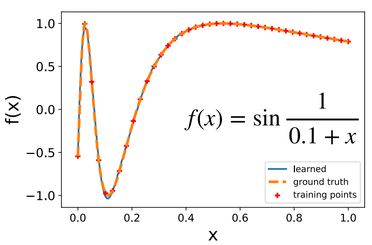
Key properties of linear models

1) They preserve the **convexity** of the ℓ_i

2) They have **good approximation properties :**

Example : gaussian kernel of width σ :

$$k_\sigma(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$



$\mathcal{H}$ **infinite dimensional**

We can approximate **any continuous function**

3) A **finite n-dimensional representation** $\theta_\star = \sum_{i=1}^n \alpha_i\varphi(x_i)$

$$\alpha_\star \arg\min_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \ell_i([K\alpha]_i) + \frac{\lambda}{2}\alpha^\top K\alpha, \qquad K = (k(x_i, x_j))_{1\leq i,j\leq n}$$

What if we want to learn non-negative functions ?

$$f_\star \in \arg\min_{\substack{f\in\mathcal{F} \\ \forall x\in\mathcal{X},\ f(x)\geq 0}} \frac{1}{n}\sum_{i=1}^n \ell_i(f(x_i)) + \Omega(f)$$

Existing methods have issues

1) Generalised linear models (GLM)

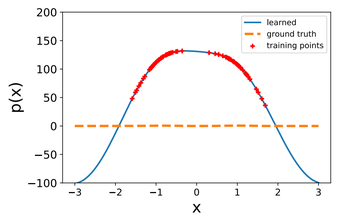
$$g_\theta(x) = \exp(\theta^\top\varphi(x))$$

2) Linear models (LM)

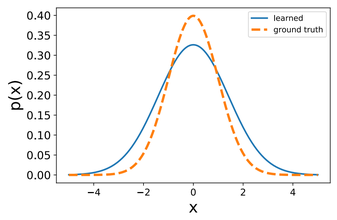
$$\theta^\top\varphi(\tilde{x}_j) \geq 0,\ 1 \leq j \leq M,\ (\tilde{x}_j) \text{ finite grid}$$

3) Nadarawa-Watson type estimators (NW)

$$f_\alpha(x) = \sum_{i=1}^n \alpha_i k(x_i, x),\ k \geq 0,\ \alpha \geq 0$$



Example 1 : Non-negativity issues using a linear model to learn a density



Example 2 : Approximation issues using NW to learn a gaussian of width σ_1 with a gaussian kernel of k_{σ_2} with $\sigma_2 > \sigma_1$

Properties	GLM	NW	LM	Our model
Non-negativity	✓	✓	✗	✓
Convexity	✗	✓	✓	✓
Approximation	✓	✗	✓	✓
Finite dimensional	✓	✓	✓	✓

Our model for non-negative functions

Intuition

$$g_\theta(x) = f_\theta(x)^2 = (\theta^\top\varphi(x))^2$$

- ✓ Non negativity
- ✓ Finite dimensional representation
- ✓ Approximation properties
- ✗ Convexity of the loss function

We circumvent this problem with a matrix A

- ✗ Quadratic : $g_\theta(x) = \varphi(x)^\top\theta\theta^\top\varphi(x)$
- ✓ Linear : $f_A(x) = \varphi(x)^\top A\varphi(x),\ A \in \mathcal{S}(\mathcal{H}),\ A \succeq 0$

$$A \succeq 0 \implies \forall x \in \mathcal{X},\ f_A(x) \geq 0$$

Our model has the nice properties of linear models

$$A_\star \in \arg\min_{A\geq 0} \frac{1}{n}\sum_{i=1}^n \ell_i(\varphi(x_i)^\top A\varphi(x_i)) + \Omega(A)$$

1) They preserve the **convexity** of the ℓ_i

2) They have **good approximation properties**

3) A **finite n^2-dimensional representation**

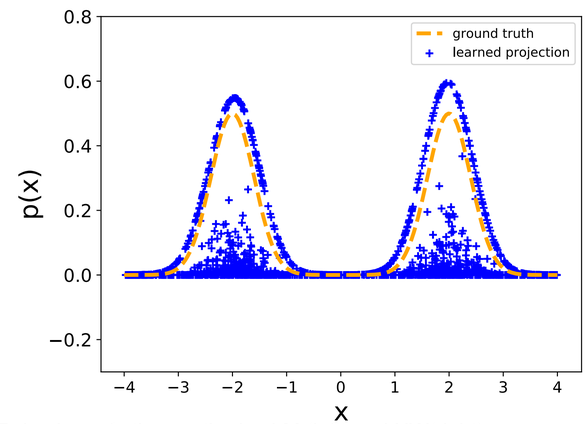
$$A_\star = \sum_{i,j=1}^n \boldsymbol{A}_{ij}\varphi(x_i)\varphi(x_j)^\top$$

$$\boldsymbol{A}_\star \in \arg\min_{A\geq 0} \frac{1}{n}\sum_{i=1}^n \ell_i([KAK]_{ii}) + \Omega(\boldsymbol{A})$$

4) And many more : a finite n-dimensional dual representation, good Rademacher complexity …

Example : estimating the density of a gaussian mixture in dimension 10 with 1000 data points

$$A_\star \in \arg\min_{\substack{A\succeq 0 \\ \int\varphi(x)^\top A\varphi(x)=1}} \frac{1}{n}\sum_{i=1}^n -\log(\varphi(x_i)^\top A\varphi(x_i)) + \mu\|A\|_\star + \frac{\lambda}{2}\|A\|_F^2$$



Behaviour of other methods : LM (left) and NW(right)