Non-parametric Models for Non-negative Functions

Ulysse Marteau-Ferey, Francis Bach, Alessandro Rudi

NeurIPS 2020, December 2020

INRIA Paris – Département d'informatique de l'ENS Paris

Outline

Linear models : a machine learning workhorse with great properties

But what if we want to learn non-negative functions ?

We model positive functions with the same good properties

Outline

Linear models : a machine learning workhorse with great properties

But what if we want to learn non-negative functions ?

We model positive functions with the same good properties

Prototypical machine learning task

Goal : find $f_\star : \mathcal{X} \to \mathbb{R}$ using n training points $(x_i)_{1 \leqslant i \leqslant n}$.

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

Prototypical machine learning task

Goal : find $\theta_\star \in \mathcal{H}$ using n training points $(x_i)_{1 \leqslant i \leqslant n}$.

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

Linear Models

Features : $\varphi(x) \in \mathcal{H}$ (built features, kernels...)
Parametrization : by a vector $\theta \in \mathcal{H}$, $f_\theta : \mathcal{X} \to \mathbb{R}$

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

Linear models have great properties

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

- **They preserve the convexity of the loss function**

If the ℓ_i are **convex**, then (2) is convex

* convex **analysis**
* convex **optimization**

Linear models have great properties

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

- They preserve the convexity of the loss function
- **They can approximate rich classes of functions when $\mathcal{H}$ is infinite dimensional**

Example

φ **feature map** associated to the **gaussian kernel** $k(x, y) = \exp(-\|x - y\|^2)$ on $\mathbb{R}^d$

The linear model can approximate **any continuous function** :

$\boldsymbol{f}_\theta = \theta^\top \varphi(x)$ for $\theta \in \mathcal{H}$
$\boldsymbol{f}$ a continuous function on $\mathbb{R}^d$

$$\exists (\theta_n) \in \mathcal{H}^{\mathbb{N}},\ f_{\theta_n} \underset{n \to +\infty}{\longrightarrow} f \qquad \text{uniformly on compact sets}$$

Linear models have great properties

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

- They preserve the convexity of the loss function
- They can approximate rich classes of functions when $\mathcal{H}$ is infinite dimensional
- **There is a finite dimensional representation with n degrees of freedom**

$$\theta_\star = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$$

Linear models have great properties

$$\alpha_\star = \arg\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \ell_i([\boldsymbol{K}\alpha]_i) + \Omega(\alpha), \;\; \boldsymbol{K} \in \mathbb{R}^{n \times n} \tag{2}$$

- They preserve the convexity of the loss function
- They can approximate rich classes of functions when $\mathcal{H}$ is infinite dimensional
- **There is a finite dimensional representation with n degrees of freedom**

$$\theta_\star = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$$

(2) is now a problem in $(\alpha_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$

Outline

Linear models : a machine learning workhorse with great properties

But what if we want to learn non-negative functions ?

We model positive functions with the same good properties

What if we want to learn non-negative functions ?

What if we want $f \geqslant 0$?

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

Linear models do not work anymore !

Classical models lack nice properties of linear models

Some models to solve :

$$f_\star \in \arg\min_{\substack{f \in \mathcal{F} \\ f \geqslant 0}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(f(x_i)) + \Omega(f) \tag{3}$$

Classical models lack nice properties of linear models

$$\theta_\star \in \underset{\substack{\theta \in \mathcal{H} \\ f_\theta \geqslant 0}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell_i(f_\theta(x_i)) + \Omega(\theta) \tag{3}$$

Generalized linear models : $f_\theta(x) = \exp(\theta^\top \varphi(x))$

Advantages :

- Automatically have $\boldsymbol{f}_\theta \geqslant 0$
- Good approximation properties
- Finite dimensional representer theorem

Classical models lack nice properties of linear models

$$\theta_\star \in \arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\exp(\theta^\top \varphi(x_i))) + \Omega(\theta) \tag{3}$$

Generalized linear models : $f_\theta(x) = \exp(\theta^\top \varphi(x))$

Advantages :

- Automatically have $\boldsymbol{f}_\theta \geqslant 0$
- Good approximation properties
- Finite dimensional representer theorem

Main drawback : (3) is not convex in θ when the ℓ_i are convex

Classical models lack nice properties of linear models

$$\theta_\star \in \arg \min_{\substack{\theta \in \mathcal{H} \\ \theta^\top \varphi(\tilde{x}) \geqslant 0,\ \tilde{x} \in G}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(f_\theta(x_i)) + \Omega(\theta) \quad (3)$$

Linear models $f_\theta(x) = \theta^\top \varphi(x)$ with constraints on a grid

Advantages :

- Preserved convexity
- Good approximation properties
- Finite dimensional representer theorem

Classical models lack nice properties of linear models

$$\theta_\star \in \arg \min_{\substack{\theta \in \mathcal{H} \\ \theta^\top \varphi(\tilde{x}) \geqslant 0,\ \tilde{x} \in G}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(f_\theta(x_i)) + \Omega(\theta) \quad (3)$$

Linear models $f_\theta(x) = \theta^\top \varphi(x)$ with constraints on a grid

Advantages :

- Preserved convexity
- Good approximation properties
- Finite dimensional representer theorem

Main drawback : $f_\theta \not\geq 0$, grid size untractable in high dimensions

$$\theta_\star \in \arg\min_{\substack{\theta\in\mathbb{R}^n \\ \theta \geqslant 0}} \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\theta(x_i)) + \Omega(\theta) \tag{3}$$

Nadaraya Watson type estimators with positive kernel k

$$f_\theta(x) = \sum_{i=1}^{n} \theta_i k(x - x_i),\ k \geqslant 0,\ \theta \in \mathbb{R}^n,\ \theta \geqslant 0.$$

Advantages :

- Preserved convexity
- $f_\theta \geqslant 0$ guaranteed

Classical models lack nice properties of linear models

$$\theta_\star \in \arg\min_{\substack{\theta\in\mathbb{R}^n \\ \theta\geqslant 0}} \frac{1}{n}\sum_{i=1}^{n} \ell_i([\boldsymbol{K}\theta]_i) + \Omega(\theta), \;\; \boldsymbol{K} \in \mathbb{R}^{n\times n} \tag{3}$$

Nadaraya Watson type estimators with positive kernel k

$$f_\theta(x) = \sum_{i=1}^{n} \theta_i k(x - x_i), \;\; k \geqslant 0, \;\; \theta \in \mathbb{R}^n, \;\; \theta \geqslant 0.$$

Advantages :

- Preserved convexity
- $f_\theta \geqslant 0$ guaranteed

Main drawbacks : Poor approximation due to the "width" of k

Classical models lack nice properties of linear models

$$f_\star \in \arg\min_{\substack{f \in \mathcal{F} \\ f \geqslant 0}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(f(x_i)) + \Omega(f) \tag{3}$$

- Generalized linear models do not preserve convexity
- Linear models on a grid do not guarantee non-negativity and are not tractable in high dimensions
- Nadaraya-Watson type kernels have poor approximation and computational properties

Proposed model for non-negative functions

Idea : start from the following GLM :

$$f_\theta(x) = (\theta^\top \varphi(x))^2, \qquad \theta \in \mathcal{H}$$

It has all the good properties... except for **convexity** :

$$\theta_\star \in \arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_i((\theta^\top \varphi(x_i))^2) + \Omega(\theta) \tag{3}$$

Proposed model for non-negative functions

Rewrite it differently :

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

$\theta\theta^\top$ is a positive semi-definite rank 1 operator :

$$\theta_\star \in \arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\varphi(x_i)^\top \theta\theta^\top \varphi(x_i)) + \Omega(\theta) \tag{3}$$

Proposed model for non-negative functions

Change to a linear parametrization :

$$f_A(x) = \varphi(x)^\top A \varphi(x), \qquad A \in \mathcal{S}(\mathcal{H}),\ A \succeq 0,\ \mathrm{rk}(A) \leqslant 1$$

The following problem is now convex in A...

$$A_\star \in \arg \min_{\substack{A \in \mathcal{S}(\mathcal{H}) \\ A \succeq 0 \\ \mathrm{rk}(A) \leqslant 1}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\varphi(x_i)^\top A \varphi(x_i)) + \Omega(A) \tag{3}$$

... except for the $\mathrm{rk}(A) \leqslant 1$ constraint.

Our model for non-negative functions :

$$f_A(x) = \varphi(x)^\top A \varphi(x), \qquad A \in \mathcal{S}(\mathcal{H}),\ A \succeq 0$$

The problem is now convex in A

$$A_\star \in \arg \min_{\substack{A \in \mathcal{S}(\mathcal{H}) \\ A \succeq 0}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\varphi(x_i)^\top A \varphi(x_i)) + \Omega(A) \tag{3}$$

Non-negativity : $\boxed{A \succeq 0 \implies f_A \geqslant 0}$

Outline

Linear models : a machine learning workhorse with great properties

But what if we want to learn non-negative functions ?

We model positive functions with the same good properties

The proposed model keeps the interesting properties

$$A_\star \in \arg \min_{\substack{A \in \mathcal{S}(\mathcal{H}) \\ A \succeq 0}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\varphi(x_i)^\top A \varphi(x_i)) + \Omega(A) \qquad (3)$$

We prove that our model has the good properties of linear models :

- (3) is convex in A if the ℓ_i are convex
- approximation properties **match those of linear models**, **$\mathcal{H}$ infinite dimensional**
- **finite dimensional representation** with $\boldsymbol{n^2}$ **parameters**:
- **dual representation** using only **$\boldsymbol{n}$ parameters**;
- statistical complexity **matches that of linear models**
- ... and more !

Approximation properties

Approximation properties **match those of linear models**
when $\mathcal{H}$ is infinite dimensional :

With a certain feature maps φ our model can approximate
any non-negative continuous function

Finite dimensional representations

- **finite dimensional representation** with $\boldsymbol{n}^2$ parameters:

$$A_\star \in \arg \min_{\substack{A \in \mathcal{S}(\mathcal{H}) \\ A \succeq 0}} \frac{1}{n} \sum_{i=1}^{n} \ell_i(\varphi(x_i)^\top A \varphi(x_i)) + \Omega(A) \tag{3}$$

$A_\star$ can be parametrized by $\boldsymbol{B} \in \mathbb{R}^{n \times n}$:

$$A_\star = \sum_{i,j=1}^{n} \boldsymbol{B}_{ij} \; \varphi(x_i)\varphi(x_j)^\top, \;\; \boldsymbol{B} \in \mathbb{R}^{n \times n}$$

Finite dimensional representations

- **finite dimensional representation** with $\boldsymbol{n}^2$ parameters:

$$\boldsymbol{B}_\star \in \arg \min_{\substack{\boldsymbol{B}\in\mathbb{R}^{n\times n}\\ \boldsymbol{B}\succeq 0}} \frac{1}{n}\sum_{i=1}^{n} \ell_i([\boldsymbol{K}\boldsymbol{B}\boldsymbol{K}]_{ii}) + \Omega(\boldsymbol{B}),\ \ \boldsymbol{K}\in\mathbb{R}^{n\times n} \tag{3}$$

(3) is now a problem in $\boldsymbol{B} \in \mathbb{R}^{n\times n}$

$$A_\star = \sum_{i,j=1}^{n} \boldsymbol{B}_{ij}\ \varphi(x_i)\varphi(x_j)^\top,\ \ \boldsymbol{B}\in\mathbb{R}^{n\times n}$$

Finite dimensional representations

- finite dimensional representation with $\boldsymbol{n}^2$ parameters:
- **dual representation** using only **$\boldsymbol{n}$ parameters**;

$$\alpha_\star \in \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \ell_i^*(n\alpha_i) + \Omega_+^*\left[\sum_{i=1}^n \alpha_i \varphi(x_i)\varphi(x_i)^\top\right] \tag{3dual}$$

$A_\star$ can be recovered from $\alpha_\star$ for certain Ω.

Is this model computable/tractable ?

Yes !

Is this model computable/tractable ?

Example with a density estimation problem :

$$f_\star \in \arg \min_{\substack{f \geqslant 0 \\ \int_{\mathbb{R}^d} f(x)dx = 1}} -\frac{1}{n} \sum_{i=1}^{n} \log f(x_i)$$

Is this model computable/tractable ?

Example with a density estimation problem :

$$A_\star \in \arg \min_{\substack{A \succeq 0 \\ A \cdot \int_{\mathbb{R}^d} \varphi(x)\varphi(x)^\top dx = 1}} -\frac{1}{n}\sum_{i=1}^{n} \log(\varphi(x_i)^\top A \varphi(x_i))$$

Is this model computable/tractable ?

Example with a density estimation problem :

$$A_\star \in \arg \min_{\substack{A \succeq 0 \\ A \cdot \int_{\mathbb{R}^d} \varphi(x)\varphi(x)^\top dx = 1}} -\frac{1}{n}\sum_{i=1}^{n} \log(\varphi(x_i)^\top A \varphi(x_i))$$

Other examples:

- Heteroscedastic regression : **guarantee the variance is non-negative**
- Quantile regression : **guarantee that quantiles do not intersect**

Is this model computable/tractable ?

Toy problem : retrieve density of a mixture of Gaussians