Mathematical Foundation of Machine Learning Spring 2025: Assignment III

1. (30%) Let $\{x_i, f(x_i)\}_{i\in I}$ be a dataset with $x_i \in [0,1]^n$ and $f:[0,1]^n \to \mathbb{R}^L$. Let $\mathcal{F}$ be a forest constructed over this data. For any $m>0$, let $\tilde{x}_i = (x_i, z_i) \in [0,1]^{n+m}$, $z_i \in \mathbb{R}^m$, $i \in I$ and $\tilde{f}:[0,1]^{n+m} \to \mathbb{R}^L$, defined by $\tilde{f}(\tilde{x}) := f(\tilde{x}_1,\ldots,\tilde{x}_n)$. Let $\tilde{\mathcal{F}}$ be the natural extension of $\mathcal{F}$ over $[0,1]^{n+m}$ using the same trees with the same subdivisions over the first n dimensions. Prove that $N_\tau(\tilde{f},\tilde{\mathcal{F}}) = N_\tau(f,\mathcal{F})$, for any $\tau>0$.

Hints/comments

a. The wavelets do change with the increase of the dimension. So, you need to show the invariance of the wavelet norms.
b. This shows some invariance of the sparsity/smoothness indicators under dimension embeddings. Moreover, if the impactful features are only a subset of lower dimension (not necessarily the first n features), then the sparsity/smoothness indicators will only be determined by them.

2. (40%) [Back propagation] Let $f(x) = h_\theta(f_2(f_1(x,w_1),w_2))$, $w := (w_1,w_2) \in \mathbb{R}^2$, $\theta = (\beta,\beta_0) \in \mathbb{R}^2$, be a binary classification model consisting of 2 neural network layers and a logistic regression function

$$h_\theta(t) = \frac{1}{1+e^{-(\beta t+\beta_0)}}.$$

Compute the gradient with respect to the negative log-likelihood loss, for the 4 weights at some point $(w^*,\theta^*) \in \mathbb{R}^4$, using the training dataset $\{x_i, y_i\}_{i\in I}$, $x_i \in \mathbb{R}$, $y_i \in \{0,1\}$.

2. (40%) Prove that for any $\alpha < 1/\tau$, there exists a constant $c(\alpha,\tau)>0$, such that

$$|\Delta_j|_{B^\alpha_\tau([0,1])} \le c 2^{j\alpha},$$

where Δ_j is the sawtooth function with 2^{j-1} teeth.

Hints/comments

a. For simplicity, you can prove the case $0<\alpha<2$, which allows you to use

$$|\Delta_j|_{B^\alpha_\tau([0,1])} = \left(\int_0^\infty \left(t^{-\alpha}\omega_2(\Delta_j,t)_\tau\right)^\tau \frac{dt}{t}\right)^{1/\tau}.$$

For higher values of α, the definition calls for higher orders of the modulus $r \ge \lfloor \alpha \rfloor + 1 > 2$.

b. It is convenient to split the integration in t to: $\int_0^\infty = \int_0^{2^{-(j+1)}} + \int_{2^{-(j+1)}}^\infty$.

c. There is also a lower bound that gives the equivalence we stated in class $|\Delta_j|_{B^\alpha_\tau([0,1])} \sim 2^{j\alpha}$.