Mathematical Foundation of Machine Learning Spring 2024: Final assignment

1. [10%] Assume you have trained a logistic regression model. Explain what its ROC curve is and how exactly is it computed. What is an 'optimal' ROC curve?

2. [10%]. For any $N \geq 1$, find an optimal adaptive N-term piecewise constant approximation to $f(x) = x^2$ in $L_\infty[0,1]$.

3. [10%] We have a classification dataset with n features and L classes.

    a. Explain in detail how to evaluate feature importance using the wavelet method.
    b. Explain how to use an evaluation dataset (not used during training), to ensure the feature importance method is less susceptible to noise.

4. [15%] Prove that for any $\alpha < 1/\tau$, there exists a constant $c(\alpha, \tau) > 0$, such that

$$\left|\Delta_j\right|_{B_\tau^\alpha([0,1])} \leq c 2^{j\alpha},$$

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

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

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

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 $\left|\Delta_j\right|_{B_\tau^\alpha([0,1])} \sim 2^{j\alpha}$.

5. [15%] 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.

6. [10%] Show that for any $0<\varepsilon<1$, there exists a neural network $\tilde{D}$ with $O\left(\log^2\left(\varepsilon^{-1}\right)\right)$ weights that approximates the function $D(x_1,x_2)=e^{x_1}x_2$, $\left\|D-\tilde{D}\right\|_{L_\infty\left([-1,1]^2\right)}\le\varepsilon$.

Hint/comment

You may use the following: For any $n\ge 1$, there exist neural networks $D_{1,n}, D_{2,n}$, each with $O(n)$ weights, such that $\max_{-1\le x\le 1}\left|e^x-D_{1,n}(x)\right|\le c_1e^{-n}$ and $\max_{-2e\le x_1,x_2\le 2e}\left|x_1x_2-D_{2,n}(x_1,x_2)\right|\le c_2e^{-\sqrt{n}}$.

7. [10%] Design a NN that will solve the following computer vision problem: You are given a blurry real-life RGB image, and you need to unblur/sharpen it.
   a. Describe how to create a training dataset for this problem.
   b. Describe a candidate NN architecture and explain its logic.
   c. Describe what loss function(s) can be used during training.

Hint/comment

You may assume the blurred images were generated through a Gaussian blur, that is, each color channel of the blurry image is the convolution of a `sharp' real-life image with a fixed 2D Gaussian.

8. [10%] Design a NN $\tilde{u}$ and a loss function for the following problem. You need to solve the 2D heat equation

$$\frac{\partial u}{\partial t}(x,t)=\frac{\partial^2 u}{\partial^2 x_1}(x,t)+\frac{\partial^2 u}{\partial^2 x_2}(x,t),$$

in the domain $x=(x_1,x_2)\in[0,1]^2$, $t\in[0,10]$, with initial conditions $u(x,0)=f(x)$, and (compatible) boundary conditions

$$u((0,x_2),t)=g_1(t),\ u((1,x_2),t)=g_2(t),\ u((x_1,0),t)=g_3(t),\ u((x_1,1),t)=g_4(t),\ t\in[0,10]$$

The input to the network is $x=(x_1,x_2)\in[0,1]^2$, $t\in[0,10]$, and the output is a scalar $\tilde{u}(x,t)$. Explain the role of automatic differentiation during training.

9. [10%] You are given the problem of finding the point source location at time 0, of the 3D wave equation in the 3D domain $[0,1]^3$, given only samples of the solution at fixed time T

$$\left\{u(x_i,T):x_i=\left(\frac{i_1}{N},\frac{i_2}{N},\frac{i_3}{N}\right),\ 0\le i_1,i_2,i_3\le N\right\}.$$

   a. Describe how you would create synthetic dataset for this problem (to be split later to training and testing datasets).
   b. Provide a sketch design for the network architecture and write the loss function.