Mathematical Foundation of Machine Learning Spring 2022: Assignment II

1. (40%) [Besov smoothness of indicator functions] Let $f(x) = 1_{\tilde{\Omega}}(x)$, where $\tilde{\Omega} \subset [0,1]^2$ is a domain with a smooth boundary.
   a. Prove that for an isotropic tree $\mathcal{T}_I$, that applies non-adaptive subdivisions at dyadic partitions of the two variables, we get that $f \in B_\tau^\alpha(\mathcal{T}_I)$, for $\alpha < 1/(2\tau)$.
   b. Assume that one can construct an anisotropic tree $\mathcal{T}_A$, such that from some level $k_0 \geq 1$, there are only $\leq c_1 2^m$ domains at the level $k_0 + 2m$ that intersect the boundary of $\tilde{\Omega}$, each of area $\leq c_2 2^{-3m}$. Prove that $f \in B_\tau^\alpha(\mathcal{T}_A)$, where $\alpha < 2/(3\tau)$. You can assume that there exists $0 < \rho < 1$, such that for any child Ω' of Ω, $|\Omega'| \leq \rho|\Omega|$.

Hints:

(i) For both cases you need to follow the proof of the case of the classic Besov smoothness with dyadic cubes.

(ii) The case (a) is very similar to the classic case, because $B_\tau^{2\alpha} \sim B_\tau^\alpha(\mathcal{T}_I)$.

(iii) In both cases, you first need to bound the contribution of the odd levels of the tree by the even levels (use the ρ volume condition for (b) and properties of the modulus).

2. (25%) [Convolutions] The convolution of $f, g \in L_1(\mathbb{R}^n)$ is defined by $f * g(x) := \int_{\mathbb{R}^n} f(x-y)g(y)dy$.

   (i) Prove that $f * g \in L_1(\mathbb{R}^n)$,

   (ii) Prove that $f * g = g * f$,

   (iii) The Fourier Transform is defined by $\hat{f}(w) = \int_{\mathbb{R}^n} f(x)e^{-i\langle w,x\rangle}dx$, for $w \in \mathbb{R}^n$. Show that

$$(f * g)^\wedge(w) = \hat{f}(w)\hat{g}(w), \quad \forall w \in \mathbb{R}^n.$$

   (iv) Let $f \in L_1(\mathbb{R}^2)$ be a piecewise constant function. Design a 'filter' $g \in L_\infty(\mathbb{R}^2)$, with support in $[-\varepsilon/2, \varepsilon/2]^2$, for some $\varepsilon > 0$, such that $f * g$ is 'significant' only in ε neighborhoods of points where f has 'almost' vertical edges.

3. (25%) [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 $\{x_i, y_i\}_{i \in I}$, $x_i \in \mathbb{R}$, $y_i \in \{0,1\}$.

4. (10%) The random forest model is a stronger learner than the linear regression model. Can one build a regression model composed of a neural network with a random forest at the last layer instead of a linear regression model?