

**STATISTICAL MACHINE LEARNING (WS2024)  
SEMINAR ON ENSEMBLING**

**Assignment 1.** Consider regression with training datasets  $\mathcal{T}^m$  of size  $m$  generated as:

$$y = f(x) + \epsilon, \quad (1)$$

where  $\epsilon$  is the noise having  $\mathbb{E}[\epsilon] = 0$  and  $\text{Var}(\epsilon) = \sigma^2$ . Derive bias-variance decomposition for  $k$ -nearest-neighbor regression. The response of the  $k$ -NN regressor is defined as:

$$h_m(x) = \frac{1}{k} \sum_{i=1}^k y_{n(x,i)} = \frac{1}{k} \sum_{i=1}^k f(x_{n(x,i)}) + \epsilon, \quad (2)$$

where  $n(x, i)$  gives the index of  $i$ -th nearest neighbor of  $x$  in  $\mathcal{T}^m$ . For simplicity assume that all  $x_i$  are the same for all training datasets  $\mathcal{T}^m$  in consideration, hence, the randomness arises from the noise  $\epsilon$ , only.

Give bias<sup>2</sup>:

$$\mathbb{E}_x \left[ \left( g_m(x) - f(x) \right)^2 \right] = \mathbb{E}_x \left[ \left( \mathbb{E}_{\mathcal{T}^m} [h_m(x)] - f(x) \right)^2 \right] \quad (3)$$

and variance:

$$\text{Var}_{x, \mathcal{T}^m} (h_m(x)). \quad (4)$$

**Assignment 2.** The output of a regression tree is defined as:

$$h(\mathbf{x}) = \sum_{r=1}^M c_r \mathbb{I}\{\mathbf{x} \in R_r\} \quad (5)$$

where  $R_r$  is an input space region defined by the  $r$ -th tree leaf and  $c_r \in \mathbb{R}$  the corresponding region's response. The tree is trained using set  $\mathcal{T}^m = \{(\mathbf{x}_i, y_i) \mid i = 1, \dots, m\}$ . Show that the sum of squares loss function  $\sum_{i=1}^m (y_i - h(\mathbf{x}_i))^2$  is minimized by choosing the following region responses:

$$c_r = \frac{1}{|S_r|} \sum_{\mathbf{x}_i \in R_r} y_i \quad (6)$$

where  $S_r = \{(\mathbf{x}_i, y_i) : (\mathbf{x}_i, y_i) \in \mathcal{T}^m \wedge \mathbf{x}_i \in R_r\}$ .

**Assignment 3.** What is an optimal value of  $c_r$  when the sum of absolute deviations  $\sum_{i=1}^m |y_i - h(\mathbf{x}_i)|$  is used instead of the squared loss?

**Assignment 4.** Bootstrapping is a method which produces  $K$  datasets  $\mathcal{T}_i^m$  for  $i = 1, \dots, K$  by uniformly sampling the original dataset  $\mathcal{T}^m$  with replacement. Bootstrap datasets have typically the same size as the original dataset  $|\mathcal{T}_i^m| = |\mathcal{T}^m| = m$ . Show that as  $m \rightarrow \infty$  the fraction of unique samples in  $\mathcal{T}_i^m$  approaches  $1 - \frac{1}{e} \approx 63.2\%$ .

Hint: apply exponential of a logarithm to a limit which emerges in a last step in order to solve it.

**Assignment 5.** Consider the Huber loss:

$$\ell(y, h(x)) = \begin{cases} (y - h(x))^2 & \text{for } |y - h(x)| \leq \delta \\ 2\delta|y - h(x)| - \delta^2 & \text{otherwise.} \end{cases} \quad (7)$$

Define Gradient Boosting Machine using the Huber loss and discuss differences to the squared loss GBM.