STATISTICAL MACHINE LEARNING (WS2021)
SEMINAR ON ENSEMBLING

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

$$y = f(x) + \epsilon, \tag{1}$$

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

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

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 ϵ, only.

Give bias^2:

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

and variance:

$$\mathrm{Var}_{x,\mathcal{T}^m}\Big(h_m(x)\Big). \tag{4}$$

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

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

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 $\mathcal{T}^m = \{(\boldsymbol{x}_i, y_i) \mid i = 1, \ldots, m\}$. Show that the sum of squares loss function $\sum_{i=1}^{m}(y_i - h(\boldsymbol{x}_i))^2$ is minimized by choosing the following region responses:

$$c_r = \frac{1}{|S_r|}\sum_{\boldsymbol{x}_i \in R_r} y_i \tag{6}$$

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

Assignment 3. What is an optimal value of c_r when the sum of absolute deviations $\sum_{i=1}^{m} |y_i - h(\boldsymbol{x}_i)|$ 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 $m \to \infty$ 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:

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

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