SML 2024, Monsoon, Quiz 2, Dur. 1 hr 10 mins.

Q1. Explain the difference between bagging and random forest. [0.5]

Q2. Explain OOB error. [1]

Q3. Suppose the regression model is $\hat{y} = w_0 \ln(w_1 e^x)$. Find w_0 and w_1 for the given data $D = \{(x, y)\} = \{(0, 1); (1, 3)\}$. [1]

Q4. Find the regression decision stump for data $D = \{(x_1, x_2, y)\} = \{(0, 0, 1); (1, 1, 3); (2, 1, 3), (0, 2, 2)\}$. Consider splitting at $(0, 1.5)$ and $(1.5, 0)$. What are the predictions of the terminal node of the decision stump? [1.5]

Q5. In Adaboost.M1, each tree and its confidence is obtained by minimizing a weighted mis-classification error. Suppose instead of this weighted mis-classification error we choose total weighted Gini-index (WGI) to be minimized. For Gini index (GI), we need to compute probability of each class. Given a node, this can be done as the ratio of number of samples of a given class to the total number of samples. For node m, this can be written as $p_{mk} = \sum_{i=1}^{N_m} \mathcal{I}(x_i = k)/N_m$, where $\mathcal{I}(x_i = k)$ is an indicator function with value 1 when the sample x_i belongs to class k, and N_m is the number of samples in node m. Then GI for node m can be written as $\sum_{k=1}^{K} p_{mk}(1 - p_{mk})$. Extending the notion to WGI, WGI is defined as $\sum_{k=1}^{K} p'_{mk}(1 - p'_{mk})$, where $p'_{mk} = \frac{\sum_{i=1}^{N_m} w_i \mathcal{I}(x_i = k)}{\sum_{i=1}^{N_m} w_i}$. Note that you need to compute total WGI for a given cut and seek the cut that minimizes total WGI. Total WGI will be computed in a manner similar to that of total GI.

a. Using the above definition of WGI, find a boosted tree for $D = \{(x, y)\} = \{(1, 1), (2.5, 1), (3.5, 1), (5, -1), (6.5, -1)\}$. Consider decision stumps with cuts at 2 and 3. You need to perform two iterations, that is find α_1, $h_1(x)$, α_2, $h_2(x)$. [3]

b. Find the prediction of sample $x = 4$. [.5]