Pattern Recognition: Homework 4

Due date: 2023.3.23

Problem 1 (40 pt)

We have learned that the Bagging method can reduce the error of the model, and now we prove this theoretically. Taking regression problems as an example, first we use bootstrap method to sample and generate M sub-datasets on the original dataset, and train prediction models $y_m, m = 1, \ldots, M$ on these M sub-datasets respectively. For any sample x, the prediction result of Bagging method is defined as the average of multiple prediction model outputs:

$$y_{COM}(x) = \frac{1}{M} \sum_{m=1}^{M} y_m(x) \tag{1}$$

Let d(x) be the true regression value of sample x, then the error of each prediction model is:

$$\epsilon_m(x) = y_m(x) - d(x) \tag{2}$$

For M separate prediction models, their average mean square error can be expressed as:

$$E_{AV} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)]^2 \tag{3}$$

where $\mathbb{E}_x$ represents taking expectation on the distribution of sample x, $0 \leq y_m(x) \leq 1$. Similarly, the expected error of Bagging method is:

$$E_{COM} = \mathbb{E}_x \left[\frac{1}{M} \sum_{m=1}^{M} \epsilon_m(x) \right]^2 \tag{4}$$

(1) Assuming that all prediction model errors are zero-mean and uncorrelated, i.e.,

$$\mathbb{E}_x[\epsilon_m(x)] = 0, \quad \mathbb{E}_x[\epsilon_m(x)\epsilon_l(x)] = 0, \ m \neq l \tag{5}$$

Try to prove:

$$E_{COM} = \frac{1}{M} E_{AV} \tag{6}$$

(2) In practice, the errors of various prediction models are often highly correlated. Please prove that in case (1)'s assumption does not hold, the following formula still holds:

$$E_{COM} \leq E_{AV} \tag{7}$$

At this point we have theoretically proved the effectiveness of Bagging method in reducing model errors.
Hint:

Jensen's inequality: for any convex function $f(x)$, we have $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$

Problem 2 (60 pt)

In the folder `data`, there is a CSV file containing 500 entries. Each entry contains a person's information, including their gender, height, weight, and obesity index. Persons with an obesity index of ≥ 4 are considered obese. We aim to build a decision tree from scratch that correctly classifies each person as obese or not based on their gender, height, and weight. The code and instructions are in the file `decision_tree.ipynb`. Please read the code and complete the following tasks:

(1) Complete code marked as # TODO in the following 3 functions (each has one): `gini_impurity`, `information_gain` and `max_information_gain_split`.

(2) Read the code of the function `train_tree` carefully, then answer what each input variable and each return value does in function `make_split` and function `get_best_split`.

For example, for function `make_prediction`, a proper answer would be:

Input: `data`, which is a pandas series variable. Each item of `data` represents whether the person is obese (1) or not (0).

Return: `pred`, which is the prediction based on the most frequent value of `data`.

(3) Setting the value of `max_depth` to 1, 2, 4, 8, 16 when training the decision tree and plot the validation accuracy *vs.* `max_depth` curve. What is your observation? Does overfitting occurs? If not, try explaining the reason.