Problem Set 5

Oct. 28, 2021

1. Bagging is another ensemble algorithm designed to improve the accuracy of classification/regression by voting. We now prove its property theoretically. Considering a regression problem in which we perform sampling with replacement on dataset D to generate M datasets $\{D_m\}_{m=1}^{\mathrm{M}}$. We train a predictive model $y_m(x)$ on each resampled dataset D_m $m = 1, \ldots, \mathrm{M}$. For any sample $\boldsymbol{x}$, the prediction result of Bagging is given by
$$y_{\mathrm{BAG}}(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M} y_m(\boldsymbol{x}).$$
Suppose the true regression function is $d(\boldsymbol{x})$. The error of each prediction model is given by
$$\epsilon_{\mathrm{m}}(\boldsymbol{x}) = y_m(\boldsymbol{x}) - d(\boldsymbol{x})$$
For M predictive models, their average sum-of-squares error then takes the form
$$E_{\mathrm{AV}} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\boldsymbol{x}}\{[\epsilon_{\mathrm{m}}(\boldsymbol{x})]^2\}$$
where $\mathbb{E}_{\boldsymbol{x}}\{\}$ denotes the expectation with respect to the distribution of the input vector $\boldsymbol{x}$.
Similarly, the expected error of Bagging algorithm is given by
$$E_{\mathrm{BAG}} = \mathbb{E}_{\boldsymbol{x}}\left\{\left[\frac{1}{M}\sum_{m=1}^{M} y_m(\boldsymbol{x}) - d(\boldsymbol{x})\right]^2\right\} = \mathbb{E}_{\boldsymbol{x}}\left\{\left[\frac{1}{M}\sum_{m=1}^{M} \epsilon_{\mathrm{m}}(\boldsymbol{x})\right]^2\right\}.$$
(1) If we assume that the errors have zero mean and are uncorrelated, i.e.,
$$\mathbb{E}_{\boldsymbol{x}}[\epsilon_{\mathrm{m}}(\boldsymbol{x})] = 0$$
$$\mathbb{E}_{\boldsymbol{x}}[\epsilon_{\mathrm{m}}(\boldsymbol{x})\epsilon_l(\boldsymbol{x})] = 0 \qquad \mathrm{m} \neq l$$
Prove that:
$$E_{\mathrm{BAG}} = \frac{1}{\mathrm{M}} E_{\mathrm{AV}}$$
(2) In practice, the errors are often correlated. Prove that in the case that the assumption of (1) is not true, the following still holds:
$$E_{\mathrm{BAG}} \leq E_{\mathrm{AV}} \ .$$

2. Deciding whether orange is sweet based on its appearance is a challenge. Here is a table of several features of 17 oranges and whether they're sweet or not. Please draw a decision tree to determine orange sweetness for this table using information gain.

No.	Weight	Size	Touch	Texture	Sweet
1	heavy	big	hard	smooth	Yes
2	light	big	hard	smooth	Yes
3	light	big	hard	smooth	Yes
4	heavy	big	hard	smooth	Yes
5	average	big	hard	smooth	Yes
6	heavy	medium	hard	fine	Yes
7	light	medium	normal	fine	Yes
8	light	medium	hard	smooth	Yes
9	light	medium	normal	smooth	No
10	heavy	small	hard	fine	No
11	average	small	spongy	smooth	No
12	average	big	spongy	fine	No
13	heavy	medium	normal	smooth	No
14	average	medium	normal	smooth	No
15	light	medium	hard	fine	No
16	average	big	spongy	smooth	No
17	heavy	big	normal	smooth	No

Due date: Nov. 3 (Wednesday) 23:00 Beijing time