Model Zoo 2

Decision Trees

Describe the decision tree as implemented in `sklearn.tree.DecisionTreeClassifier`.

Theorem 1 (universal approximation)**.** Let g be the ERM hypothesis for the class of binary decision trees with k nodes. Then,

$$\lim_{k\to\infty} E_{\text{in}}(g) = 0. \tag{1}$$

Theorem 2. Let k be the number of nodes in a binary decision tree. Then the VC-dimension is bounded by

$$d_{\text{VC}} = O(k \log(kd)). \tag{2}$$

Proof. See "Decision trees as partitioning machines to characterize their generalization properties," NeurIPS 2020. □

Corollary 1. Let j be the height of a binary decision tree. Then the VC dimension is bounded by

$$d_{\text{VC}} = O(2^j j \log d). \tag{3}$$

Proof. The number of nodes $k = O(2^j)$. Substituting into Equation (2) gives Equation (3). □

Note 1. These results directly contradict the advice given in scikit-learn's "Tips for Practical Use": `https://scikit-learn.org/stable/modules/tree.html#tips-on-practical-use`.

Problem 1. Describe how changes to the following hyperparameters to `sklearn.tree.DecisionTreeClassifier` affect the VC dimension (increase, decrease, stays the same).

1. `criterion`

2. `max_depth`

3. `max_features`

4. `max_leaf_nodes`

5. `min_samples_leaf`

6. `min_samples_split`

7. `random_state`

Problem 2. If you double the height of a decision tree from 3 to 6, approximately how much more data do you need to achieve the same generalization error?

Problem 3. What is the VC dimension of a decision tree with k nodes where the polynomial feature map of degree p has been applied to the data?

Ensemble Methods

The hypothesis class of ensemble methods is

$$L(B, T) = \left\{ \mathbf{x} \mapsto \operatorname{sign}\left(\sum_{t=1}^{T} w_t h_t(\mathbf{x}) \right) : \mathbf{w} \in \mathbb{R}^T, h_t \in B \right\} \tag{4}$$

where B is a set of "base" hypothesis classes and $T \in \mathbb{Z}$ is the number of hypotheses from B to combine.

Theorem 3 (universal approximation)**.** Let g be the ERM hypothesis for $L(B,T)$. Then

$$\lim_{T\to\infty} E_{\text{in}}(g) = 0 \tag{5}$$

for any hypothesis class B that is a weak learner. A *weak learner* is any hypothesis class capable of achieving better than random error for any dataset. (All infinite hypothesis classes we've seen are examples of weak learners.)

Lemma 1. The VC-dimension of $L(B,T)$ is

$$d_{\text{VC}}(L(B,T)) = O(T d_{\text{VC}}(B) \log(T d_{\text{VC}}(B))) \tag{6}$$

Proof. See Chapter 10, Lemma 10.3 of *Understanding Machine Learning: From Theory to Algorithms.* □

Fact 1. There are two main categories of ensemble algorithms:

1. boosting (e.g. `AdaBoostClassifier`, `GradientBoostingClassifier`, `XGBoost`, `LightGBM`), and

2. bagging (e.g. `BaggingClassifier`, `RandomForestClassifier`).

Problem 4. Decision trees are some of the most commonly "boosted" models.

1. Provide a tight upper bound on the VC dimension for an ensemble of decision trees.

2. If you increase the number of decision trees T in the ensemble, then how should you adjust the number of nodes k in the decision trees?

3. If you increase the number of nodes k in the base decision trees, then how should you adjust the number of decision trees in the ensemble T?