Метод построения случайного леса на основе отдаления друг от друга базовых моделей

Дмитриев Леонид Алексеевич
Сенько Олег Валентинович

Московский государственный университет им. М. В. Ломоносова

15 февраля 2023 г.

Введение

Случайный лес - широко известный алгоритм машинного обучения. Его способность хорошо аппроксимировать данные за счёт уменьшения разброса основана на предположении, что все деревья ансамбля независимы и различны.

Проблема: практическое уменьшение разброса значительно меньше теоретического, так как на деле получаемые деревья обучаются на объектах из одного и того же множества.

Цель работы: предложить другой способ повышения разнообразия деревьев внутри леса и показать его применимость.

Идея: вместо независимой и параллельной генерации деревьев, будем на каждом шагу добавлять дерево, сильно отличающееся от уже созданного ансамбля, с помощью специального функционала, учитывающего ответы предыдущих моделей.

Постановка задачи

Имеется выборка $\tilde{S} = \{(y_j, x_j, G_1(x_j), G_2(x_j)), j = \overline{1, m}\}$, где

- y_j - значение переменной Y для объекта с номером j
- $x_j = (x_{j1}, \ldots, x_{jn})$ - вектор значений признаков $X_1, \ldots, X_n$ для объекта с номером j
- $G_1(x_j)$ - значение функции G_1 в точке x_j
- $G_2(x_j)$ - значение функции G_2 в точке x_j

Предлагается построить дерево $T(x)$, для которого достигается минимум функционала:

$$\Phi(\tilde{S}, T) = \Sigma_{j=1}^{m} \{\gamma_1 [T(x_j) - y_j]^2 + \gamma_2 [T(x_j) - G_2(x_j)]^2 - \mu [T(x_j) - G_1(x_j)]^2\}$$

где $\gamma_1 + \gamma_2 = 1; \quad \gamma_1, \gamma_2, \mu \in [0, 1]$

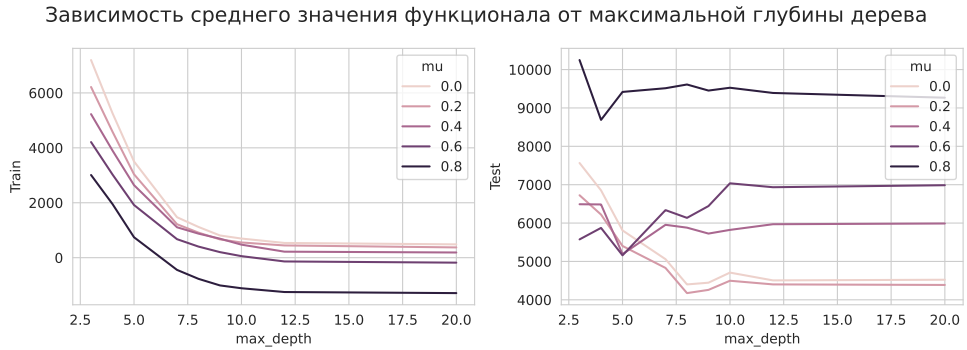
Вместо независимой и параллельной генерации деревьев, будем на каждом шагу добавлять дерево, сильно отличающееся от уже созданного ансамбля, с помощью специального функционала, учитывающего ответы предыдущих моделей.

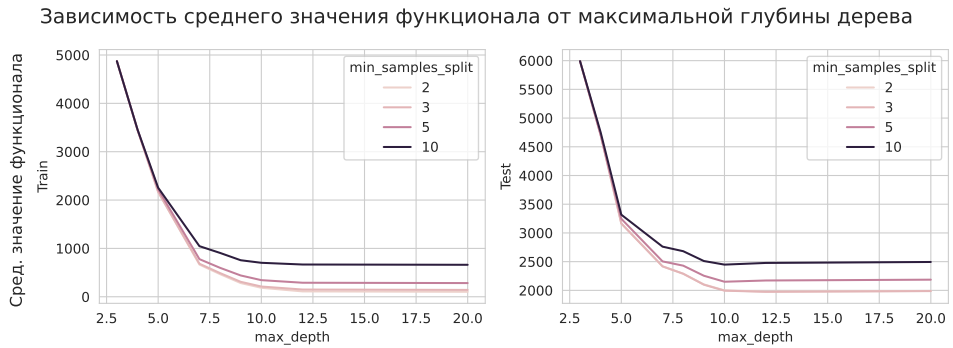
Реализация

При построении дерева был использован "жадный" метод оптимизации целевого функционала: на каждом шагу к дереву добавляется узел, обеспечивающий наибольшее снижение используемого функционала Ф.
Предположим, что на каком-то шаге дерево T_k содержит k концевых узлов, которым соответствуют концевые выборки $S_1^k, \ldots, S_k^k$.
Новое дерево T_{k+1} строится через добавление к дереву T_k дополнительного узла u.
Узел u получается из некоторого концевого узла g с помощью порогового правила вида $X_u > \delta_u$, где X_u и δ_u признак и порог к нему соответственно.
Правило $X_u > \delta_u$ расщепляет выборку S_g^k на две подвыборки.
Признак X_u и порог δ_u ищутся из условия максимизации разности $\Phi(\tilde{S}, T_k) - \Phi(\tilde{S}, T_{k+1})$.
Процесс построения может быть прекращен при выполнении одного из перечисленных условий:

- На очередном шаге не удается уменьшить функционал
- На очередном шаге произошло изменение функционала, меньшее чем некоторое пороговое значение
- Кол-во объектов внутри узла меньше некоторого порогового значения

Слайд об исследованиях





Были проведены эксперименты примения метода для задачи регрессии. Для тестирования и исследования в качестве модели G1 (к зависимости которой дерево должно приближаться) был взят градиентный бустинг, а в качестве модели G2 (от зависимости которой необходимо удаляться) случайный лес. Параметры данных моделей были подобраны на кросс-валидации, где использовалась стандартная метрика MSE. Был произведен ряд экспериментов с различными значениями гиперпараметров модели.

Выводы

В данной работе исследовался метод построения дерева с помощью специального функционала, позволяющего отдаляться от уже построенного ансамбля, с целью повышения разнообразия итогового леса. Исследование производилось на данных о химических соединениях, где целевой переменной была выбрана температура плавления. Были проведены эксперименты, показавшие корректность реализованной модели, выявившие некоторые отличия от стандартного функционала и показавшие применимость метода.