Метод повышения эффективности обучения градиентного бустинга, основанный на модифицированных функциях потерь

Королев Николай Сергеевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: korolev.nikolay.s@gmail.com

Научный руководитель — *Сенько Олег Валентинович*

Решается задача регрессии и классификации, а именно восстановления функции $f(x)$, по множеству точек x_i, y_i, таких что $y_i = f(x_i)$.

Градиентный бустинг [2] основан на итеративном построении функции $f(x)$ за счёт использования большого количества деревьев решений, каждое из которых исправляет ошибки предыдущих. При этом для решения задачи задаётся оптимизируемый функционал ошибки. Одним из стандартных оптимизируемых функционалов является среднеквадратичная ошибка $L(f(x), X, Y) = \frac{1}{T}\sum_{i=1}^{T}(f(x_i) - y_i)^2$.

В работе предлагается использовать модифицированные функционалы ошибки.

Будем использовать в качестве $L(f(x), X, Y) = \frac{1}{T}\sum_{i=1}^{T}(\alpha f(x_i) - y_i)^2$, где $\alpha \in \mathbb{R}_{++}$. Основная идея данного подхода заключается в том, чтобы добавить шума в обучаемые деревья решений, для того чтобы уменьшить корреляцию между выходами различных деревьев решений в итоговом лесе, что позволяет увеличить обобщающую способность обучаемой модели [1].

Как отмечалось ранее, для повышения обобщающей способности леса деревьев решений необходимо уменьшать корреляцию между деревьями решений, поэтому разумно на каждом шаге оптимизировать функцию (1).

$$\frac{1}{T}\sum_{i=1}^{T}\left[(h(x_i) + f(x_i) - y_i)^2 - \gamma(h(x_i) - f(x_i))^2\right] \tag{1}$$

В данной функции есть дополнительная добавка $-\gamma(h(x_i) - f(x_i))^2$, позволяющая добиться дополнительной регуляризации за счёт различия между новым обучаемым деревом и уже обученным лесом решающих деревьев.

Отметим, что в такой ситуации оказывается, что использование данной функции потерь абсолютно эквивалентно использованию смещённой среднеквадратичной ошибки с параметром $\alpha = 1+\gamma$ (при дополнительном шкалировании learning-rate'а градиентного бустинга). Кроме того, можно доказать, что $0 \leq \gamma \leq 1$. В соответствии с границами изменения γ, а также выведенной зависимостью $\alpha = 1+\gamma$ получаем, что имеет смысл рассматривать лишь $\alpha \in [1; 2]$.

Для проверки качества работы представленного метода будем решать различные задачи классификации и регрессии используя обычный градиентный бустинг, сравнивая результаты работы с градиентным бустингом с использованием смещённой квадратичной ошибки. Кроме того, проводились вычислительные эксперименты с использованием среднеквадратичной ошибки с удалением от обученного ансамбля, но их результаты полностью совпадают с использованием обычной смещённой квадратичной ошибки, что соответствует теории.

Наиболее хороших результатов по увеличению предсказательной способности на различных задачах удалось достичь для $\alpha = 1.1$.

Итоговые результаты экспериментов представлены в таблице 1.

Набор данных	Ср-кв. ошибка	Смещ. ср-кв. ошибка	α
Аритмия	0.89	**0.90**	1.7
Ледники	0.72	**0.75**	1.1
Продажи	0.21	**0.26**	1.1
Сист. давл.	0.41	**0.46**	1.1

Таблица 1: Целевая метрика на тестовой выборке для лесов, обученных стандартной процедурой градиентного бустинга (столбец «Ср-кв. ошибка») и с использованием смещённой среднеквадратичной ошибки (столбец «Смещ. ср-кв. ошибка»)

Литература

1. Докукин А. А., Сенько О. В. Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности // Ж. вычисл. матем. и матем. физ., М.: Математический институт им. В. А. Стеклова РАН, 2011. — С. 1751–1760.
2. Jerome H. Stochastic Gradient Boosting // Computational Statistics & Data Analysis, 2002. — С. 367–378.