Градиентный бустинг

Градиентный бустинг - один из способов композиции базовых моделей.

Каждый из базовых алгоритмов независит от других, пытается уменьшить ошибку ансамбля, состоящего из предыдущих заюзанных базовых алгоритмов.

Gradient Boosting on Decision Trees (GBDT) - бустинг на деревьях, хорошо работает на неоднородных данных.

Можно делать градиентный бустинг и над другими видами базовых моделей, на рассмотрим только бустинги на деревьях.

Интуиция

Пусть, мы решаем задачу регрессии. Если применим одно дерево, то предсказательная способность очевидна будет низкой (пусть предсказания нашего дерева отклонены от целевых на каку-то константу), тогда второе запущенное дерево, зная о предсказаниях предшественника постарается эту разницу исправить и т. д.

Более формальное объснение фактически повторяет 1 в 1 интуицию.

Почему бустинг градиентный? Потому что каждая следующая базовая модель предсказывает антиградиент функции потерь (математически обосновывается).

А если мы будем настраивать следующие на очереди базовые алгоритмы на минимизацию не функции расстояния между объектами, а делание шага в направление антиградиента функции потерь, то такие шаги будут более выгодными.

В итоге обучение базового алгоритма проходит в два шага:

— по **функции потерь** вычисляется целевая переменная для обучения следующего базового алгоритма:

$$g_i^k = \frac{\partial \mathcal{L}(y_i, z)}{\partial z}\Big|_{z=a_k(x_i)}$$

— строится регрессионное дерево на обучающей выборке $(x_i, -g_i^k)$, минимизирующее выбранную **оценочную функцию**.

На практике построение такого ансамбля дольше линейных моделей, но меньше чем нейронки.

Learning rate

Чтобы избежать переобучения можно уменьшить глубину деревьев и можно уменьшить learning rate. Learning rate - коэффициент вклада последующего дерева, обычно темп обучения выбирается пользователем, но в некоторых библиотеках он хорошо подбирается автоматически.

Feature importance

Из-за большого количества деревьев возникает вопрос в интерпретации, так как графически уже ничего не показать. Заметим, что признаки, которые находятся близко к корню дерева влияют на предсказание сильнее, чем те, которые ближе к листам. Тогда если мы возьмём средние значения встречаемости признаков у корня деревьев в ансамбле, мы можем сказать какие более важны, а какие менеее - MDI (mean decrease in impurity).

Реализации

В sklearn градиентный бустинг реализован так себе, поэтому используем LightGBM/XGBoost/CatBoost, они +- одинаковы.

На практике

На практике фактически применяются два метода - градиентный бустинг и нейронки. Градиентный бустинг относительно прост по сравнению с нейронками и на неоднородных данных показывает хороший результат, хотя нейронки сейчас потихоньку вытесняют бустинг.