XGBoost   (Tianqi Chen)

XGBoost сокращение “Extreme Gradient Boosting”

XGBoost развивает идею “Gradient Boosting” предложенную в статье Friedman'a

Greedy Function Approximation: A Gradient Boosting Machine.

Tianqi Chen, Carlos Guestrin

XGBoost: A Scalable Tree Boosting System

R. Bekkerman. "The present and the future of the kdd cup competition: an outsider’s perspective."

Kaggle 2015

17 из 29 победителей (1-3 места) использовали XGBoost.

8 из этих 17 использовали только XGBoost,

остальные — в комбинации с нейронными сетями.

2-й по популярности метод - deep neural nets использовался 11-ю победителями.

KDDCup 2015

XGBoost использовался всеми 10-ю победителями.

Кирилл Антонов: LightGBM быстрее и лучше XGBoost

Examples of the problems in these winning solutions include:

store sales prediction;

high energy physics event classification;

web text classification;

customer behavior prediction;

motion detection;

ad click through rate prediction;

malware classification;

product categorization;

hazard risk prediction;

massive online course dropout rate prediction.

Таблица данных размер n x (m+1): n наблюдений, m+1 переменная

$$D=\{(\boldsymbol{x_i}, y_i)\},(|D|=n, \boldsymbol{x_i} \in R^m, y_i \in R)$$

<u>Шаг 1 Обозначения</u>

F - Множество (пространство ?) регрессионных деревьев, построенных по методу CART

$$F=\{f(x)=w_{q(x)}\} \quad (q: R^m \to \{1:T\}, \quad w \in R^T)$$

При этом f_k -дерево, построенное по методу CART., его описание разделяем на части: структура дерева *q* и *w.*

T — число конечных узлов дерева.

Описание дерева разделяем на части: структура дерева *q* и значение, приписанное заданному конечному узлу w.

Структура дерева *q* – правило, которое определяет номер конечного узла, в который попадает объект *x*.

w_i - вес узла, значение, которое приписывается наблюдению, попавшему в конечный узел *I*, *i*=1,2, …, T.

Вектор весов конечных узлов $w=(w_1, w_2, \ldots, w_T)$ - *вектор* значений, которые приписываются наблюдению, попавшему в узел 1, 2, ...T, соответственно.

Вопрос: Будет ли *F* пространством при фиксированном значении Т?

Шаг 2 Классификатор XGBoost. Общее описание

Классификатор будет искаться в виде суммы деревьев, то есть наблюдению x_i будет сопоставляться значение

$$\hat{y}_i^K = \varphi_K(x_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x_i}), \quad f_k \in F \qquad (1)$$

Вопрос 1: Почему в (1) без весов, почему сумма, а не линейная комбинация?

Вопрос 2: Почему хочется улучшать gbm? В чем неэффективность gbm?

Шаг 3. Критерий качества

Критерий качества ***всегда должен состоять из двух частей:*** традиционного критерия качества и регуляризации.

$$Obj(\theta) = L(\theta) + \Omega(\theta)$$

где L традиционный критерий качества, Ω регуляризация.

Примеры критерия качества

Пример 1 mean squared error (MSE).

$$L(\theta) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Пример 2 logistic loss for logistic regression (результирующая переменная y принимает значения 1 и -1)

$$L(\theta) = \sum_{i=1}^{n} (y_i \cdot \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \cdot \ln(1 + e^{\hat{y}_i}))$$

Шаг 4   Критерий качества при построении XGBoost

$$L(\varphi)=\sum_i l(\hat{y}_i, y_i)+\sum_i \Omega(f_k) \qquad (2)$$

где $\Omega(f)=\gamma \cdot T+\frac{1}{2}\lambda \cdot \|w\|^2$

При этом *l* — дифференцируемая выпуклая функция, которая измеряет различие между $\hat{y}_i$ и y_i .

Слагаемое Ω предотвращает переподгонку.

Определение (2) предварительное, будем подправлять.

Пусть построено *K*-1 дерево.

Добавляем (жадно!) к классификатору φ_{K-1} *K*-ое дерево f_K .

Дерево строим так, чтобы минимизировать $L(\varphi_K)$

При построении K-ого дерева известно не только y_i , но и $\hat{y}_i^{(K-1)}$, равный $\varphi_{K-1}(x_i)$

Чтобы найти $\hat{y}_i^{(K)}$, достаточно найти $f_K(x_i)$.

Перепишем критерий качества

$$L(\varphi_K)=\sum_{i=1}^{n} l(y_i, \hat{y}_i^{(K-1)}+f_K(\boldsymbol{x}_i))+\Omega(f_K)$$

Заменяем функцию l на ее разложение в ряд Тейлора.

$$L(\varphi_K)=\sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(K-1)})+g_i f_K(\boldsymbol{x}_i)+\frac{1}{2}h_i f_K^2(\boldsymbol{x}_i)]+\Omega(f_K)$$

$$g_i=\frac{\partial}{\partial \hat{y}_i} l(y_i, \hat{y}_i)$$

$$h_i=\frac{\partial^2}{\partial \hat{y}_i^2} l(y_i, \hat{y}_i)$$

Почему дифференцирование не по $f_K(x_i)$?

Отбросим постоянные слагаемые.

$$\widetilde{L}(\varphi_K)=\sum_{i=1}^{n}[g_i f_K(\boldsymbol{x}_i)+\frac{1}{2}h_i f_K^{\,2}(\boldsymbol{x}_i)]+\Omega(f_K) \qquad (3)$$

Перегруппируем слагаемые в (3).

Зададим $I_j=\{i \mid q(\boldsymbol{x}_i)=j\}$ - множество тех наблюдений, которые дерево относит к конечному узлу *j*.

$$\widetilde{L}(\varphi_K)=\sum_{i=1}^{n}[g_i f_K(\boldsymbol{x}_i)+\frac{1}{2}h_i f_K^{\,2}(\boldsymbol{x}_i)]+\gamma\cdot T+\frac{1}{2}\lambda\cdot\|w\|^2=$$
$$=\sum_{i=1}^{n}[g_i f_K(\boldsymbol{x}_i)+\frac{1}{2}h_i f_K^{\,2}(\boldsymbol{x}_i)]+\gamma\cdot T+\frac{1}{2}\lambda\cdot\sum_{j=1}^{T}w_j^{\,2}= \qquad (4)$$
$$=\sum_{j=1}^{T}[(\sum_{i\in I_j}g_i)w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)w_j^{\,2}]+\gamma\cdot T$$

Считая дерево (структуру дерева $q(x)$) фиксированной , находим вес w_j^{opt} конечного узла номер j, который минимизирует критерий качества

$$w_j^{opt}=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda} \qquad (5)$$

Минимальное значение критерия качества будет равно

$$\widetilde{L}(t)=-\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma\cdot T \qquad (6)$$

И это не все! Правая часть уравнения (6) будет использоваться как критерий "чистоты" при расщеплении дерева, при нахождении структуры дерева q.

Таки образом критерий чистоты будет определяться критерием качества.

Уточним вид критерия чистоты.

Определим *I* как множество наблюдений из обучающей выборки, попавших в родительский узел. Обозначим I_L и I_R множество наблюдений, попавших в левый и правый (соответственно) узлы потомки. Тогда увеличение чистоты при расщеплении будет находиться по формуле

$$L_{split}=\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda}+\frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda}-\frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i+\lambda}\right]-\gamma \qquad (7)$$

Заметим, что формула явно не учитывает число наблюдений, попавших в каждый из потомков.