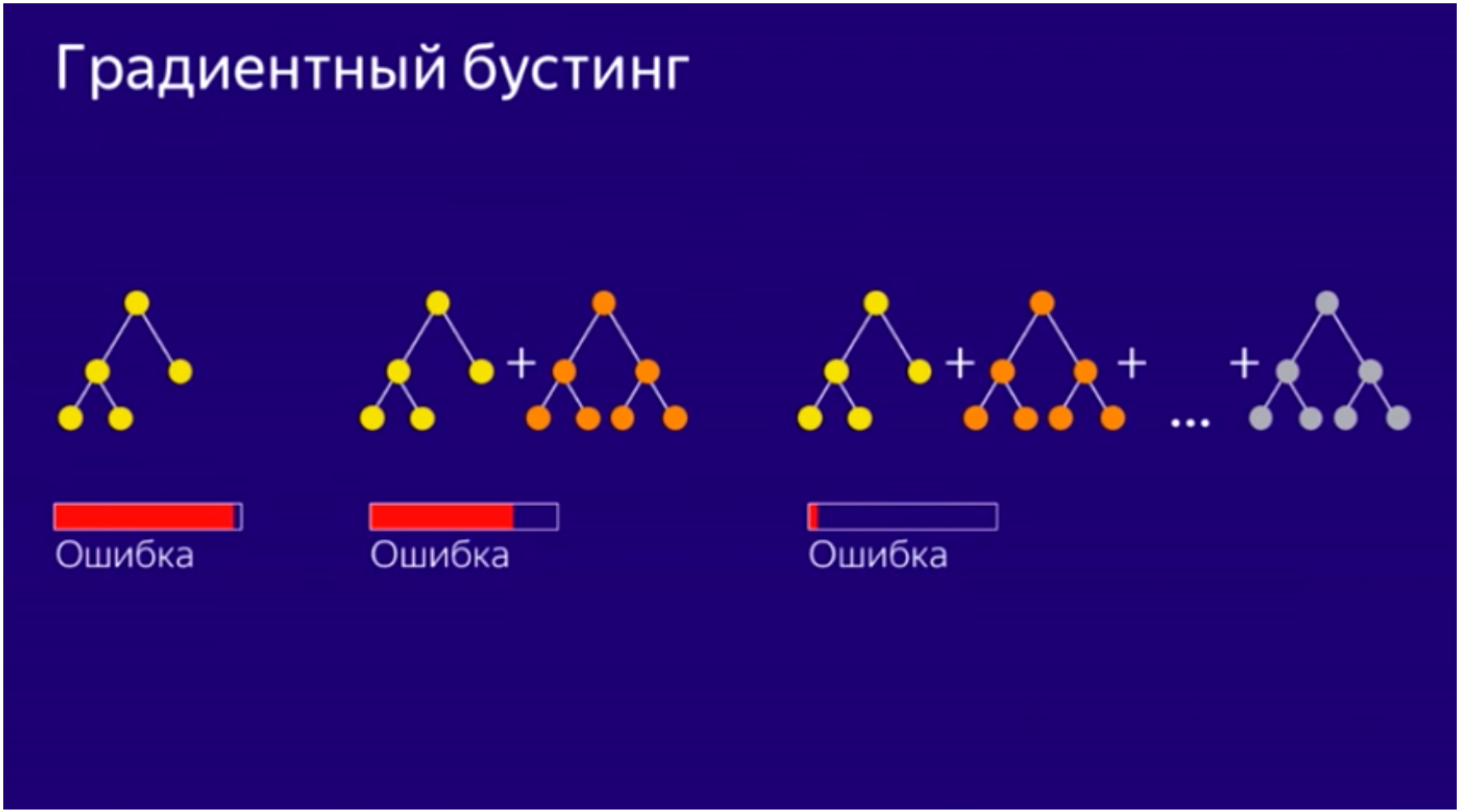
XGBoost, CatBoost

Кантонистова Елена

ekantonistova@hse.ru

19 мая 2018

Схема работы градиентного бустинга



Главные особенности XGBoost

- Базовый алгоритм приближает направление, посчитанное с учетом вторых производных функции потерь.
- Функционал регуляризируется - добавляются штрафы за количество листьев и за норму коэффициентов.
- При построении дерева используется критерий информативности, зависящий от оптимального вектора сдвига.
- Критерий останова при обучении дерева также зависит от оптимального сдвига.

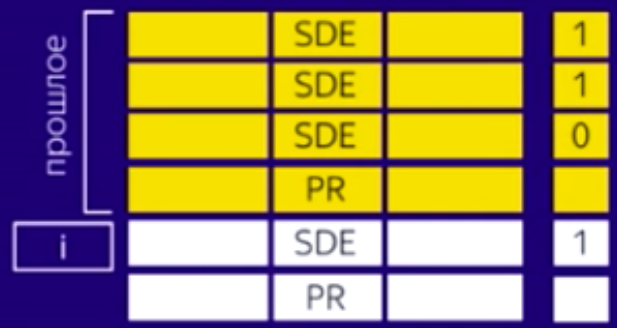
CatBoost

CatBoost - алгоритм, разработанный в Яндексе. Алгоритм является оптимизацией XGBoost, однако, в отличие от XGBoost он умеет работать с категориальными признаками.

1.Симметричные деревья решений

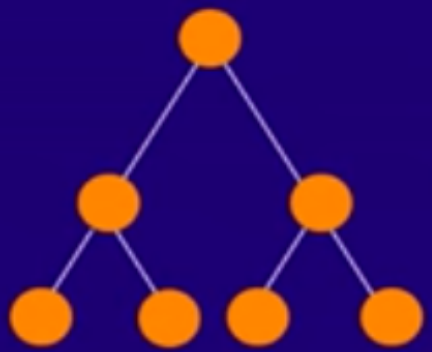
Статистики по категориальным факторам

- One-hot кодирование
- Статистики без использования таргета
- Статистики по случайным перестановкам
- Комбинации факторов



3.Динамический бустинг

Динамический бустинг



$$\text{leafValue(doc)} = \sum_{i=1}^{doc} \frac{\text{g(approx(i), target(i))}}{docs\ in\ the\ past}$$

Бонусы реализации

- Поддержка пропусков в данных
- Обучается быстрее, чем xgboost
- Показывает хороший результат даже без подбора параметров
- Удобные методы: проверка на переобученность, вычисление значений метрик, удобная кросс-валидация и др.

Сравнение алгоритмов

	CatBoost	LightGBM		XGBoost		H2O	
Adult	0.269741	0.276018	+ 2.33 %	0.275423	+ 2.11%	0.275104	+ 1.99%
Amazon	0.137720	0.163600	+ 18.79 %	0.163271	+ 18.55%	0.162641	+ 18.09%
Appet	0.071511	0.071795	+ 0.40 %	0.071760	+ 0.35%	0.072457	+ 1.32%
Click	0.390902	0.396328	+ 1.39 %	0.396242	+ 1.37%	0.397595	+ 1.71%
Internet	0.208748	0.223154	+ 6.90 %	0.225323	+ 7.94%	0.222091	+ 6.39%
Kdd98	0.194668	0.195759	+ 0.56 %	0.195677	+ 0.52%	0.195395	+ 0.37%
Kddchurn	0.231289	0.232049	+ 0.33 %	0.233123	+ 0.79%	0.232752	+ 0.63%
Kick	0.284793	0.295660	+ 3.82 %	0.294647	+ 3.46%	0.294814	+ 3.52%

Logloss

Ссылка

Видео про catboost с объяснением алгоритма:

https://www.youtube.com/watch?v=Q$_x$a4RvnDcY

Презентация сделана с использованием материалов из этого видео.