Оптимизация CART-дерева

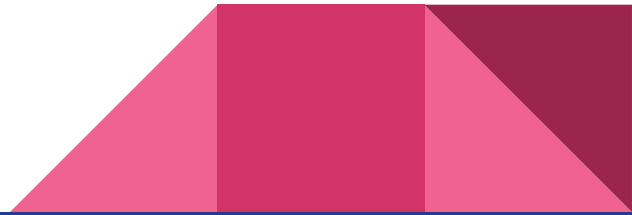
СПбАУ, МИТ, 3 курс, осень

Автор: Бугакова Надежда

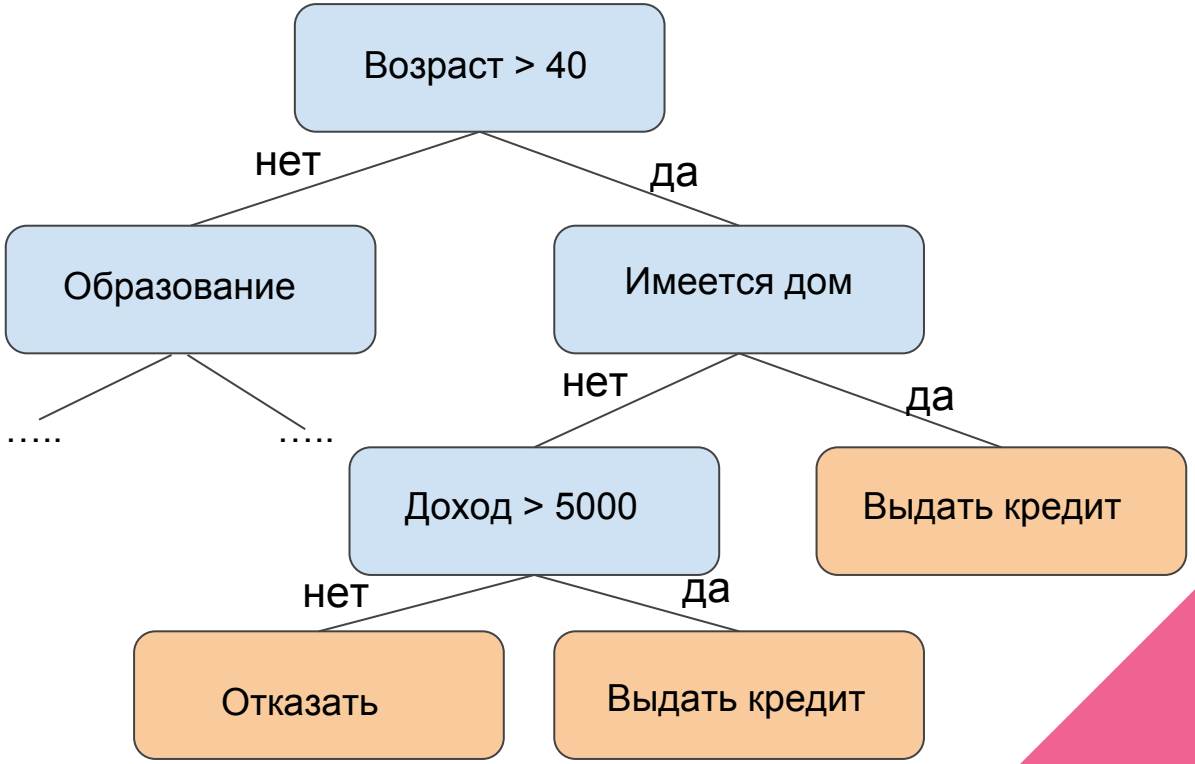
Руководитель: Кураленок Игорь Евгеньевич

Цель

- Разобраться в такой области машинного обучения, как обучение деревьев решений.
- Улучшить стандартную реализацию алгоритма построения деревьев решений CART.

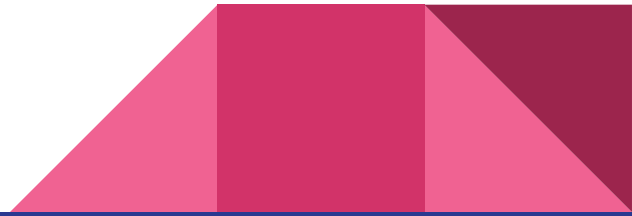


Что такое деревья решений?



Способы построения дерева решений

1. Простые деревья
   - ID3
   - C4.5
   - CART
   - ...
2. Ансамбли
   - Bagging
   - Boosting



CART

- Бинарное
- Может решать, как задачу классификации, так и регрессии
- Жадное построение

Целевая функция: $$T(f) = \sum (f(x) - y)^2$$

Минимизируем при делении листа:

$$\sum_{l \in left} (\bar{y}_{left} - y_l)^2 + \sum_{r \in right} (\bar{y}_{right} - y_r)^2$$

Оптимизация: целевая функция

Улучшение:

- Live one out $$\left(\frac{|left|}{|left|-1}\right)^2 \sum_{l \in left} (\bar{y}_{left} - y_l)^2$$

- Sat $$\frac{|left|(|left|-2)}{|left|^2 - 3|left|+1}^2 \sum_{l \in left} (\bar{y}_{left} - y_l)^2$$

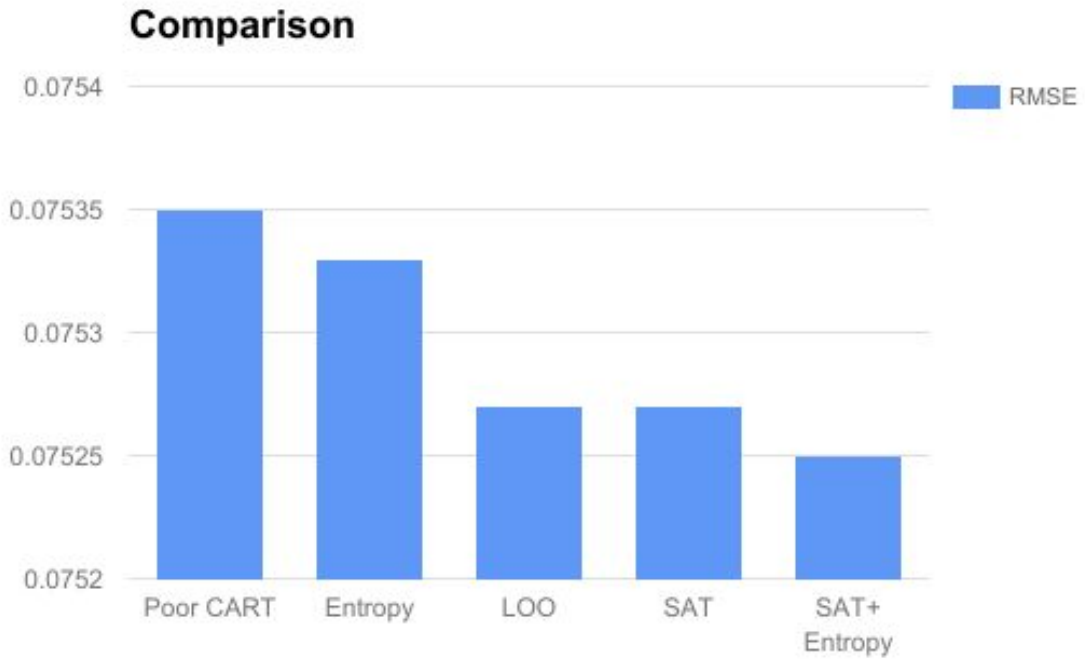
Оптимизация: энтропия

$T_{new}(f) = T_{old}(f) + \lambda H(f)$

$$p_1 = \frac{leftCount + 1}{genCount + 2} \cdot \frac{genCount + 1}{n + leafCount}$$

$$p_2 = \frac{rightCount + 1}{genCount + 2} \cdot \frac{genCount + 1}{n + leafCount}$$

Сравнение

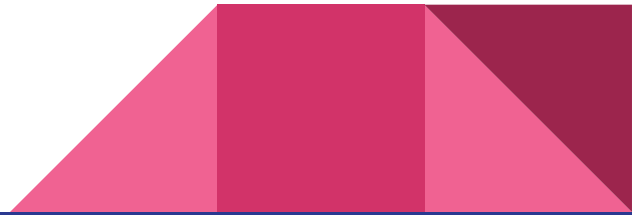

Comparison
0.0754
0.07535
0.0753
0.07525
0.0752
RMSE
Poor CART
Entropy
LOO
SAT
SAT+
Entropy

Возникшие сложности

- Нехватка знаний
- Интегрированность в библиотеку
- Скорость
- Долгое тестирование

Новое

- Работа с уже имеющейся библиотекой(интеграция кода)
- Деревья решений и принципы их работы
- Работа со средними объёмами данных(1М точек)



Ссылка на репозиторий:

https://github.com/spbsu-ml-community/jmll/tree/nadya/

Спасибо за внимание!