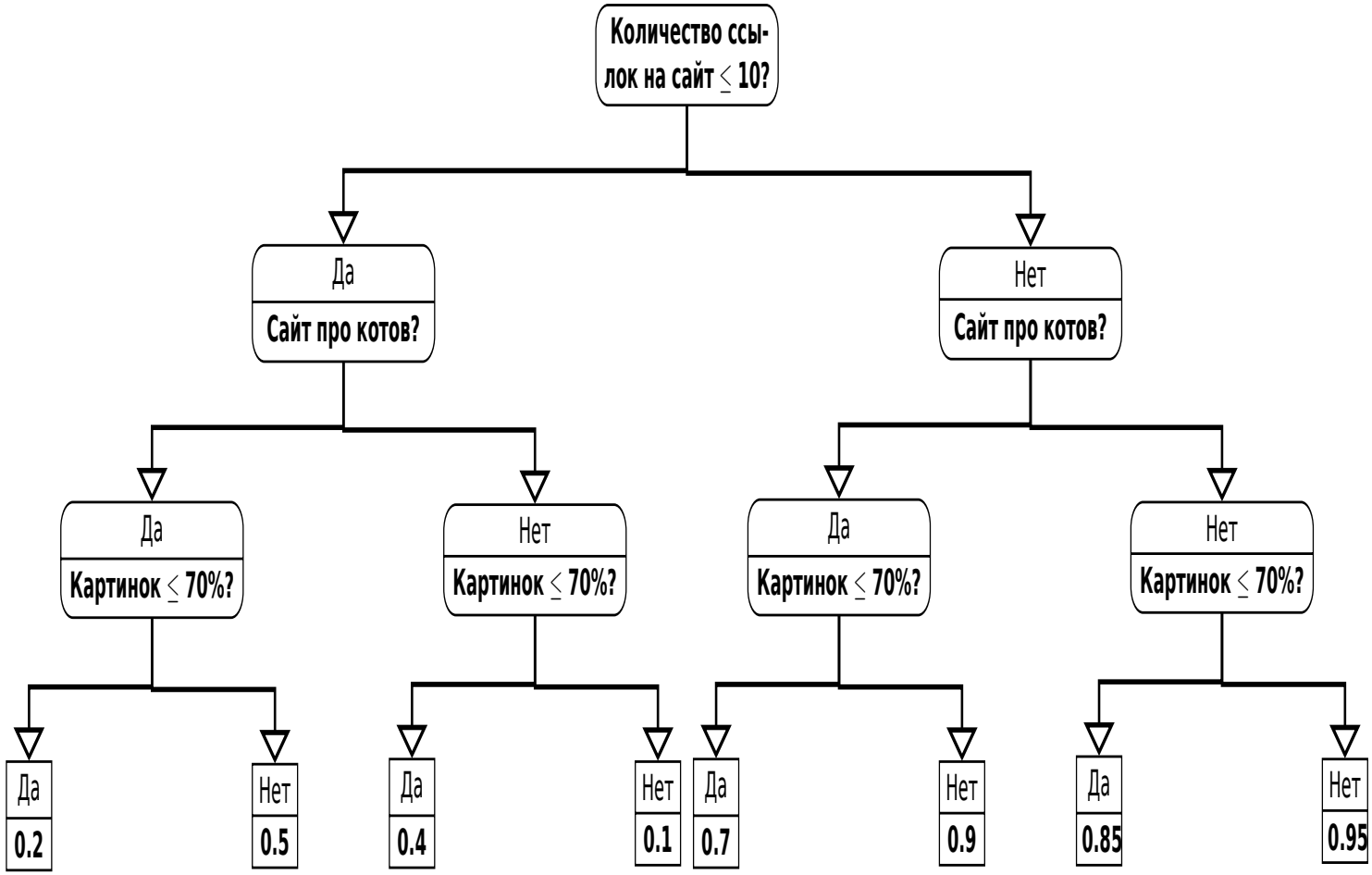
Использование регрессии в алгоритмах машинного обучения

Михаил Белоус
группа 544
руководитель д.ф.-м.н, проф. А.Н. Терехов
рецензент к.ф.-м.н. И.Е. Куралёнок

15 июня 2015 г.

Забывчивые деревья решений



Строим жадно, минимизируя дисперсию

Градиентный бустинг

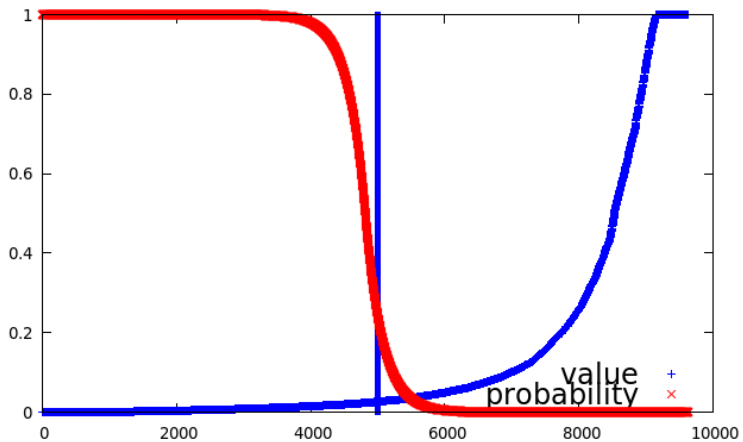
- Функция потерь $\Psi(a, b)$, например $(a - b)^2$
- Базовой алгоритм h обучается на оценках $target$ $h(target) \to F$
- Текущая оценка y
- Псевдооценка $target'_i = -\frac{\partial \Psi(target_i, y_i)}{\partial y_i}$
- Алгоритм обучается на псевдооценках

Постановка задачи

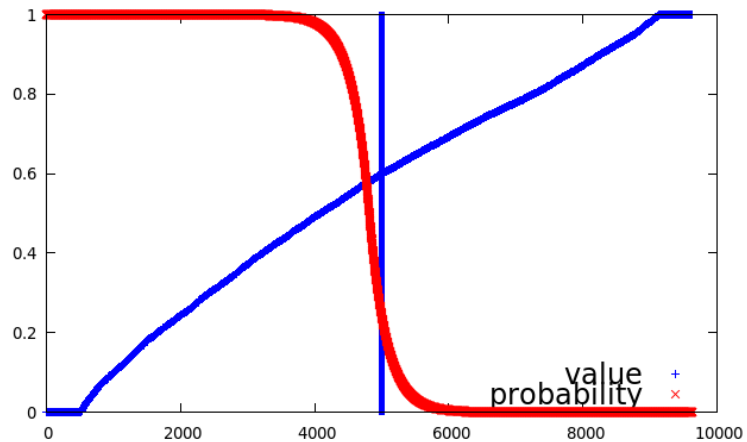
- Создать меру близости точек к листьям деревьев решений
- Разработать алгоритм построения деревьев решений, учитывающий близкие точки
- Реализовать алгоритм полиномиальной регрессии значимых факторов в листьях деревьев решений
- Провести тестирование на открытом наборе данных
- Сравнить с существующими методами обучения ранжированию

Нечеткие границы

Вместо строгих условий вероятность
У разных факторов разное распределение



Длина ссылки



Длина документа

Полиномиальная регрессия

Регрессия вычисляется по формуле

$$F(x, y^L) = \sum_{i_1, i_2 \ldots i_n} x_{i_1} \cdot x_{i_2} \cdots x_{i_n} \cdot y_l^L$$

Коэффициенты находятся как минимум функции ошибки

$$y_1^L = \underset{y}{\operatorname{argmin}} \sum_{x \in Learn, x \in R^L} (F(x, y) - target(x))^2$$

Учтем значения в близких регионах

$$y_2^L = \underset{y}{\operatorname{argmin}} \sum_{x \in Learn, x \in R^L} (F(x, y) - target(x))^2 + \lambda \sum_n ||y - y_1^n||^2$$

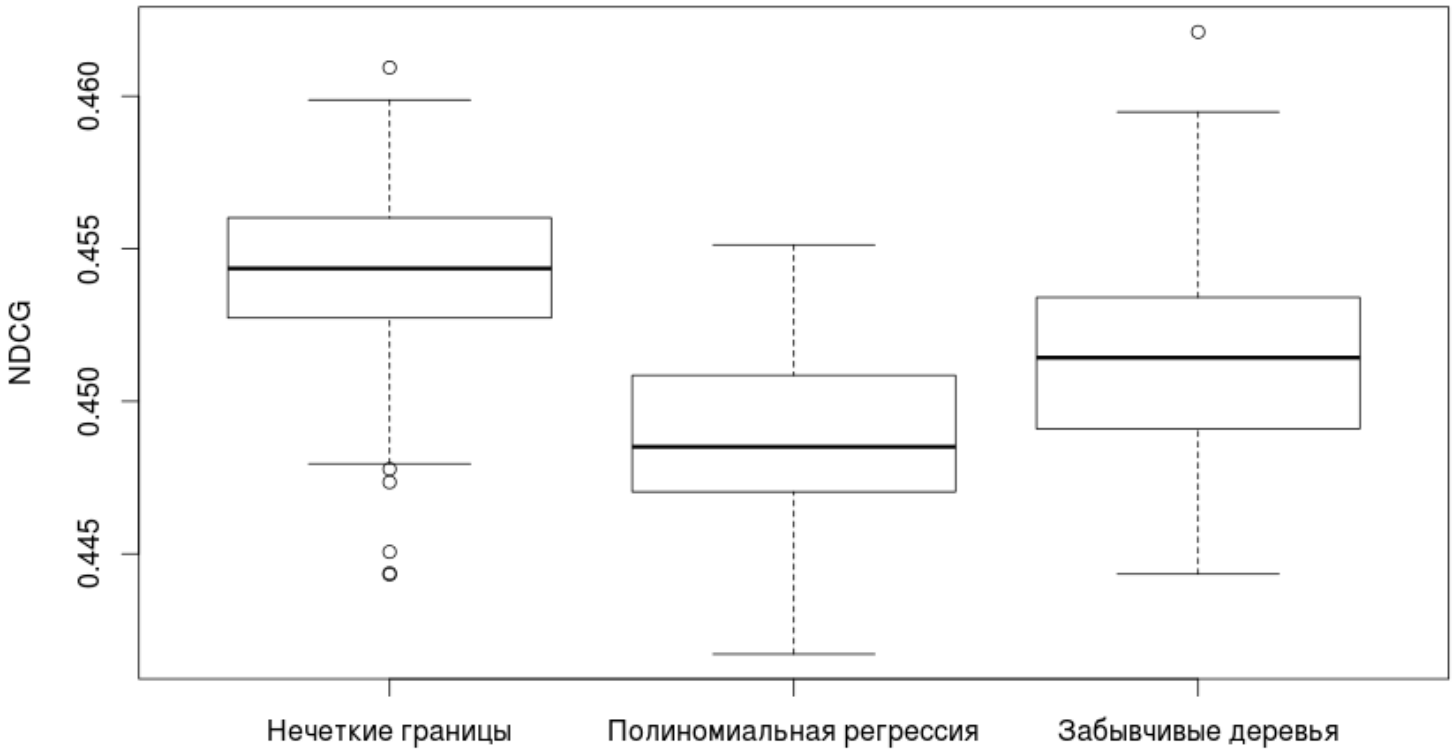
Тестирование

$$DCG_n(q) = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

$$NCDG_n(q) = \frac{DCG_n(q)}{IDCG_n(q)}$$

Примеры факторов из Letor	
Номер фактора	Описание фактора
16	Длина основной части страницы
20	Длина всей страницы
41	Количество ссылок на страницу
42	Количество ссылок на странице
46	Количество подстраниц страницы

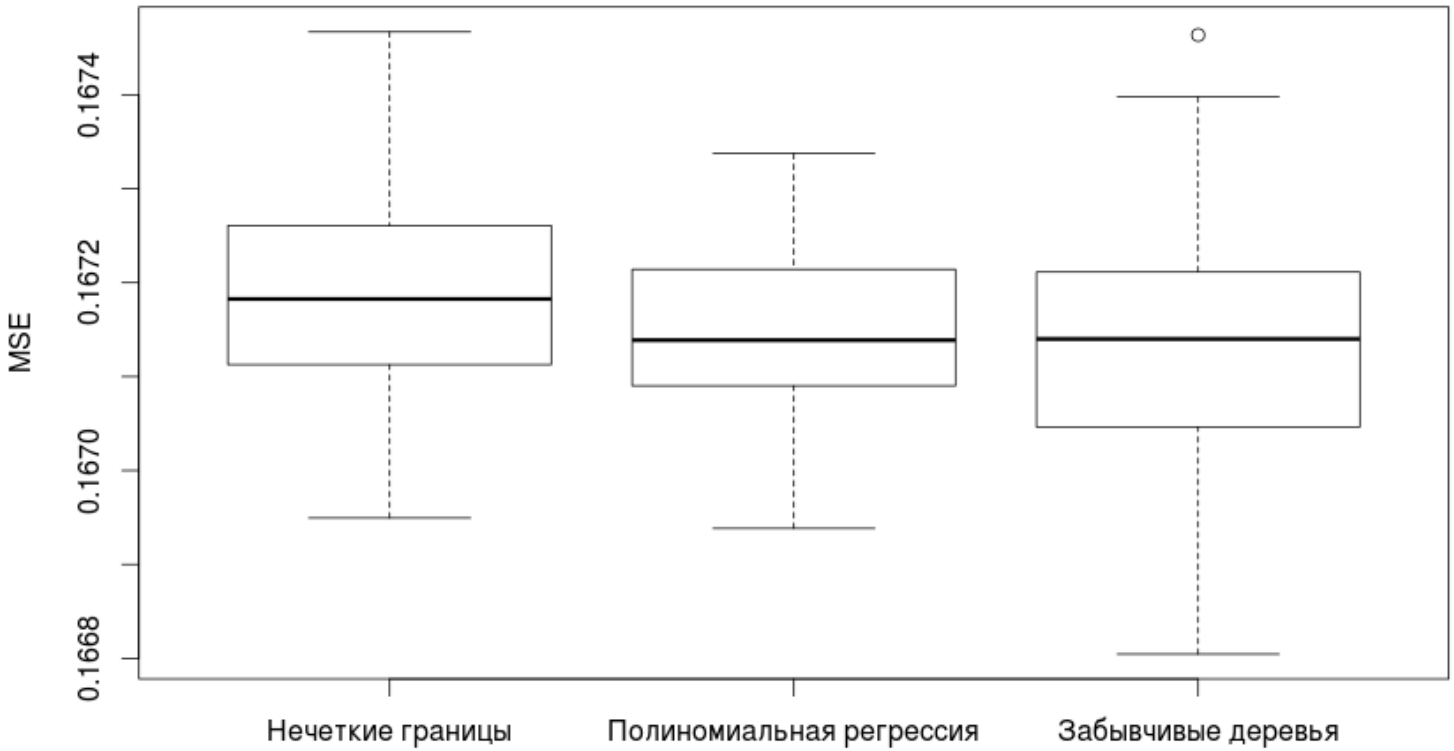
Сравнение методов по NDCG



Сравнение методов по NDCG

Метод	$NDCG_5$
Полиномиальная регрессия	0.4485 ± 0.0006
Забывчивые деревья	0.4514 ± 0.0007
Нечеткие границы	0.4543 ± 0.0005
ListNet	0.4565

Сравнение методов по MSE



Результаты

- Создана мера близости точек к листьям деревьев решений
- Разработан алгоритм построения деревьев решений, учитывающий близкие точки
- Реализован алгоритм полиномиальной регрессии значимых факторов в листьях деревьев решений
- Алгоритмы протестированы на открытом наборе данных
- Проведено сравнение существующими методами обучения ранжированию