CART деревья

Кручинин Дмитрий
Научный руководитель: Кураленок И.Е.

СПбАУ

26 декабря 2016 г.

Содержание

Введение

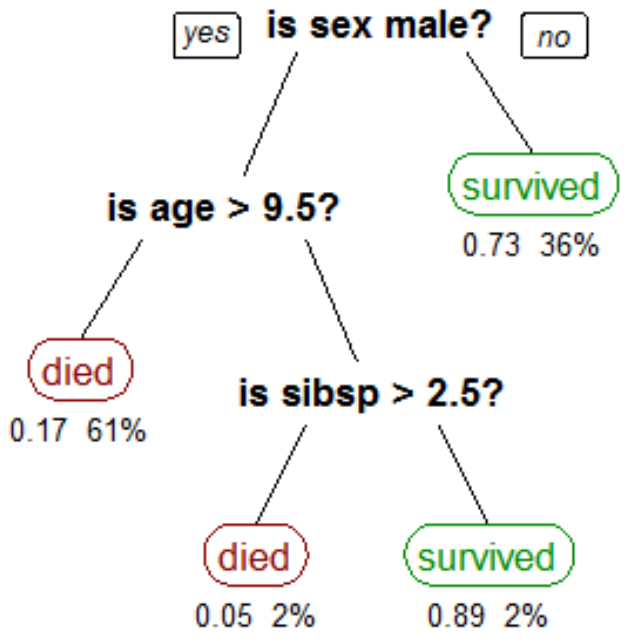


Рис. 1: Пример дерева решений

Построение дерева

Как искать оптимальное разбиение? –

1. Сортировка по каждой компоненте.
2. Полный перебор по каждой из них.
3. Среди результатов выбираем наилучший.

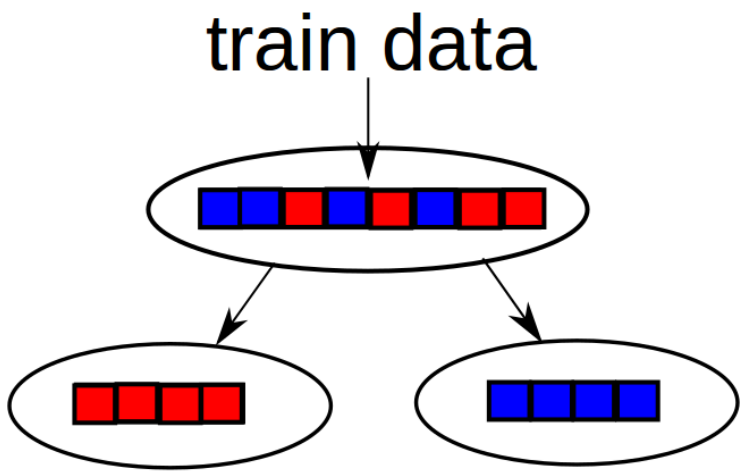
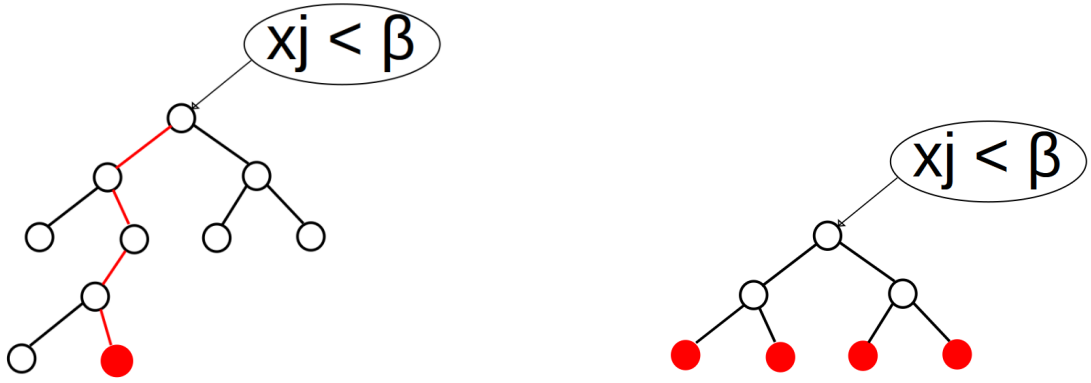


Рис. 2: Шаг построения дерева

Эффективная реализация CART-деревьев



(a) "В глубину" (b) "За один проход"

Рис. 3: Раскрытие листьев

В случае раскрытия "в глубину" – вынужденная сортировка в каждом узле.

Идея: обновляем все листья одновременно, запоминая, для каждого семпла номер его листа.

Получаем: $n \log(n) \leq n \log(\sum_{i=1}^{L} l_i)$ вместо $\sum_{i=1}^{L} l_i \log(l_i)$,
l_i - количество семплов в узле i, L - количество узлов.

Парадокс Штейне

Пусть $\theta, \hat{\theta}$ – истинное значение среднего и его оценка.
$L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$. $R(\theta, \hat{\theta}) = E\{L(\theta, \hat{\theta})\}$

$X = (X_1, X_2, ..., X_N), X_i \sim \mathcal{N}(\mu, \sigma^2)$, все X_i – попарно независимы. Штейн показал, что для X оценка выборочного среднего не является оптимальной относительно R.

$$\widehat{\theta}_{Stein} = \left(1 - \frac{(N-2)\sigma^2}{\|X\|^2}\right) X$$

Связь с деревьями регрессии

Обучающая выборка $\mathcal{X} = \{x_i\}$ для x_i известен "ответ" – $y_i \in \mathbb{R}$

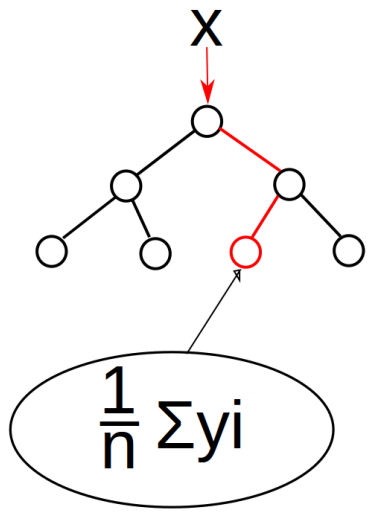


Рис. 4: Ответ на входной вектор x

Пусть L - один из листьев дерева, а y_i^L - все y попавшие в него. Теперь на любой x попавший в лист L наше дерево выдает среднее значение по всем y_i^L.
Предположение: наш $y_i \sim y_i^{truth} + \mathcal{N}(0, \sigma^2)$.

Парадокс Штейне и реальность

Вообще говоря, не все данные распределены нормально.

$$\widehat{\theta}_{Stein} = \left(1 - \frac{(N-2)\sigma^2}{\|X\|^2}\right) X$$

Основная идея оценки заключается в стягивании среднего значения к нулю. Попробуем использовать эту идею в наших целях (здесь мы ипользовали функцию насыщения).

$$\hat{\theta} = X \frac{N}{N+1}$$

Результаты (одно дерево). CT slices.

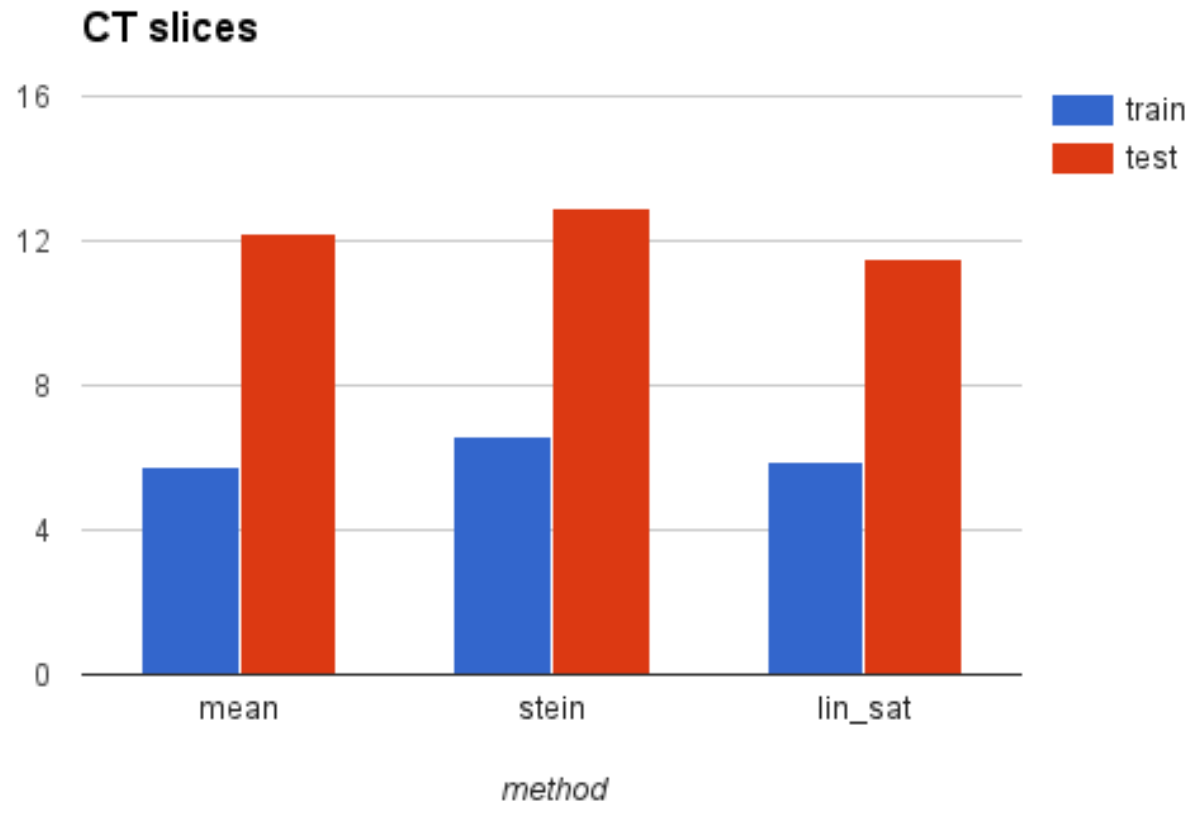


Рис. 5: Оценка местоположения среза КТ (RMSE score)

Результаты (бустинг). House pricing.

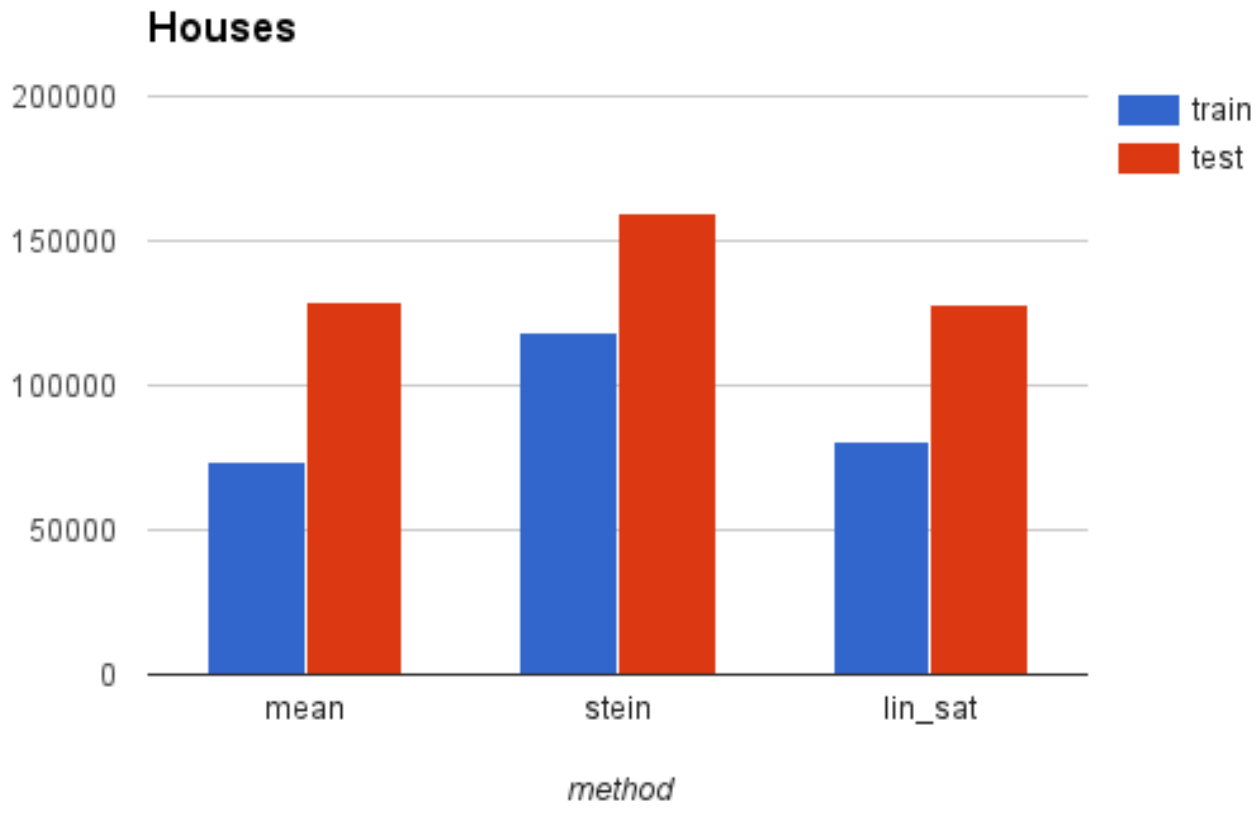


Рис. 6: Оценка стоимости недвижимости (RMSE score)

Результаты (бустинг). Breast cancer.

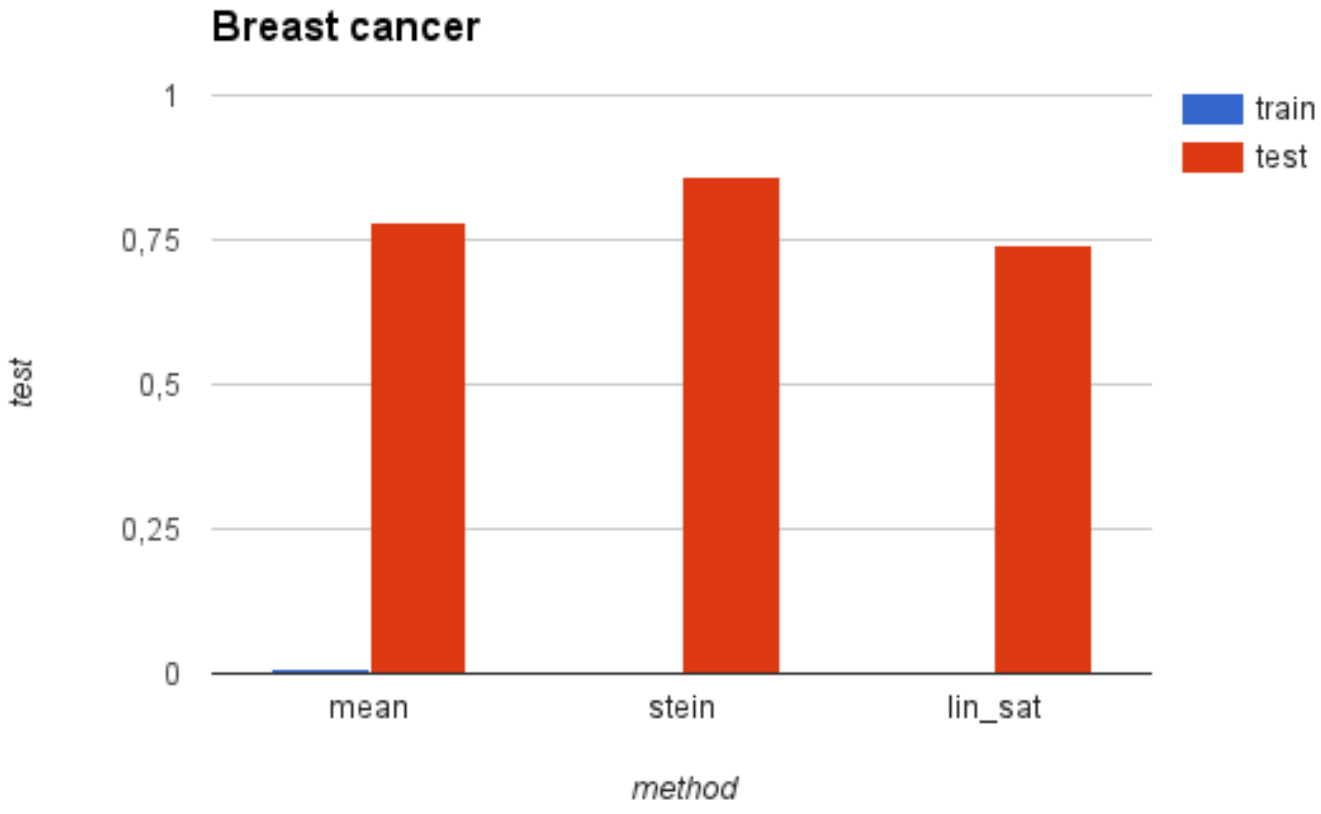


Рис. 7: Предсказание вероятности рака (RMSE score)

Результаты (бустинг). Diabete.

method	**mean**	**stein**	**lin_sat**
train	0,1737461149	0,1908068693	0,1807691669
test	0,4063224416	0,4099584351	0,4029915315

Таблица 1: Предсказание вероятности диабета (RMSE score)

Заключение

В рамках проделанной работы были достигнуты следующие результаты.

- Написана реализация CART-деревьев на Java.
- Проведена оптимизация построения дерева.
- Достигнуто более высокое качество в сравнении со стандартными реализациями.