Разработка методов машинного обучения

Белоус. М. А.
Научный руководитель к.ф.-м.н. Куралёнок И.Е

Санкт-Петербургский Государственный университет

4 іюня 2014 г.

Предметная область

- Задача ранжирования

Предметная область

- Задача ранжирования
- *Оценка качества:*

$$pFound = \sum_{i=0}^{n} pLook_i \cdot pRel_i \tag{1}$$

Базовый метод

- Дерево решений

Базовый метод

- Дерево решений
- Забывчивые деревья

Базовый метод

- Дерево решений
- Забывчивые деревья
- Градиентный бустинг

Базовый метод

- Дерево решений
- Забывчивые деревья
- Градиентный бустинг
- Matrix Net

Постановка задачи

- Построить модификации деревьев, дающие лучший результат

Постановка задачи

- Построить модификации деревьев, дающие лучший результат
- Достичь ускорения, используя GPU

Линейная аппроксимация

x - значимые факторы, $x_0 = 1$

$$F(x, y^l) = \sum_{i=0}^{depth} x_i \cdot y_i^l, x \in R^l \tag{2}$$

Линейная аппроксимация

x - значимые факторы, $x_0 = 1$

$$F(x, y^l) = \sum_{i=0}^{depth} x_i \cdot y_i^l, x \in R^l \tag{2}$$

Метод наименьших квадратов

Линейная аппроксимация

x - значимые факторы, $x_0 = 1$

$$F(x, y^l) = \sum_{i=0}^{depth} x_i \cdot y_i^l, x \in R^l \tag{2}$$

Метод наименьших квадратов
Система линейных уравнений

Квадратичная аппроксимация с граничными условиями

Квадратичная аппроксимация значимых факторов x, $x_0 = 1$

Квадратичная аппроксимация с граничными условиями

Квадратичная аппроксимация значимых факторов x, $x_0 = 1$

$$F(x, y^l) = \sum_{i=0,j=0}^{depth} x_i \cdot x_j \cdot y^l_{i,j}, x \in R^l \tag{3}$$

Квадратичная аппроксимация с граничными условиями

Квадратичная аппроксимация значимых факторов x, $x_0 = 1$

$$F(x, y^l) = \sum_{i=0,j=0}^{depth} x_i \cdot x_j \cdot y^l_{i,j}, x \in R^l \tag{3}$$

- Условия непрерывности

Квадратичная аппроксимация с граничными условиями

Квадратичная аппроксимация значимых факторов x, $x_0 = 1$

$$F(x, y^l) = \sum_{i=0,j=0}^{depth} x_i \cdot x_j \cdot y^l_{i,j}, x \in R^l \tag{3}$$

- Условия непрерывности
- Система уравнений размера $2^{depth} \cdot depth^2$

Квадратичная аппроксимация с граничными условиями

Квадратичная аппроксимация значимых факторов x, $x_0 = 1$

$$F(x, y^l) = \sum_{i=0,j=0}^{depth} x_i \cdot x_j \cdot y^l_{i,j}, x \in R^l \tag{3}$$

- Условия непрерывности
- Система уравнений размера $2^{depth} \cdot depth^2$
- Требуется ускорение

Программная реализация на GPU

Требуется подсчитать следующие статистики:

$$\sum_{x \in Learn, x \in R^l} \sum_{i,j=0}^{depth} x_i \cdot x_j \cdot target(x) \tag{4}$$

$$\sum_{x \in Learn, x \in R^l} \sum_{i,j,f,g=0}^{depth} x_i \cdot x_j \cdot x_f \cdot x_g \tag{5}$$

LQ - разложение Хаусхолдера

Эксперименты

Маленький набор данных Обучающая выборка — 12465 образцов, тестовая 46596 образцов.
13 бинарных факторов, 20 непрерывных.

Эксперименты

Маленький набор данных Обучающая выборка — 12465 образцов, тестовая 46596 образцов.
13 бинарных факторов, 20 непрерывных.
Большой набор данных Обучающая выборка — $\geq$ 1500000 образцов $\geq$ 700 факторов.

Эксперименты

Маленький набор данных Обучающая выборка — 12465 образцов, тестовая 46596 образцов.
13 бинарных факторов, 20 непрерывных.
Большой набор данных Обучающая выборка — $\geq$ 1500000 образцов $\geq$ 700 факторов.

- Линейная апроксимация *pFound* -2%.

Эксперименты

Маленький набор данных Обучающая выборка — 12465 образцов, тестовая 46596 образцов.
13 бинарных факторов, 20 непрерывных.
Большой набор данных Обучающая выборка — $\geq$ 1500000 образцов $\geq$ 700 факторов.

- Линейная апроксимация *pFound* -2%.
- Квадратичная апроксимация
  - Прирост pFound 1.5%.
  - Сбор статистик ускорено в $\approx$ 48 раз.
  - LQ разложение ускорено в $\approx$ 100 раз

Результаты

- Разработаны модификации деревьев, дающие лучшие результаты

Результаты

- Разработаны модификации деревьев, дающие лучшие результаты
- Достигнуто ускорение средствами GPU в ≈ 75 раз