

Структурное обучение для генерации моделей

Артем Максимович Бочкарев

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Научный руководитель: д.ф.-м.н. В.В. Стрижов

Выпускная квалификационная работа магистра

Москва 2018

Введение

Проблема

- Генетический алгоритм символьной регрессии находит точные модели аппроксимации, но требует значительных вычислительных ресурсов
- При аппроксимации выборки не учитываются модели, полученные на похожих задачах

Цель работы

- Автоматизировать построение моделей аппроксимации
- Ускорить нахождение моделей символьной регрессии

Методы

- Прогнозирование структуры модели в виде дерева
- Мета-обучение на парах задача аппроксимации – оптимальная модель

● Символьная регрессия

- ▶ Kulunchakov, A. S., & Strijov, V. V. (2017). Generation of simple structured information retrieval functions by genetic algorithm without stagnation. *Expert Systems with Applications*, 85, 221-230.
- ▶ Koza, J. R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4(2), 87-112.

● Мета-обучение

- ▶ Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.
- ▶ Lemke, C., Budka, M., & Gabrys, B. (2015). Metalearning: a survey of trends and technologies. *Artificial intelligence review*, 44(1), 117-130.

● Прогнозирование структуры деревьев

- ▶ Alvarez-Melis, D., & Jaakkola, T. S. (2016). Tree-structured decoding with doubly-recurrent neural networks.
- ▶ Jin, W., Barzilay, R., & Jaakkola, T. (2018). Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv preprint arXiv:1802.04364*.

Постановка задачи

Задача аппроксимации

Обозначим \mathbf{X} – матрица объект-признак, \mathbf{y} – вектор зависимой переменной.

Описание задачи аппроксимации – выборка $D = (\mathbf{X}, \mathbf{y})$.

Задача аппроксимации должна удовлетворять следующим условиям.

Требования к задаче аппроксимации

- \mathbf{x}_i неслучайны
- $\{\mathbf{x}_i\}_{i=1}^n$ – упорядоченное множество
- \mathbf{y} случайны
- $y_i = f(\mathbf{x}_i) + \varepsilon_i$
 - ▶ ε_i независимы
 - ▶ ε_i гомоскедастичны
 - ▶ $\varepsilon_i \sim \mathcal{N}(0, \sigma)$

Постановка задачи

Выборка мета-обучения

Набор пар $\mathcal{D} = \{D_i = (\mathbf{X}_i, \mathbf{y}_i), f_i\}_{i=1}^m$ назовем мета-выборкой. Она удовлетворяет условиям:

- $\text{dom}(\mathbf{x}_i) = \text{dom}(\mathbf{x}_j) \forall i, j$ (все \mathbf{X} имеют одну область определения)
- f_i является оптимальной моделью для D_i в пространстве \mathfrak{F} :

$$f_i = \arg \min_{f \in \mathfrak{F}} \text{MSE}(\mathbf{y}_i, f_i(\mathbf{X}_i))$$

Задача мета-обучения

Для мета-выборки \mathcal{D} найти оптимальную мета-модель $g : D \rightarrow f$, минимизирующую ошибку на всех задачах аппроксимации:

$$\mathcal{L}(g, \mathcal{D}) = \frac{1}{m} \sum_{i=1}^m \text{MSE}(\mathbf{y}_i, g(D_i)(\mathbf{X}_i))$$

Постановка задачи

В качестве моделей для задачи аппроксимации рассматривается пространство \mathcal{F} моделей символьной регрессии.

Модель символьной регрессии

- Порождается грамматикой G :

$$g \rightarrow B(g, g) | U(g) | S,$$

где B – бинарные функции $(+, *)$, U – унарные функции $(\text{sqrt}, \text{log}, \text{exp})$, а S – множество переменных.

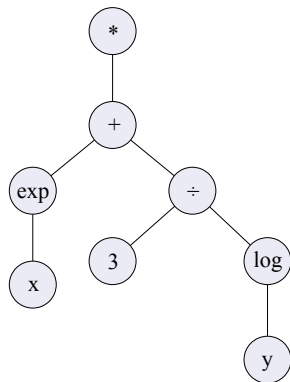
- $f = g_1 \circ g_2 \circ \dots \circ g_k$
- Модель символьной регрессии f представляется в виде дерева Γ_f

Постановка задачи

Дерево Γ_f удовлетворяет следующим условиям:

Дерево Γ_f

- 1 Символ $*$ является корнем дерева;
- 2 Листья Γ_f содержат переменные $x \in S$.
- 3 Узлы дерева v содержат соответствующие функции g ;
- 4 $\#child(v) = arity(g)$;
- 5 Если v_j дочерняя вершина к v_i , то $dom(g_j) \supset cod(g_i)$;
- 6 Дочерние вершины g упорядочены;



$$f = e^x + \frac{3}{\log(y)}$$

Решение

Представление модели

Матрица смежности \mathbf{Z}_f для дерева Γ_f .

Представление задачи аппроксимации

Вектор $\mathbf{d} = [\text{vec}(\mathbf{X}), \mathbf{y}]^T$ является представлением выборки для задачи аппроксимации.

Декомпозиция мета-модели

Мета-модель $g : D \rightarrow f$ является отображением из пространства векторов \mathbb{R}^n в пространство матриц смежности деревьев \mathbb{Z} .

Мета-модель g является суперпозицией двух функций:

$$f = g(D) = g_{\text{rec}}(g_{\text{clf}}(D))$$

Решение

Классификация

Функция классификации $g_{\text{clf}} : \mathbb{R}^n \rightarrow \mathbb{P}$ является отображением из пространства представлений задачи аппроксимации в пространство матриц вероятностей.

$$g_{\text{clf}}(\mathbf{d}) = \mathbf{P}_f,$$

где \mathbf{P}_f – матрица вероятностей ребер в дереве Γ_f . g_{clf} – алгоритм классификации (логистическая регрессия, нейронная сеть).

Восстановление структуры

Функция восстановления структуры $g_{\text{rec}} : \mathbb{P} \rightarrow \mathbb{Z}$ это отображение из пространства матриц вероятностей ребер в пространство матриц смежности дерева. Предлагается два метода восстановления структуры дерева g_{rec} :

- Жадный алгоритм
- Динамическое программирование

Восстановление структуры

Жадный алгоритм

Алгоритм начинает восстановление из корня *. На каждом шаге достраивается ребро с наибольшей вероятностью. Конец работы, если достигли максимальной глубины дерева или в листьях только переменные.

Динамическое программирование

На каждом шаге задача разбивается на подзадачи, соответствующие возможным поддеревьям, для каждой ищется решение, максимизирующее $s(f)$.

- $s(f) = \prod_{e \in f} P_e$ – правдоподобие дерева;
- $s(f) = \frac{1}{n} \sum_{e \in f} P_e$ – средняя вероятность ребра в дереве.

Параметризация

Чтобы метод работал на реальных данных, необходима параметризация. Пусть лучшая непараметрическая модель представима в суперпозиции $f = f_1 \circ \dots \circ f_n$.

Параметризация

Введем параметры для каждой элементарной функции f_i :

$$f_i(\mathbf{x}, \alpha_{i1}, \alpha_{i0}) = \alpha_{i1} f_i(\mathbf{x}) + \alpha_{i0}.$$

Параметрами суперпозиция f являются параметры ее элементарных функций:

$$f(\mathbf{x}) \rightarrow f(\mathbf{x}, \alpha)$$

Полученная функция дифференцируема, вектор параметров α находится градиентным спуском.

Обзор метода

Обучение

- 1 Удалить константы из моделей f
- 2 Обучить g_{clf} на предсказание матрицы вероятностей \mathbf{P}

Тестирование

- 1 Предсказать матрицу \mathbf{P} используя g_{clf}
- 2 Восстановить модель f с помощью g_{rec}
- 3 Параметризовать модель f
- 4 Найти оптимальные параметры α используя метод градиентного спуска

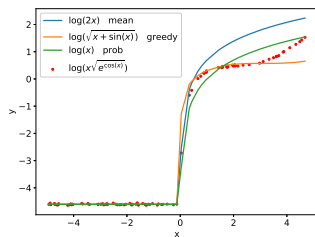
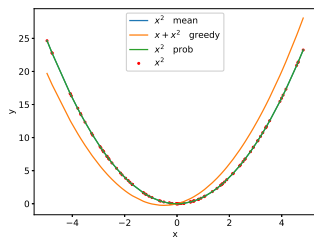
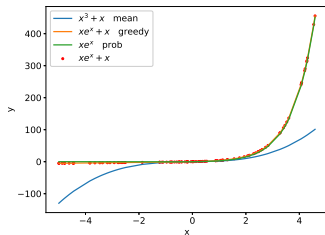
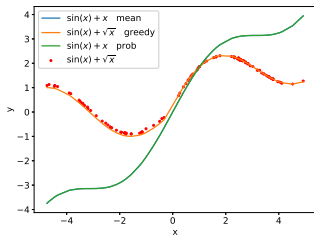
Вычислительный эксперимент

Сегменты сгенерированного временного ряда, порожденные случайными моделями символьной регрессии.

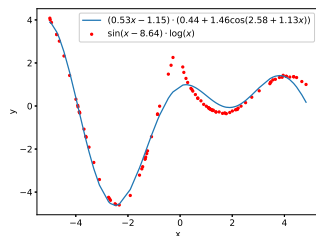
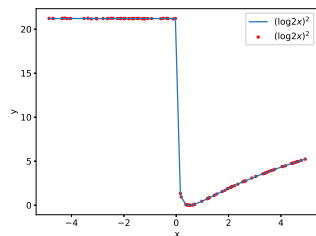
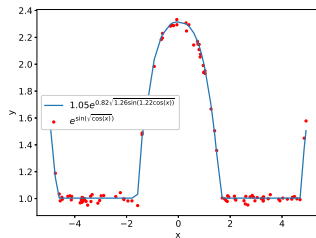
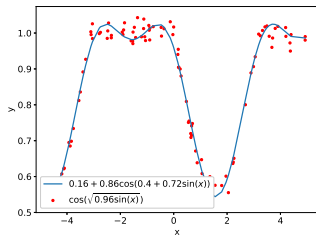
Схема эксперимента

- 1 Сгенерировать ≈ 5000 1-D задач аппроксимации
 - ▶ \mathbf{X} равномерно распределен на $[-5, 5]$
 - ▶ f – случайная модель символьной регрессии
 - ▶ $\mathbf{y} = f(\mathbf{X}) + \mathcal{N}(0, 0.05)$
- 2 Разделить задачи аппроксимации на обучение и контроль
- 3 Обучить g_{clf}
- 4 Предсказать матрицы вероятностей \mathbf{P} для тестовых задач и восстановить модели с помощью g_{rec}

Непараметрический случай



Параметрический случай



Результаты

Ошибка на тестовых данных.

Непараметрический случай

	<i>Random Forest</i>	<i>NN</i>	<i>Logistic regression</i>
Greedy algorithm	5.45 ± 0.04	5.81 ± 0.03	6.3 ± 0.07
DP (tree likelihood)	5.41 ± 0.05	5.65 ± 0.02	5.97 ± 0.04
DP (mean probability)	5.32 ± 0.04	5.72 ± 0.04	6.12 ± 0.06

Параметрический случай.

	<i>Random Forest</i>	<i>NN</i>	<i>Logistic regression</i>
Greedy algorithm	7.02 ± 0.06	7.13 ± 0.05	7.35 ± 0.05
DP (tree likelihood)	6.88 ± 0.05	6.93 ± 0.03	7.01 ± 0.06
DP (mean probability)	6.92 ± 0.07	6.94 ± 0.07	6.99 ± 0.09

Реальные данные

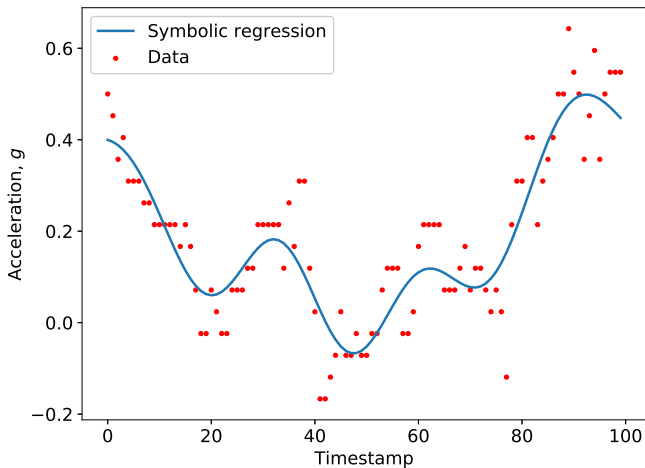


Рис.: Сегмент временного ряда акселерометра

Реальные данные

Вычислительный эксперимент проводился на трех выборках: временные ряды с акселерометра, курса валют и стоимости акций на бирже. Сравнивались результаты и время работы предложенного метода и генетического алгоритма.

Результаты на реальных данных.

	Акселерометр		Курс валют		Курс акций	
	MSE	t, sec	MSE	t, sec	MSE	t, sec
Genetic	0.052 ± 0.02	5.12 ± 0.07	0.012 ± 0.002	6.02 ± 0.05	3.13 ± 0.06	6.3 ± 0.1
Meta	0.054 ± 0.004	0.23 ± 0.02	0.014 ± 0.001	0.28 ± 0.04	3.28 ± 0.03	0.3 ± 0.01

Заключение

- Предложен метод предсказания структуры дерева модели символьной регрессии
- Метод мета-обучения используется для прогнозирования оптимальных моделей аппроксимации
- Вычислительный эксперимент на реальных данных показал значительный прирост скорости в сравнении с генетическим алгоритмом