Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2019

Problem statement for machine learning

Formal problem statement, an analyst has to set

1) an algebraic structure for the dataset from measurements
2) a data generation hypothesis from 1)
3) a model, or a mixture from 2)
4) an error function (quality criteria with restrictions) from 2)
5) an optimization algorithm from 3) and 4)

The result of the model construction is a Cartesian product

$$\{\textbf{models} \times \textbf{data sets} \times \textbf{quality critea}\}.$$

Def: *Big data rejects the i.i.d. (independent and identically distributed random variables) data generation hypothesis from 2). It requests a mixture model.*

Некторые задачи машинного обучения

- Задача оценки параметров модели,
- задача выбора признаков или объектов выборки,
- задача выбора модели оптимальной сложности,
- задача построения и выбора структуры модели,
- задача проверки гипотезы порождения данных.

Предполагается, что функция ошибки $S(\mathbf{w}|D, f)$ задана исходя из

- гипотезы порождения данных,
- либо из практических соображений.

Задача нахождения наиболее правдоподобных параметров

Задана выборка $D = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I}$, функция ошибки модели S и модель — параметрическое семейство функций $f(\mathbf{w}, \mathbf{x})$. Требуется найти такие параметры $\mathbf{w}$ модели, которые бы доставляли минимум функции ошибки

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w}|D, f). \tag{1}$$

Функция ошибки, определенная посредством логарифмической функции правдоподобия

$$S(\mathbf{w}) = -\ln\big(p(D|\mathbf{w}, f)\big),$$

обеспечивает максимизацию правдоподобия параметров. Параметры, найденные минимизацией этой функции ошибок, будут называться наиболее правдоподобными.

Примеры функции ошибки в регрессии и классификации

Регрессия

Гипотеза порождения данных: $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{I})$.
Функция ошибки:

$$S(\mathbf{w}) = \|\mathbf{y} - \mathbf{f}\|_2^2.$$

Классификация

Гипотеза порождения данных: $\mathbf{y} \sim \mathcal{B}(f, 1 - f)$.
Функция ошибки:

$$S(\mathbf{w}) = \sum_{i \in \mathcal{I}} y_i \ln f(\mathbf{w}^\mathsf{T}\mathbf{x})_i + (1 - y_i) \ln\bigl(1 - f(\mathbf{w}^\mathsf{T}\mathbf{x})_i\bigr).$$

Задача выбора оптимального набора признаков

- Задана выборка $D = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I}$.
- Задано случайное разбиение множество индексов элементов выборки $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$.
- Множество независимых переменных $\mathbf{x} = [x_1, \ldots, x_j, \ldots, x_n]$ проиндексировано $j \in \mathcal{J} = \{1, \ldots, n\}$.
- Задано множество моделей-претендентов $\mathfrak{F} = \{f(\mathbf{w}, \mathbf{x})\}$.
- Модель — параметрическое семейство функций $f(\mathbf{w}, \mathbf{x}) = \mu(\mathbf{w}^\mathsf{T}\mathbf{x})$, где μ — функция связи (в случае регрессии $\mu = \mathrm{id}$, в случае классификации $\mu = \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x})}$.
- Структура модели $f_\mathcal{A}$ задана множеством индексов $\mathcal{A} \subseteq \mathcal{J}$ и означает включение переменных $\mathbf{x}_\mathcal{A}$. Иначе, используются только признаки-столбцы матрицы $\mathbf{X}$ с индексами из множества $\mathcal{A}$.
- Задана функция ошибки S.

Задача выбора оптимального набора признаков

Требуется найти такое подмножество индексов $\mathcal{A} \subseteq \mathcal{J}$, которое бы доставляло минимум функции

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(f_{\mathcal{A}} | \mathbf{w}^*, D_{\mathcal{C}})$$

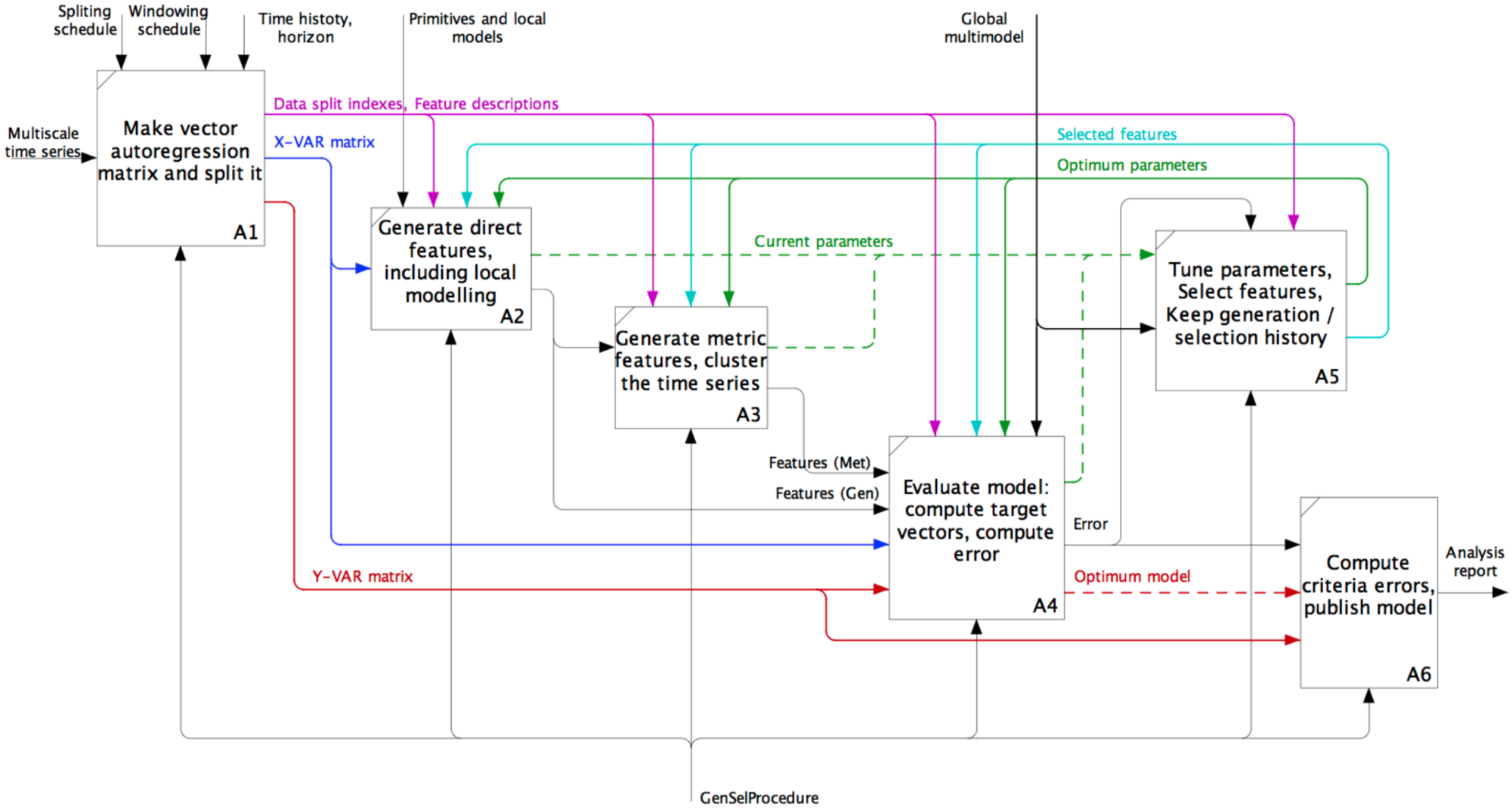
на разбиении выборки D, определенном множеством индексов $\mathcal{C}$.

При этом параметры $\mathbf{w}^*$ модели должны доставлять минимум функции

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w} | D_{\mathcal{L}}, f_{\mathcal{A}})$$

на разбиении выборки, определенном множеством $\mathcal{L}$.

Порождение и выбор моделей, лист A0



Преобразование шкал

- Область деятельности заемщика, номинальная шкала

Nominal	Tourism	Banking	Education
John	1	0	0
Thomas	0	1	0
Sara	0	0	1

- Образование заемщика, ординальная шкала

Ordinal	Primary	Secondary	Higher
John	1	0	0
Thomas	1	1	0
Sara	1	1	1

Группировка признаков: оптимизационная задача

Мы имеем начальную модель, заданную набором индексов $\mathcal{A}$. Добавим полученные в результате группировки признаки и рассмотрим улучшение функционала качества.

$$\begin{array}{rccccccl} \xi = & 1 & 2 & 3 & \dots & c, & c \text{ число категорий}, \xi \in C; \\ & \downarrow & \downarrow & \downarrow & & \downarrow & \\ x_j = & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_c, & |\Gamma| \text{ число групп}, \gamma \in \Gamma. \end{array}$$

Требуется найти функцию

$$h : C \to \Gamma.$$

Задача оптимизации ставится так:

$$(h, |\Gamma|) = \arg \max_{h \in H} S(w)_{\mathcal{A} \cup j}$$

и решается методом полного перебора или генетическим алгоритмом.

Discrete genetic algorithm for grouping

1. There are set of binary vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_P\}$, $\mathbf{a} \in \{1, \ldots, k\}^n$;
2. get two vectors $\mathbf{a}_p, \mathbf{a}_q$, $p, q \in \{1, \ldots, P\}$;
3. chose random number $\nu \in \{1, \ldots, n-1\}$;
4. split both vectors and change their parts:

$$[a_{p,1}, \ldots, a_{p,\nu}, a_{q,\nu+1}, \ldots, a_{q,n}] \to \mathbf{a'}_p,$$

$$[a_{q,1}, \ldots, a_{q,\nu}, a_{p,\nu+1}, \ldots, a_{p,n}] \to \mathbf{a'}_q;$$

5. choose random numbers $\eta_1, \ldots, \eta_Q \in \{1, \ldots, n\}$;
6. replace values in positions $\eta_1, \ldots, \eta_Q$ of the vectors $\mathbf{a'}_p, \mathbf{a'}_q$ for random values from $\{1, \ldots, k\}$;
7. repeat items 2-6 $P/2$ times;
8. evaluate the obtained models.

Repeat R times; here P, Q, R are the parameters of the algorithm and k is desired number of categories.