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

$$\{\textbf{models} \times \textbf{datasets} \times \textbf{quality criteria}\}.$$

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

Quality criteria for model generation and selection

Three sources of quality criteria

1. Business: model operation productivity, agent impact to environment
2. Theory: statistical hypothesis, bayesian inference
3. Technology: optimization requirements, resources

The main criteria of model quality

- Precision: MAPE, AUC
- Stability (diversity): std deviation for prediction, covariance of parameters
- Complexity: structure complexity, MDL, evidence of model

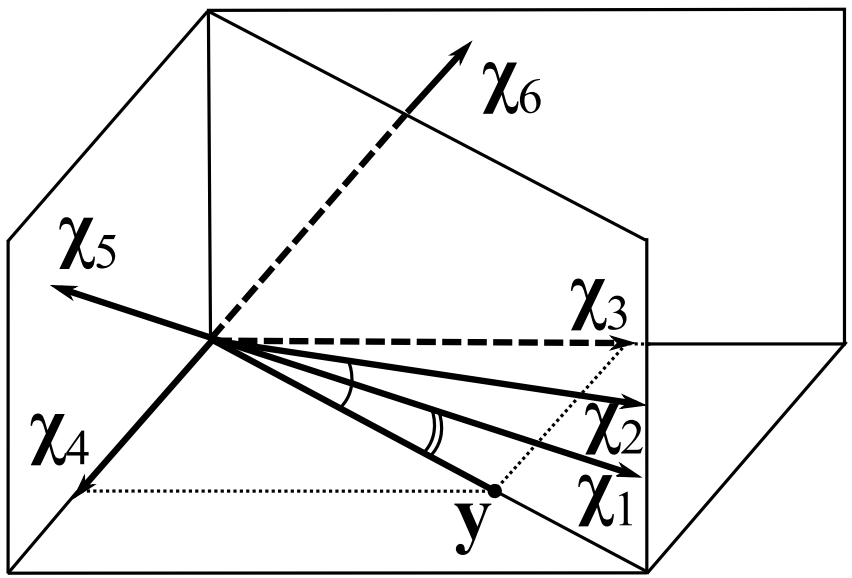
To start an applied project an expert and an analyst set

1. Project goal (the expected result of development)
   main purpose of research
2. Project application (how the project result will be applied)
   environment of measures and impacts
3. Historical data description (data formats and timing)
   algebraic structures of data
4. Quality criteria (how the project quality is measured)
   error function
5. Feasibility of the project (how to prove the project feasibility, list possible risks)
   error analysis

How long the model lives after being put on operation? What replaces it after?

Выбор устойчивой и точной модели

Выборка содержит мультикоррелирующие χ_1, χ_2 и устойчивые χ_5, χ_6 признаки — столбцы матрицы «объект-признак» $\mathbf{X}$. Требуется выбрать два признака из шести.



Точность и устойчивость при заданной сложности

Решение: χ_3, χ_4 — набор ортогональных признаков с наименьшим значением функции ошибки.

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 $\mathcal{L}$.