Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2019

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 $\mathbf{x} = [x_1, \dots, x_j, \dots, x_n]$ проиндексировано $j \in \mathcal{J} = \{1, \dots, n\}$.
- 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

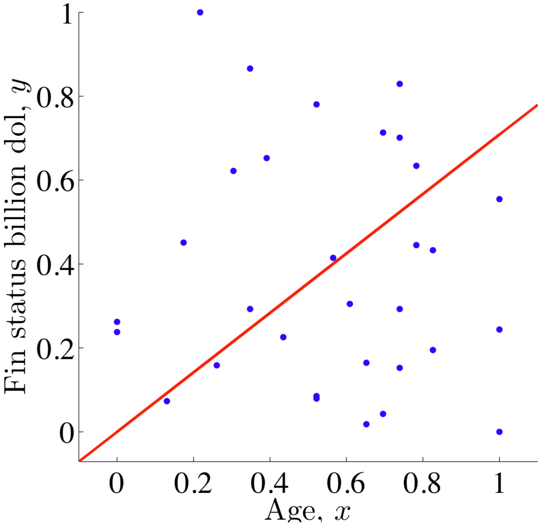
$$\mathcal{A}^* = \arg\min_{\mathcal{A}\subseteq\mathcal{J}} S(f_{\mathcal{A}}|\mathbf{w}^*, D_{\mathcal{C}})$$

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 $\mathcal{L}$.



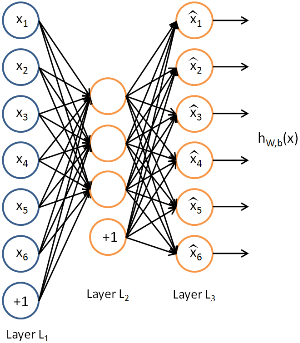
Fin status billion dol, y
Age, x
0
0.2
0.4
0.6
0.8
1

Автокодировщик **h** это монотонное нелинейное отображение входного вектора свободных переменных $\mathbf{x} \in \mathbb{R}^n$ в скрытое представление $\mathbf{h} \in \mathbb{R}^\nu$ следующего вида:

$$\mathbf{h}(\mathbf{x}) = \boldsymbol{\sigma}(\underset{\nu \times n}{\mathbf{W}}\,\mathbf{x} + \mathbf{b}).$$

Скрытое представление **h** создает линейную реконструкцию вектора **x**:

$$\mathbf{r}(\mathbf{x}) = \underset{n \times \nu}{\mathbf{W}'}\mathbf{h} + \mathbf{b}'.$$



Структура автокодировщика

Настройка параметров

Параметры автокодировщика

$$\boldsymbol{\lambda} = \{\mathbf{W}', \mathbf{W}, \mathbf{b}', \mathbf{b}\}$$

оптимизированы таким образом, чтобы сделать реконструкцию $\mathbf{r}(\mathbf{x})$ максимально близкой к $\mathbf{x}$. Функция ошибки автокодировщика:

$$S(\boldsymbol{\lambda}) = \frac{1}{2m}\sum_{i=1}^{m}\|\mathbf{r}(\mathbf{x}_i)-\mathbf{x}_i\|^2+\|\mathbf{W}\|_F^2+\beta\sum_{j=1}^{m}\left[\rho\log\frac{\rho}{\hat{\rho}_j}+(1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}\right],$$

где m — количество элементов в обучающей выборке, β — вес разреживающего слагаемого, ρ — параметр разреженности, желаемое среднее значение каждой компоненты скрытого представления $\mathbf{h}$, а $\hat{\rho}_j$ — среднее значение j-ой компоненты вектора $\mathbf{h}$.

Автоэнкодер и нейросеть: LS, PCA, AE, SAE, 2NN, DNN

$$f = \sigma' \circ \mathbf{w}'^{\mathsf{T}} \boldsymbol{\sigma} \circ \underset{n' \times n}{\mathbf{W}} \mathbf{x}$$

$$+\mathbf{b}$$

$$S = \Sigma_{i \in \mathcal{I}} \big(y_i - f(\mathbf{x}_i)\big)^2 \qquad E_x = \Sigma_{i \in \mathcal{I}} \|\mathbf{x}_i - \mathbf{r}(\mathbf{x}_i)\|_2^2$$