Линейная регрессия и немного алгебры

Александр Сенов

16 мая 2017 г.

E-Contenta

Table of contents

Задача обучения

Обучение

*Говорят, что компьютерная программа обучается на основе опыта **E** по отношению к некоторому классу задач **T** и меры качества **P**, если качество решения задач из **T**, измеренное на основе **P**, улучшается с приобретением опыта **E**.*

— T.M. Mitchell Machine Learning. McGraw-Hill, 1997.

- **E** — ? **T** — ? **P** — ?

Обучение с учителем

- Опыт **E**: $\{x^{(i)}, y^{(i)}\}_{i=1}^{N}$, $x^{(i)} \in \mathcal{X}$, $y^{(i)} \in \mathcal{Y}$.
- Мера качества **P**: $\sum_{i=1}^{N} loss\left(y^{(i)}, \hat{y}^{(i)}\right)$
  - Регрессия:

    $$loss\left(y^{(i)}, \hat{y}^{(i)}\right) = \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

  - Классификация:

    $$loss\left(y^{(i)}, \hat{y}^{(i)}\right) = y^{(i)} \log\left(\hat{y}^{(i)}\right) + (1 - y^{(i)}) \log\left(1 - \hat{y}^{(i)}\right)$$

- Оптимальная $\hat{f}$ из заданного семейства $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{N} loss\left(y^{(i)}, f(x^{(i)})\right)$$

— принцип минимизации *эмпирического риска*

Линейная регрессия

- Рассмотрим класс задач **T**:

$$y = x_1\theta_1^* + \ldots + x_d\theta_d^* + \varepsilon.$$

- Семейство $\mathcal{F}$:

$$\mathcal{F} = \{f(\cdot \mid \theta)\}_\theta\,,$$
$$f(x \mid \theta) = x_1\theta_1 + \ldots + x_d\theta_d\,.$$

Линейная алгебра

Скаляр

- Целые: $-1, 0, 1, 2, 3, \ldots \in \mathbb{Z}$;
- Вещественные: $0.99, 3.141592.., 6.666.., \ldots\ldots \in \mathbb{R}$;
- Комплексные: $a + ib$, $a \in \mathbb{R}$, $b \in \mathbb{R}$
- Операции: $+, -, \cdot, /, \hat{}$

Вектор

$\mathbf{x} \in \mathbb{R}^d$

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_d \end{bmatrix}$$

$\mathbf{x}_i \in \mathbb{R} \quad \forall i = 1 \dots d.$

Матрица

$\mathbf{X} \in \mathbb{R}^{n \times m}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{1,1} & \mathbf{X}_{1,2} & \cdots & \mathbf{X}_{1,m} \\ \mathbf{X}_{2,1} & \mathbf{X}_{2,2} & \cdots & \mathbf{X}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_{n,1} & \mathbf{X}_{n,2} & \cdots & \mathbf{X}_{n,m} \end{bmatrix}$$

$\mathbf{X}_{i,j} \in \mathbb{R} \quad \forall i = 1 \dots n,\ j = 1 \dots m$

$\mathbf{X}_{i,:} \in \mathbb{R}^m$ — строки, $\mathbf{X}_{:,j} \in \mathbb{R}^n$ — столбцы

Тензор

- $\mathbf{X} \in \mathbb{R}^{d_1 \times \ldots \times d_K}$

$$\mathbf{X} = [x_{i_1,\ldots,i_K}]_{i_1=1,\ldots,i_K=1}^{i_1=d_1,\ldots,i_K=d_K}$$

  $x_{i_1,\ldots,i_K} \in \mathbb{R} \quad \forall i_1 = 1 \ldots d_1, \ldots, i_K = 1 \ldots d_K$
- Поэлементные операции: $\mathbf{X} + \mathbf{Y}, \mathbf{X} - \mathbf{Y}$
- Векторные операции: $\mathbf{X}^\top, \mathbf{XY}, \mathbf{X}^{-1},$

Матричные операции

$\mathbf{A} \in \mathbb{R}^{n \times m}$

- Умножение на скаляр

$$(\alpha \mathbf{A})_{i,j} = \alpha \mathbf{A}_{i,j}$$

- Сложение $\mathbf{B} \in \mathbb{R}^{n \times m}$

$$(\mathbf{A} + \mathbf{B})_{i,j} = \mathbf{A}_{i,j} + \mathbf{B}_{i,j}$$

- Умножение $\mathbf{A} \in \mathbb{R}^{m \times d}$

$$(\mathbf{AB})_{i,j} = \sum_{k=1}^{m} \mathbf{A}_{i,k} \mathbf{B}_{k,j}$$

- Транспонирование

$$\left(\mathbf{A}^\top\right)_{i,j} = \mathbf{A}_{j,i}$$

Дополнительные определения

- Линейная комбинация $\{\mathbf{x}_i\}_{i=1}^{n}$: $\sum_{i=1} \alpha_i \mathbf{x}_i$
- Линейная оболочка $\{\mathbf{x}_i\}_{i=1}^{n}$: $\mathrm{span}\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) \subset \mathbb{R}^d$
- Ранг матрицы $\mathbf{X}$: $\mathrm{rank}\left(\mathbf{X}\right) =$ мощности базиса $\mathrm{span}\left(\mathbf{X}_{:,1}, \ldots, \mathbf{X}_{:,d}\right)$
- l_p–норма вектора: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{d} \mathbf{x}_i^p\right)^{\frac{1}{p}}$
- Норма Фробениуса: $\|\mathbf{X}\|_F = \sqrt{\sum\limits_{i=1,j=1}^{n,m} \mathbf{x}_{i,j}^2}$

Метод наименьших квадратов

Линейная регрессия в матричной форме

Оригинальная запись

- Модель данных: $y = x_1\theta_1^* + \ldots + x_d\theta_d^* + \varepsilon$
- Решающая функция: $f(x \mid \theta) = x_1\theta_1 + \ldots + x_d\theta_d$

Матричная запись

- Матрица $\mathbf{X}$: $\mathbf{X}_{i,:} = x^{(i)}$, вектор $\mathbf{y}_i = y^{(i)}$
- Модель данных:
$$\mathbf{y} = \mathbf{X}\theta^* + \varepsilon$$
- Решающая функция:
$$f(\mathbf{x} \mid \theta) = \mathbf{x}^\top\theta$$
- Мера качества $\mathbf{P}$:
$$\sum_{i=1}^{N}\left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \|\mathbf{y} - \mathbf{X}\hat{\theta}\|_2^2$$
$$\rightsquigarrow \mathbf{y} \sim \mathbf{X}\hat{\theta}$$

СЛАУ

$\mathbf{X} \in \mathbb{R}^{N \times d}$, $\mathbf{y} \in \mathbb{R}^N$, $\theta \in \mathbb{R}^d$

$$\mathbf{X}\theta = \mathbf{y}$$

$$\mathbf{X}_{:,1}\theta_1 + \ldots + \mathbf{X}_{:,d}\theta_d = \mathbf{y}$$

$$\begin{bmatrix} \mathbf{X}_{1,1} \\ \ldots \\ \mathbf{X}_{N,1} \end{bmatrix} \theta_1 + \cdots + \begin{bmatrix} \mathbf{X}_{1,d} \\ \ldots \\ \mathbf{X}_{N,d} \end{bmatrix} \theta_d = \begin{bmatrix} \mathbf{y}_1 \\ \ldots \\ \mathbf{y}_N \end{bmatrix}$$

Свойства

- $\exists\, \hat{\theta}: \ \mathbf{X}\hat{\theta} = \mathbf{y} \ \Leftrightarrow \ \mathbf{y} \in \operatorname{span}(\mathbf{X}_{:,1}, \ldots, \mathbf{X}_{:,d}) \subset \mathbb{R}^N$
- $\forall\, \mathbf{y}\, \exists\, \hat{\theta}: \ \mathbf{X}\hat{\theta} = \mathbf{y} \ \Leftrightarrow \ \operatorname{span}(\mathbf{X}_{:,1}, \ldots, \mathbf{X}_{:,d}) = \mathbb{R}^N$
- $d < N \Rightarrow \emptyset$
- $N = d$ и $\operatorname{rank}(\mathbf{X}) = N \Rightarrow \exists!\, \hat{\theta}$

Зачастую $N >> d$. Хотелось бы просто $\hat{\theta} = \mathbf{X}^{-1}\mathbf{y}$.

Обратная матрица

$\mathbf{I} = \mathrm{diag}(1, \dots, 1) \in \mathbb{R}^{\text{следует из контекста}}$

- Для $\mathbf{A} \in \mathbb{R}^{n \times n}$, может $\exists\ \mathbf{A}^{-1}$:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

- Для $\mathbf{A} \in \mathbb{R}^{n \times m}$, может $\exists\ \mathbf{A}^{-1}_{\mathit{left}}$ и $\exists\ \mathbf{A}^{-1}_{\mathit{right}}$:

$$\begin{aligned} \mathbf{A}^{-1}_{\mathit{left}}\mathbf{A} &= \mathbf{I}, & \mathbf{A}^{-1}_{\mathit{left}} &= (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}, \\ \mathbf{A}\mathbf{A}^{-1}_{\mathit{right}} &= \mathbf{I}, & \mathbf{A}^{-1}_{\mathit{right}} &= \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}. \end{aligned}$$

Примеры

- $\mathbf{I}^{-1} = \mathbf{I}$
- $\mathrm{diag}\left(\lambda_1, \dots, \lambda_d\right)^{-1} = \mathrm{diag}\left(\lambda_1^{-1}, \dots, \lambda_d^{-1}\right)$
- $\mathbf{V}$ — ортогональная $\Leftrightarrow \mathbf{V}^{-1} = \mathbf{V}^\top$

Собственные пространство матрицы

$\mathbf{A} \in \mathbb{R}^{n \times n}$

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$$

- λ — собственное число, $\mathbf{v}$ — собственный вектор, $\|\mathbf{v}\|_2^2 = 1$
- Пусть $\mathbf{A}^\top = \mathbf{A} \Rightarrow \exists \{\lambda_i\}_{i=1}^n \mathbb{R}$, $\{\mathbf{v}^{(i)}\}_{i=1}^n \subset \mathbb{R}^n$: $\mathbf{A}\mathbf{v}^{(i)} = \lambda_i \mathbf{v}^{(i)}$
- Спектральное разложение

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}$$

  $\mathbf{V} = \left[\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)}\right]$, $\Lambda = \mathrm{diag}\left(\lambda_1, \dots, \lambda_n\right)$
- $\lambda_i \neq 0 \; \forall i = 1..n \Rightarrow \exists \, \mathbf{A}^{-1}$

МНК

Идея

$$\mathbf{X}\theta \sim \mathbf{y}$$
$$\mathbf{X}^\top \mathbf{X}\theta \sim \mathbf{X}^\top \mathbf{y}$$

предположим, что $\mathbf{X}^\top \mathbf{X}$ обратима

$$\theta \sim \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{X}_{left}^{-1} \mathbf{y}$$

$\hat{\theta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$ — оценка МНК:

$$\hat{\theta} = \operatorname{argmin} \theta \|\mathbf{y} - \mathbf{X}\theta\|_2^2$$

Свойства МНК

$$\|\mathbf{y} - \mathbf{X}\hat{\theta}\|_2^2 = \|\mathbf{X}\theta^* + \varepsilon - \mathbf{X}\hat{\theta}\|_2^2 = \|\mathbf{X}\theta^* + \varepsilon - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\|_2^2$$
$$= \|\mathbf{X}\theta^* + \varepsilon - \mathbf{X}\theta^* - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\varepsilon\|_2^2$$
$$= \|(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top)\,\varepsilon\|_2^2$$

- $\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ — проекция на $\operatorname{span}(\mathbf{X}_{:,1}, \ldots, \mathbf{X}_{:,d})$
- Если $\Sigma_\varepsilon = \sigma^2 I$, то $E\hat{\theta} = \theta^*$ и $\Sigma_{\hat{\theta}} = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$
- Пусть λ_i — с.ч. $\left(\mathbf{X}^\top\mathbf{X}\right)$, $\frac{|\lambda_{max}|}{|\lambda_{min}|}$ — число обусловленности матрицы, характеризует чувствительность ко входу $(\mathbf{A}^{-1}\mathbf{b})$
- Для $\mathbf{P}$: $\|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \alpha\|\theta\|_2^2$, оценка $\hat{\theta} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}$

Спасибо!

Вопросы?