Базовые модели машинного обучения: логистическая регрессия

Першин Антон Юрьевич, Ph.D.
Никольская Анастасия Николаевна

Программа «Большие данные и распределенная цифровая платформа»
Санкт-Петербургский государственный университет

Практика по дисциплине «Технологии ИИ»
31 марта 2023 г.

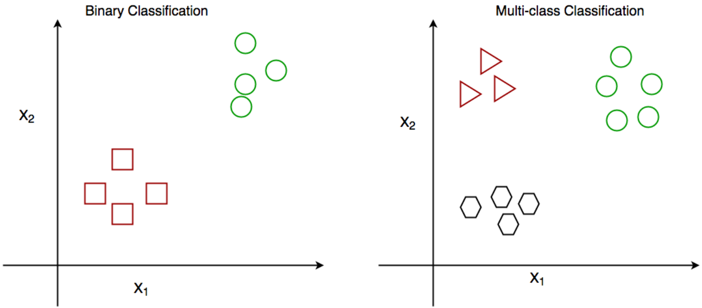
Задача классификации

→ Пусть наблюдение характеризуется признаками $\boldsymbol{x} \in \mathbb{R}^M$

→ Известно, что каждому наблюдению соответствует класс $y \in \mathcal{Y}$, где без потери общности $\mathcal{Y} = \{1, 2, \ldots, K\}$

→ Мы хотим построить правило $h : \mathbb{R}^M \to \mathcal{Y}$, позволяющее по наблюдению предсказывать его класс

→ Предполагается, что имеется тренировочный набор данных, то есть набор пар $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, y_i) | \boldsymbol{x}^{(i)} \in \mathbb{R}^M, y_i \in \mathcal{Y}\}_{i=1}^N$



Задача классификации

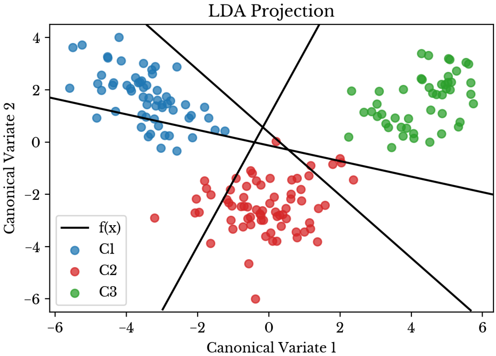
→ Если представить, что мы знаем апостериорное распределение $P(Y = y|\boldsymbol{x})$, то не составит труда задать правило $h(\boldsymbol{x})$ для предсказания:

$$h(\boldsymbol{x}) = \text{argmax}_{y \in \mathcal{Y}} P(Y = y|\boldsymbol{x}). \tag{1}$$

→ Распределения неизвестны $\implies$ можно оценить с помощью статистических методов

→ Для пространств большой размерности их оценка затруднена, поэтому чаще используют подход с построением сепарирующей поверхности

→ Например, линейный дискриминантный анализ использует линейную функцию для разделения классов



- Логистическая регрессия появляется из желания смоделировать апостериорное распределение линейной функцией, при этом гарантировав, что все вероятности суммируются в единицу и лежат в интервале $[0; 1]$
- Это возможно при использовании logit-преобразования:

$$\log \frac{P(Y=1|X=\boldsymbol{x})}{P(Y=K|X=\boldsymbol{x})} = \beta_{10} + \boldsymbol{\beta}_1^T \boldsymbol{x},$$

$$\log \frac{P(Y=2|X=\boldsymbol{x})}{P(Y=K|X=\boldsymbol{x})} = \beta_{20} + \boldsymbol{\beta}_2^T \boldsymbol{x},$$

$$\dots$$

$$\log \frac{P(Y=K-1|X=\boldsymbol{x})}{P(Y=K|X=\boldsymbol{x})} = \beta_{K-1,0} + \boldsymbol{\beta}_K^T \boldsymbol{x}.$$

Логистическая регрессия

→ Так как сумма вероятностей должна быть равна единице, легко вывести функцию вероятности:

$$P(Y = y|X = \boldsymbol{x}) = \frac{exp(\beta_{k0} + \boldsymbol{\beta}_k^T \boldsymbol{x})}{1 + \sum_{j=1}^{K-1} exp(\beta_{j0} + \boldsymbol{\beta}_j^T \boldsymbol{x})}, \quad k = 1, \dots, K-1,$$

$$P(Y = K|X = \boldsymbol{x}) = \frac{1}{1 + \sum_{j=1}^{K-1} exp(\beta_{j0} + \boldsymbol{\beta}_j^T \boldsymbol{x})}.$$

→ Таким образом, распределение $P(Y = y|X = \boldsymbol{x}) = p_y(\boldsymbol{x}; \boldsymbol{\theta})$ параметризовано $\boldsymbol{\theta} = [\beta_{10}, \boldsymbol{\beta}_1^T, \dots, \beta_{K-1,0}, \boldsymbol{\beta}_{K-1}^T]$

→ Логарифмическая функция правдоподобия в таком случае имеет вид

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(Y = y_i|X = \boldsymbol{x}^{(i)}) = \sum_{i=1}^{N} \log p_{y_i}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}). \qquad (2)$$

Логистическая регрессия для бинарного случая

- В случае бинарной классификации нам достаточно одной функции для задания распределения. Для удобства заменим класс 2 на класс 0. Тогда

$$p(\boldsymbol{x};\boldsymbol{\theta}) = P(Y=1|X=\boldsymbol{x}) = \frac{1}{1+exp(\beta_0+\boldsymbol{\beta}^T\boldsymbol{x})} \tag{3}$$

  так как $P(Y=0|X=\boldsymbol{x}) = 1 - p(\boldsymbol{x};\boldsymbol{\theta})$. Здесь $\boldsymbol{\theta} = [\beta_0, \boldsymbol{\beta}]^T$

- Дополним $\boldsymbol{x}$ единицей в начале (смещение). Тогда логарифмическая функция правдоподобия может быть записана как

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{N}\left[y_i \log p(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) + (1-y_i)\log(1-p(\boldsymbol{x}^{(i)};\boldsymbol{\theta}))\right] \\ &= \sum_{i=1}^{N}\left[y_i\boldsymbol{\beta}^T\boldsymbol{x}^{(i)} - \log(1+\exp(\boldsymbol{\beta}^T\boldsymbol{x}^{(i)}))\right]. \end{aligned} \tag{4}$$

- Эта целевая функция выпуклая ($\Longrightarrow$ гарантирована сходимость к глобальному максимуму), но нелинейная (требуется метод Ньютона). Частичное доказательство выпуклости: (1) $\boldsymbol{\beta}^T\boldsymbol{x}^{(i)}$ вогнутая/выпуклая, exp вогнутая и монотонно возрастающая $\Longrightarrow 1+\exp(\boldsymbol{\beta}^T\boldsymbol{x}^{(i)})$ вогнутая; LogSumExp вогнутая, тогда -LogSumExp выпуклая.