



Phystech @ DataScience

Логистическая регрессия



Задача классификации



Классификация

\mathcal{X} — пространство объектов,

Истинное правило классификации:

неизвестная функция $f : \mathcal{X} \rightarrow \mathcal{Y}$.

Пространство \mathcal{X} разбивается
на подпространства (*decision regions*)

$\mathcal{X}_y = \{x \in \mathcal{X} \mid f(x) = y\}$,

границы которых называются
разделяющими поверхностями
(*decision surfaces*).





Часто $\mathcal{X} \subset \mathbb{R}^d$, в т.ч. могут быть категориальные.

Типы классификации

1. Двухклассовая.

$$\mathcal{Y} = \{0, 1\} \text{ или } \mathcal{Y} = \{-1, 1\}.$$

2. Многоклассовая.

$$\mathcal{Y} = \{1, \dots, K\} \text{ или } \mathcal{Y} = \{(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, 0, \dots, 1)\}.$$

Задача классификации:

предложить оценку $\hat{f}: \mathcal{X} \rightarrow \mathcal{Y}$ правила классификации на основе обуч. выборки $(x_1, Y_1), \dots, (x_n, Y_n)$, где $x_i = (x_{i1}, \dots, x_{id}) \in \mathcal{X}$, $Y_i \in \mathcal{Y}$, как можно точнее приближающую неизвестное правило классиф-ции.

Оценку правила классификации чаще будем называть **моделью**.



Вероятностная природа

Часто предполагается случайная принадлежность классу:

функция f при повторении эксперимента один и тот же объект $x \in \mathcal{X}$ может отнести как одному классу, так и к другому.

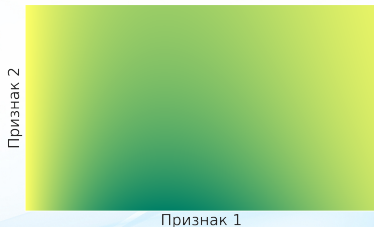
\Rightarrow имеет смысл **предсказывать вероятность** $P_x(Y = y)$
принадлежности объекта x каждому из классов.

Точечная оценка: $\arg \max_{y \in \mathcal{Y}} P_x(Y = y)$

Если классы неравнозначны:

$\arg \max_{y \in \mathcal{Y}} [w_y P_x(Y = y)],$

w_y — приоритетность класса



Примеры:

1. $P(Y = 0 \mid X = x_2) = 0.95, \quad P(Y = 1 \mid X = x_2) = 0.05$
Уверенное предсказание в пользу класса 0.
2. $P(Y = 0 \mid X = x_1) = 0.55, \quad P(Y = 1 \mid X = x_1) = 0.45$
Модель не уверена в предсказании.



Линейные модели

$y(x) = \theta^T x$ — линейная модель регрессии.

Линейная модель в классификации:

Разделяющая поверхность — линейная гиперплоскость в пр-ве \mathcal{X} .

В многоклассовом случае — при дополнении до гиперплоскости.

Например, при $\mathcal{Y} = \{-1, 1\}$ линейна модель $y(x) = \text{sign}(\theta^T x)$.



Замечание.

Исходное пр-во признаков может быть предварительно преобразовано с помощью нелинейных функций, в частности можно включить константный признак. В таком случае разделяющая поверхность лин. классификатора *не будет* линейной в исходном пространстве.



Логистическая регрессия



Логистическая регрессия

Пространство объектов $x \in \mathcal{X} \subset \mathbb{R}^d$.

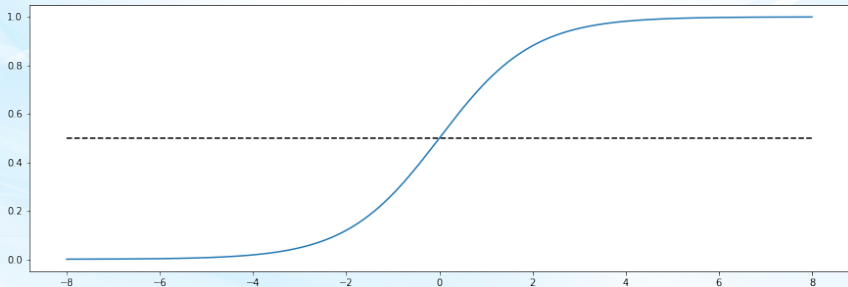
Множество классов $\mathcal{Y} = \{0, 1\}$.

Класс объекта x имеет распределение $Bern(\mu(x)) \in [0, 1]$.

Предположение:

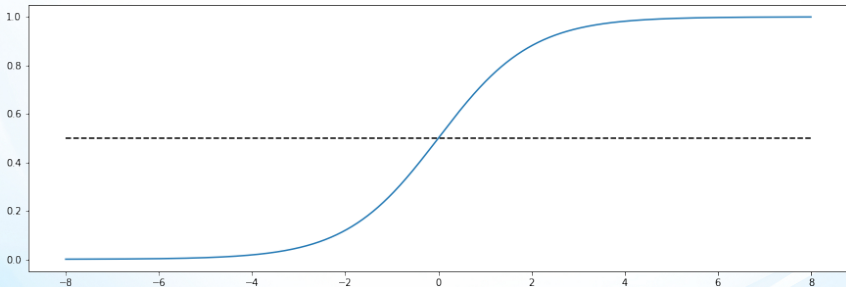
$$\mu_{\theta}(x) = \sigma(\theta^T x),$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ — логистическая сигмоида.





Логистическая регрессия



Разделяющая поверхность $\{\mu_{\theta}(x) = 1/2\} = \{\theta^T x = 0\}$ линейна, а значит логистическая регрессия является линейным классификатором.

Чем больше значение $\theta^T x$, тем более вероятен класс 1.



Свойства

Свойства:

1. $\sigma(-z) = 1 - \sigma(z)$. При $z = \theta^T x$ это — вероятность класса 0;
2. Обратная функция $z(s) = \ln \frac{s}{1-s}$ — **логит-функция**;
3. $\frac{d\sigma}{dz} = \sigma(z)(1 - \sigma(z))$.



Обучение

Пусть дана обучающая выборка $(x_1, Y_1), \dots, (x_n, Y_n)$, где $x_i = (x_{i1}, \dots, x_{id}) \in \mathcal{X}$ и случайный класс $Y_i \sim \text{Bern}(\mu_\theta(x_i))$.

Функция правдоподобия:

$$L_Y(\theta) = \prod_{i=1}^n \mu_\theta(x_i)^{Y_i} (1 - \mu_\theta(x_i))^{1-Y_i}$$

Будем ее **максимизировать** численно с помощью *градиентного подъема*. Это лучше выполнять с помощью логарифмической функции правдоподобия

$$\ell_Y(\theta) = \log L_Y(\theta) = \sum_{i=1}^n [Y_i \log \sigma(\theta^T x_i) + (1 - Y_i) \log (1 - \sigma(\theta^T x_i))]$$



Градиентный спуск

Пусть задача оптимизации имеет вид

$$f(\theta) \rightarrow \min_{\theta},$$

где $f(\theta)$ — дифференцируемая функция; $\nabla_{\theta} f(\theta)$ — градиент $f(\theta)$.

Итеративные методы оптимизации последовательно приближают текущее значение параметра θ к оптимальному θ^* .

Наблюдение (матан 1 курс): В малой окрестности точки направление скорейшего роста функции — ее **градиент** $\nabla_{\theta} f(\theta)$, направление скорейшего убывания — **антиградиент** $-\nabla_{\theta} f(\theta)$.



Градиентный спуск

Итерация:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} f(\theta_t).$$

Антиградиент вычитается с заданным малым коэффициентом η , который часто называют коэффициентом скорости обучения или **learning rate**.

Подбор η осуществляется пользователем.

Критерии останова:

1. Лимит на число итераций.
2. Early stopping. Не происходит уменьшения $f(\theta)$ в течение какого-то зафиксированного числа шагов.
3. Ограничение на норму невязки.

Норма невязки: $\|f(\theta_{t+1}) - f(\theta_t)\|$ становится ниже порога.



Максимизация $\ell_Y(\theta)$

$$\ell_Y(\theta) = \log L_Y(\theta) = \sum_{i=1}^n [Y_i \log \sigma(\theta^T x_i) + (1 - Y_i) \log (1 - \sigma(\theta^T x_i))]$$

Ее производная равна

$$\frac{\partial \ell_Y(\theta)}{\partial \theta} = \sum_{i=1}^n [Y_i - \sigma(\theta^T x_i)] x_i.$$

Получаем формулу градиентного подъема

$$\theta_{t+1} = \theta_t + \eta \sum_{i=1}^n [Y_i - \sigma(\theta_t^T x_i)] x_i$$



Максимизация $\ell_Y(\theta)$

Можно также проводить **стохастический** градиентный подъем, выбирая случайный индекс i

$$\theta_{t+1} = \theta_t + \eta [Y_i - \sigma(\theta_t^T x_i)] x_i$$

Вектор параметров сдвигается вдоль направления выбранного объекта x_i настолько, насколько модель ошибается на этом объекте.

Обозначения:

- ▶ Градиентный спуск — Gradient descent — GD;
- ▶ Стохастический градиентный спуск — Stochastic GD — SGD.



Переобучение модели

Пусть

1. Классы линейно разделимы;
2. Среди признаков есть константа;
3. $\theta : \{\theta^T x = 0\}$ в точности разделяет два класса.

Тогда $\forall c > 0$ $\{c\theta^T x = 0\}$ в точности разделяет два класса.

Но $L_Y(c\theta) = \prod_{i=1}^n \sigma(c\theta^T x_i)^{Y_i} (1 - \sigma(c\theta^T x_i))^{1-Y_i} \rightarrow 1$ при $c \rightarrow \infty$.

При конечном θ максимум функции правдоподобия **не достигается**.

Проблемы

1. Предсказания вероятностей классов близки к 0 или 1, что не информативно при решении реальных задач.
2. Может быть выбрана произвольная гиперплоскость, в точности разделяющая два класса. При разных запусках один и тот же объект между классами может относиться с вероятностью 1 как к одному классу, так и к другому.





Проблемы

Если классы близки к линейно разделимым, оптимальное θ хотя не бесконечно, но может быть достаточно большим, что также влечет подобные проблемы.

В качестве решения проблемы обычно используют регуляризацию.



ВСЁ!