Phystech@DataScience

Блок 2: линейные модели

Регуляризация

Классификация

Задача классификации

Логистическая регрессия

Градиентный спуск

Проблема: мультиколлинеарность

Мультиколлинеарность — наличие большого числа *линейно-зависимых* признаков.

Пример: среди признаков много таких, которые связаны с размером котика. Они все зависят друг от друга и несут *избыточную* информацию.

В модели линейной регрессии:

если для $\varepsilon = y - \widehat{y}$ матрица ковариаций $\Sigma = \sigma^2 I_n$, то $D\widehat{\theta} = \sigma^2 (X^T X)^{-1}$.

Если признаки мультиколлинеарны, то $X^T X$ почти вырождена и дисперсия огромна.

Решение: регуляризация.

Ridge-регрессия

Задача МНК:

$$||Y - X\theta||_2 \to \min_\theta$$

Задача Ridge-регрессии:

$$||Y - X\theta||_2 + \lambda||\theta||_2 \to \min_\theta, \lambda > 0$$

Ограничиваем коэффициенты, не позволяем им «разбрасываться».

Замечание. Предварительно необходимо

- центрировать отклик $Y := Y - \overline{Y}$ *или не накладывать ограничение на коэффициент при константе;*
- стандартизовать признаки — вычесть среднее, поделить на корень из дисперсии.

Решение задачи

Решением задачи является

$$\widehat{\theta} = (X^T X + \lambda I_d)^{-1} X^T Y$$

За счет добавки λI_d матрица стала менее вырожденной.

Вопрос в бот: посмотрите на формулу и скажите, почему признаки ОБЯЗАТЕЛЬНО надо стандартизировать?

У них могут быть разные размерности и масштаб!

Свойства

- $\lambda = 0 \implies$ МНК; $\lambda = \infty \implies \widehat{\theta} = 0$;
- При $\lambda \geq 0$ решение $\exists!$;
- $\widehat{\theta}$ может быть найдёна *итеративными* методами;
- Пусть $E\varepsilon = 0$. Оценка смещенная $E\hat{\theta} = (X^T X + \lambda I_d)^{-1} X^T X \theta$;
- Пусть $D\varepsilon = \sigma^2 I_n$. Дисперсия уменьшилась:
  $D\hat{\theta} = \sigma^2 (X^T X + \lambda I_d)^{-1} X^T X (X^T X + \lambda I_d)^{-1}$

Lasso-регрессия

Задача МНК:

$$||Y - X\theta||_2 \to \min_\theta$$

Задача Lasso-регрессии:

$$||Y - X\theta||_2 + \lambda||\theta||_1 \to \min_\theta, \lambda > 0,$$

$$||\theta||_1 = |\theta_1| + |\theta_2| + \ldots + |\theta_d|.$$

Свойства

- Решается **только** *итеративными* методами;
- Lasso-регрессия зануляет коэффициенты с ростом λ, может использоваться для отбора признаков.

Регуляризация

Классификация

Задача классификации

Логистическая регрессия

Градиентный спуск

Классификация

$\mathscr{X}$ — пространство объектов,

$\mathscr{Y}$ — конечное множество классов.

Истинное правило классификации:

неизвестная функция $f : \mathscr{X} \to \mathscr{Y}$.

Пространство $\mathscr{X}$ разбивается на *подпространства (decision regions)* $\mathscr{X}_y = \{x \in \mathscr{X} \mid f(x) = y\}$, границы которых называются *разделяющими поверхностями (decision surfaces)*.

Классификация

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

Типы классификации

1. *Двухклассовая*.
   $\mathscr{Y} = \{0, 1\}$ или $\mathscr{Y} = \{-1, 1\}$.

2. *Многоклассовая*.
   $\mathscr{Y} = \{1, ..., K\}$ или $\mathscr{Y} = \{(1, 0, ..., 0), (0, 1, ..., 0), ..., (0, 0, ..., 1)\}$.

Задача классификации:

предложить оценку $\widehat{f} : \mathscr{X} \to \mathscr{Y}$ правила классификации на основе обуч. выборки $(x_1, Y_1), ..., (x_n, Y_n)$, где $x_i = (x_{i1}, ..., x_{id}) \in \mathscr{X}$, $Y_i \in \mathscr{Y}$, 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 может отнести один и тот же объект $x \in \mathscr{X}$ как одному классу, так и к другому.

$\Longrightarrow$ имеет смысл **предсказывать вероятность** $\mathrm{P}_x(Y = y)$ принадлежности объекта x каждому из классов.

Точечная оценка: $\arg\max\limits_{y \in \mathscr{Y}} \mathrm{P}_x(Y = y)$

Если классы неравнозначны:

$\arg\max\limits_{y \in \mathscr{Y}} [w_y \, \mathrm{P}_x(Y = y)]$,

w_y — приоритетность класса

Примеры:

1. $\mathrm{P}(Y = 0 \mid X = x_2) = 0.95, \quad \mathrm{P}(Y = 1 \mid X = x_2) = 0.05$
   Уверенное предсказание в пользу класса 0.
2. $\mathrm{P}(Y = 0 \mid X = x_1) = 0.55, \quad \mathrm{P}(Y = 1 \mid X = x_1) = 0.45$
   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 $\mathscr{X}$.

В многоклассовом случае — при дополнении до гиперплоскости.

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

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.

Регуляризация

Классификация

Задача классификации

Логистическая регрессия

Градиентный спуск

Логистическая регрессия

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

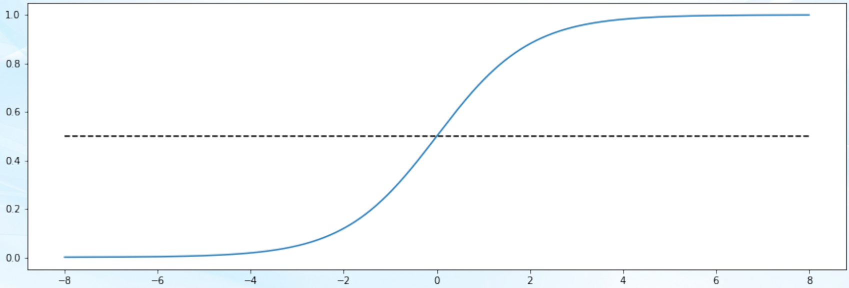
Множество классов $\mathscr{Y} = \{0, 1\}$.

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

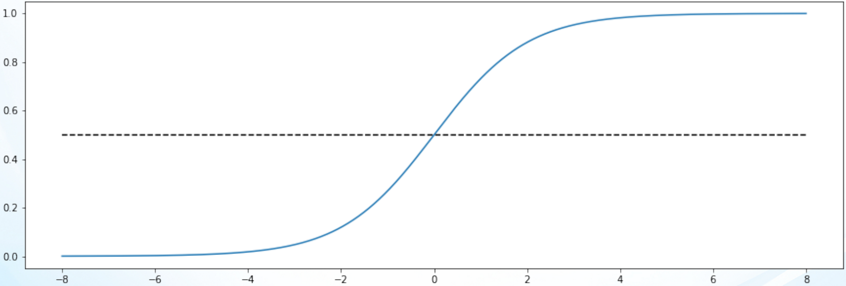
Предположение:

$$p_\theta(x) = \sigma(\theta^T x),$$

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



Логистическая регрессия



Разделяющая поверхность $\{p_\theta(x) = 1/2\} = \{\theta^T x = 0\}$ 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 $x_i = (x_{i1}, \dots, x_{id}) \in \mathscr{X}$ и случайный класс $Y_i \sim Bern(p_\theta(x_i))$.

Функция правдоподобия:

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

Что это за зверь? Вспомним формулу Бернулли: $P_n^k = C_n^k p^k (1-p)^{n-k}$

Пусть среди Y_i ровно k единиц. Если выборка уже получена, то индексы от 1 до n фиксированы, и C_n^k не нужно.

Функция правдободобия отражает вероятность получить такую реализацию!

Подробности уже этой весной, не пропустите!

Обучение

$(x_1, Y_1), ..., (x_n, Y_n)$ – *реализация* обучающей выборки,

$x_i = (x_{i1}, ..., x_{id}) \in \mathscr{X}$

$Y_i \sim Bern(p_\theta(x_i))$ – случайный класс.

Функция правдоподобия для полученных чисел:

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

Фиксируем какое-нибудь θ.

Чем больше $L_Y(\theta)$, тем «правдободобнее» это самое θ.

Будем максимизировать $L_Y(\theta)$ численно с помощью *градиентного подъема*.

Регуляризация

Классификация

Задача классификации

Логистическая регрессия

Градиентный спуск

Градиентный спуск

Пусть *задача оптимизации* имеет вид

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

где $f(\theta)$ — дифференцируемая функция;

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 $\ell_Y(\theta)$

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

Ее производная равна

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

Получаем формулу градиентного подъема:

$$\theta_{t+1} = \theta_t + \eta \underbrace{\sum_{i=1}^{n} \left[Y_i - \sigma(\theta_t^T x_i)\right] x_i}_{\nabla_\theta f(\theta_t)}$$

Максимизация $\ell_Y(\theta)$

Можно также проводить **стохастический** градиентный подъем (спуск), выбирая случайный индекс i:

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

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.

Компромисс между ними – Batch gradient descent (BGD), когда градиент на очередном шаге считается по *подмножеству* выборки (т.е. по батчу).

Переобучение модели

Пусть

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

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

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

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

Проблемы

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

В качестве решения проблемы обычно используют *регуляризацию*.


ВСЁ!