Линейные модели

Ушаков Роман

March, 2018


MACHINE
LEARNING

Сегодня в программе

1. Многомерная регрессия
   - Метод наименьших квадратов
   - Регуляризация
   - Градиентный спуск
2. Логистическая регрессия
   - Задача классификации
   - Аппроксимации эмпирического риска
   - LogLoss и правдоподобие

План

1. Многомерная регрессия
   - Метод наименьших квадратов
   - Регуляризация
   - Градиентный спуск
2. Логистическая регрессия
   - Задача классификации
   - Аппроксимации эмпирического риска
   - LogLoss и правдоподобие

Метод наименьших квадратов

- X — объекты из $\mathbb{R}^n$, Y — ответы из $\mathbb{R}$

Метод наименьших квадратов

- X — объекты из $\mathbb{R}^n$, Y — ответы из $\mathbb{R}$
- $X^l = (x_i, y_i)_{i=1}^l$ — обучающая выборка

Метод наименьших квадратов

- X — объекты из $\mathbb{R}^n$, Y — ответы из $\mathbb{R}$
- $X^l = (x_i, y_i)_{i=1}^l$ — обучающая выборка
- $f_j(x_i)$ — j-й признак на i-м объекте.

Метод наименьших квадратов

- X — объекты из $\mathbb{R}^n$, Y — ответы из $\mathbb{R}$
- $X^l = (x_i, y_i)_{i=1}^l$ — обучающая выборка
- $f_j(x_i)$ — j-й признак на i-м объекте.
- $y_i = y(x_i)$, $y : X \to Y$ — истинная зависимость.

Метод наименьших квадратов

- X — объекты из $\mathbb{R}^n$, Y — ответы из $\mathbb{R}$
- $X^l = (x_i, y_i)_{i=1}^l$ — обучающая выборка
- $f_j(x_i)$ — j-й признак на i-м объекте.
- $y_i = y(x_i)$, $y : X \to Y$ — истинная зависимость.
- Задача: восстановить $y(x)$ при помощи линейной модели и квадратичной функции потерь:

$$Q(X^l, w) = \sum_{i=1}^{l} (\langle w, f(x_i) \rangle - y_i)^2 \to \min_w$$

Метод наименьших квадратов

- Та же самая задача в матричном виде:

$$F \in \mathbb{R}^{n \times m}$$

$$F = \begin{pmatrix} f_1(x_1) & f_2(x_1) & \dots & f_m(x_1) \\ f_1(x_2) & f_2(x_2) & \dots & f_m(x_2) \\ \dots & \dots & \dots & \dots \\ f_1(x_l) & f_2(x_l) & \dots & f_m(x_l) \end{pmatrix}$$

$$||Fw - y||^2 \to \min_w$$

Метод наименьших квадратов

- Можно дифференцировать и матричные выражения

Метод наименьших квадратов

- Можно дифференцировать и матричные выражения
- Условия минимума:

$$\frac{\partial Q}{\partial w} = 2F^T(Fw - y) = 0$$

$$F^T F w = F^T y$$

Метод наименьших квадратов

- Можно дифференцировать и матричные выражения
- Условия минимума:

$$\frac{\partial Q}{\partial w} = 2F^T(Fw - y) = 0$$

$$F^T F w = F^T y$$

- Оптимальные веса:

$$w_{opt} = (F^T F)^{-1} F^T y$$

План

1. Многомерная регрессия
   - Метод наименьших квадратов
   - Регуляризация
   - Градиентный спуск
2. Логистическая регрессия
   - Задача классификации
   - Аппроксимации эмпирического риска
   - LogLoss и правдоподобие

Регуляризация

Чем плох МНК?

- Обращение матрицы $O(n^3)$ операций

Регуляризация

Чем плох МНК?

- Обращение матрицы $O(n^3)$ операций
- Обращение — неустойчивая вычислительная операция

Регуляризация

Чем плох МНК?

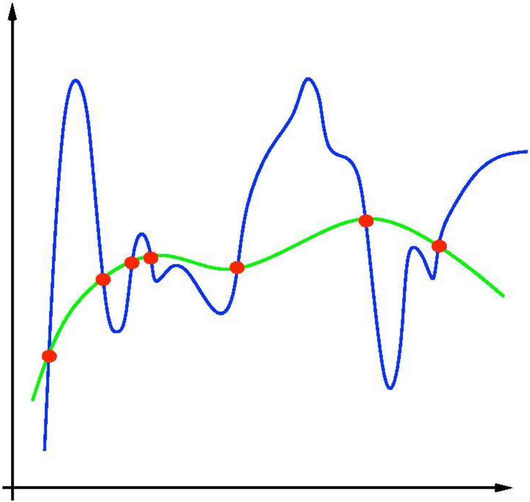
- Обращение матрицы $O(n^3)$ операций
- Обращение — неустойчивая вычислительная операция
- Если признаки скореллированы, то w_{opt} не робастно

Регуляризация

Чем плох МНК?

- Обращение матрицы $O(n^3)$ операций
- Обращение — неустойчивая вычислительная операция
- Если признаки скореллированы, то w_{opt} не робастно
- Признак переобучения — большие веса.

Регуляризация



Регуляризация

Выход есть: добавим штраф в модель!

Регуляризация

Новый функционал ошибки:

- L_1 регуляризация

$$Q(X^l, w) = \sum_{i=1}^{l} (\langle w, f(x_i) \rangle - y_i)^2 + \lambda \sum_{j=1}^{m} |w_j|$$

Регуляризация

Новый функционал ошибки:

- L_1 регуляризация

$$Q(X^l, w) = \sum_{i=1}^{l} (\langle w, f(x_i) \rangle - y_i)^2 + \lambda \sum_{j=1}^{m} |w_j|$$

- L_2 регуляризация

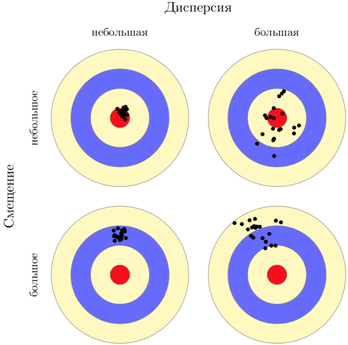
$$Q(X^l, w) = \sum_{i=1}^{l} (\langle w, f(x_i) \rangle - y_i)^2 + \lambda \sum_{j=1}^{m} w_j^2$$

Регуляризация

- Можно аналитически, что L_1 регуляризация позволяет занулять веса, обладающие низкой предсказательной способностью.
- Также можно показать, что:

$$\mathbb{E}Q = bias^2 + var$$

- *bias variance* threshold



После этой фичи скор будет расти на глазах

Надо всего лишь...

План

1. Многомерная регрессия
   - Метод наименьших квадратов
   - Регуляризация
   - Градиентный спуск

2. Логистическая регрессия
   - Задача классификации
   - Аппроксимации эмпирического риска
   - LogLoss и правдоподобие

Методы оптимизации

- Задача регрессии с L_2-регуляризацией имеет аналитическое решение:

$$w_{opt} = (F^T F + \lambda I)^{-1} F^T y$$

Но снова приходится обращаться матрицы

Методы оптимизации

- Задача регрессии с L_2-регуляризацией имеет аналитическое решение:

$$w_{opt} = (F^T F + \lambda I)^{-1} F^T y$$

Но снова приходится обращаться матрицы

- Задача L_1-регуляризации в общем случае не имеет аналитического решения

Методы оптимизации

- Задача регрессии с L_2-регуляризацией имеет аналитическое решение:
$$w_{opt} = (F^T F + \lambda I)^{-1} F^T y$$
Но снова приходится обращаться матрицы
- Задача L_1-регуляризации в общем случае не имеет аналитического решения
- Как жить и что делать?

Методы оптимизации

- Задача регрессии с L_2-регуляризацией имеет аналитическое решение:

$$w_{opt} = (F^T F + \lambda I)^{-1} F^T y$$

  Но снова приходится обращаться матрицы
- Задача L_1-регуляризации в общем случае не имеет аналитического решения
- Как жить и что делать?
- Оптимизировать!
  “Часик в радость, чифир в сладость. Градиенту ходу, лоссу в нуль уходу“.

Немного о градиентном спуске

Что такое градиент? $u = u(x, y, z)$:

$$\nabla u = \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \\ \frac{\partial u}{\partial z} \end{pmatrix}$$

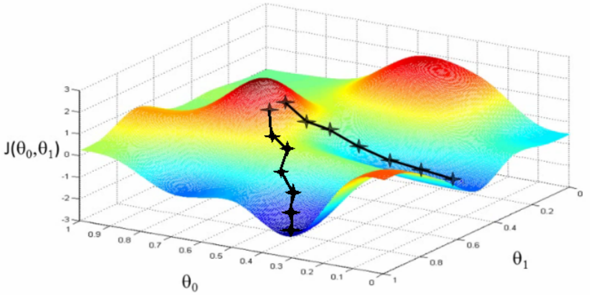
Пр.:

$$u(x, y, z) = x^2 + 2xy + y^2 + z^2$$

$$\nabla u = \begin{pmatrix} 2x + 2y \\ 2x + 2y \\ 2z \end{pmatrix}$$

Немного о градиентном спуске

- Градиент — направление наискорейшего роста функции, антиградиент — убывания
- Построим итеративную процедуру поиска оптимальных весов w:



Немного о градиентном спуске

- Инициализируем вектор весов w начальным значением w_0
- Пока $||\nabla Q(X^l, w)|| > tol$:

$$w_{k+1} = w_k - \alpha \cdot \nabla Q(X^l, w_k)$$

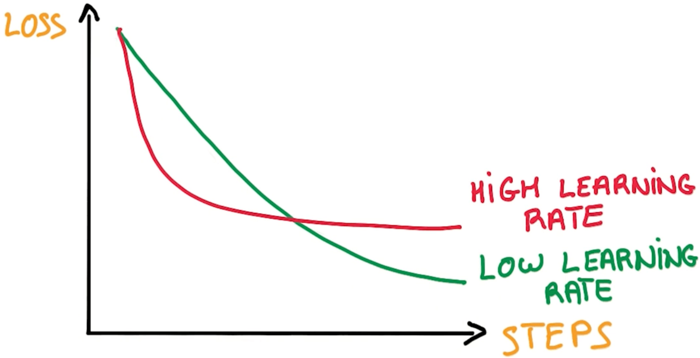
- Чтобы воспользоваться методом GD нужно уметь вычислять $\nabla Q(X^l, w)$

Немного о градиентном спуске

- Шаг GD:

$$w_{k+1} = w_k - \alpha \cdot \nabla Q(X^l, w_k)$$

  Параметр α — learning rate
- Как влияет α на скорость сходимости и качество?



Как сходится GD?

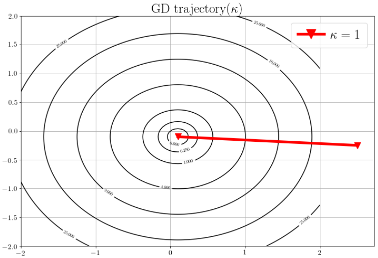


Рис.: В нормированном пространстве

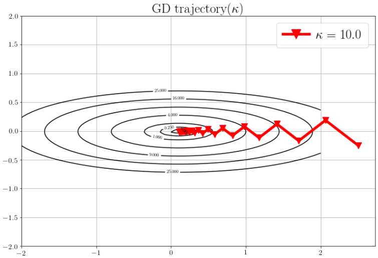
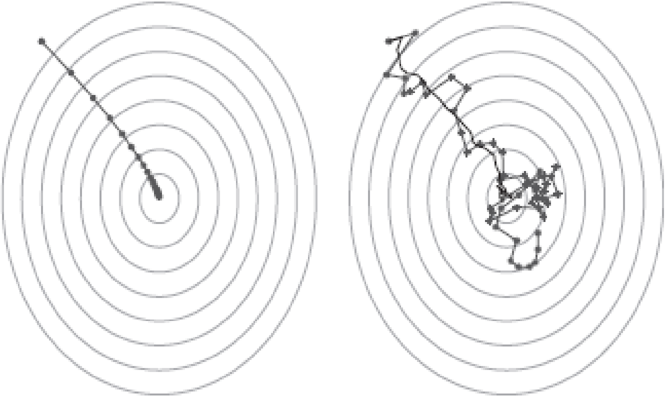


Рис.: В **не**нормированном пространстве

Немного о градиентном спуске

Ускоряем GD: берем случайную подвыборку данных и по ней считаем градиент



План

Задача классификации

- Обучающая выборка X^l, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$
- Модель классификации — линейная:

$$y_i = sign \langle w, x_i \rangle$$

- Функция потерь:

$$Q(X^l, w) = [\langle w, x_i \rangle y_i < 0] \leq \hat{L}(X^l, w),$$

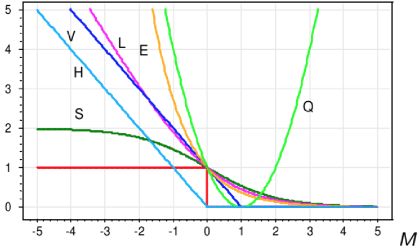
где $\hat{L}(X^l, w)$ — некоторая аппроксимация $Q(X^l, w)$.

План

Аппроксимации эмпирического риска

$M_i(w) = \langle w, x_i \rangle$ — margin (отступ).

Часто используемые непрерывные функции потерь $\mathscr{L}(M)$:



$V(M) = (1-M)_+$ — кусочно-линейная (SVM);
$H(M) = (-M)_+$ — кусочно-линейная (Hebb's rule);
$L(M) = \log_2(1+e^{-M})$ — логарифмическая (LR);
$Q(M) = (1-M)^2$ — квадратичная (FLD);
$S(M) = 2(1+e^{M})^{-1}$ — сигмоидная (ANN);
$E(M) = e^{-M}$ — экспоненциальная (AdaBoost);
$[M < 0]$ — пороговая функция потерь.

План

1. Многомерная регрессия
   - Метод наименьших квадратов
   - Регуляризация
   - Градиентный спуск

2. Логистическая регрессия
   - Задача классификации
   - Аппроксимации эмпирического риска
   - LogLoss и правдоподобие

LogLoss и правдоподобие

Рассмотрим какое-либо вероятностное распределение:

- $p(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-\frac{(x-\theta)^2}{2\sigma^2}}$ — нормальное
- $p(x) = \lambda \cdot e^{-\lambda x}$ — экспоненциальное
- $p(x) = \frac{1}{b-a}$, если $x \in [a; b]$, иначе 0 — равномерное

Пусть $X_1, \ldots, X_n$ — реализации из распределения. Насколько типична выборка $X_{i=1}^n$ для данного распределения?

LogLoss и правдоподобие

Ответ: нужно посчитать правдоподобие

$$L(X^n) = \prod_{i=1}^{n} p(X_i)$$

$$\ln L(X^n) = \sum_{i=1} p(X_i)$$

Чем больше значение $\ln L(X^n)$, тем лучше данные подходят под распределение.

Утверждение: минимизация аппроксимированного риска $L(X^{\hat{n}}, w) \Leftrightarrow$ максимизации правдоподобия $L(X^n, w)$.

LogLoss

Пусть теперь классы $Y_i \in \{0; 1\}$. Тогда можно показать, что:

$$LogLoss = \ln L(X^n, w) = \sum_{i=1}^{n} y_i \ln p_i + (1 - y_i) \ln(1 - p_i)$$

Переход от $M(X_i) = \langle w, x_i \rangle$ к $p(X_i)$ осуществляется через сигмоиду:

$$p(X_i) = \sigma(M) := \frac{1}{1 + e^{-M}}$$

Важное свойство:

$$\sigma(M) + \sigma(-M) = 1$$

Оптимизировать LogLoss можно такими же методами, как и квадратичную функцию потерь.

Итог

- МНК хорош, но не слишком
- Регуляризация уменьшает переобучение
- Градиентный спуск помогает оптимизировать
- LogLoss = логарифм правдоподобия