Линейная регрессия

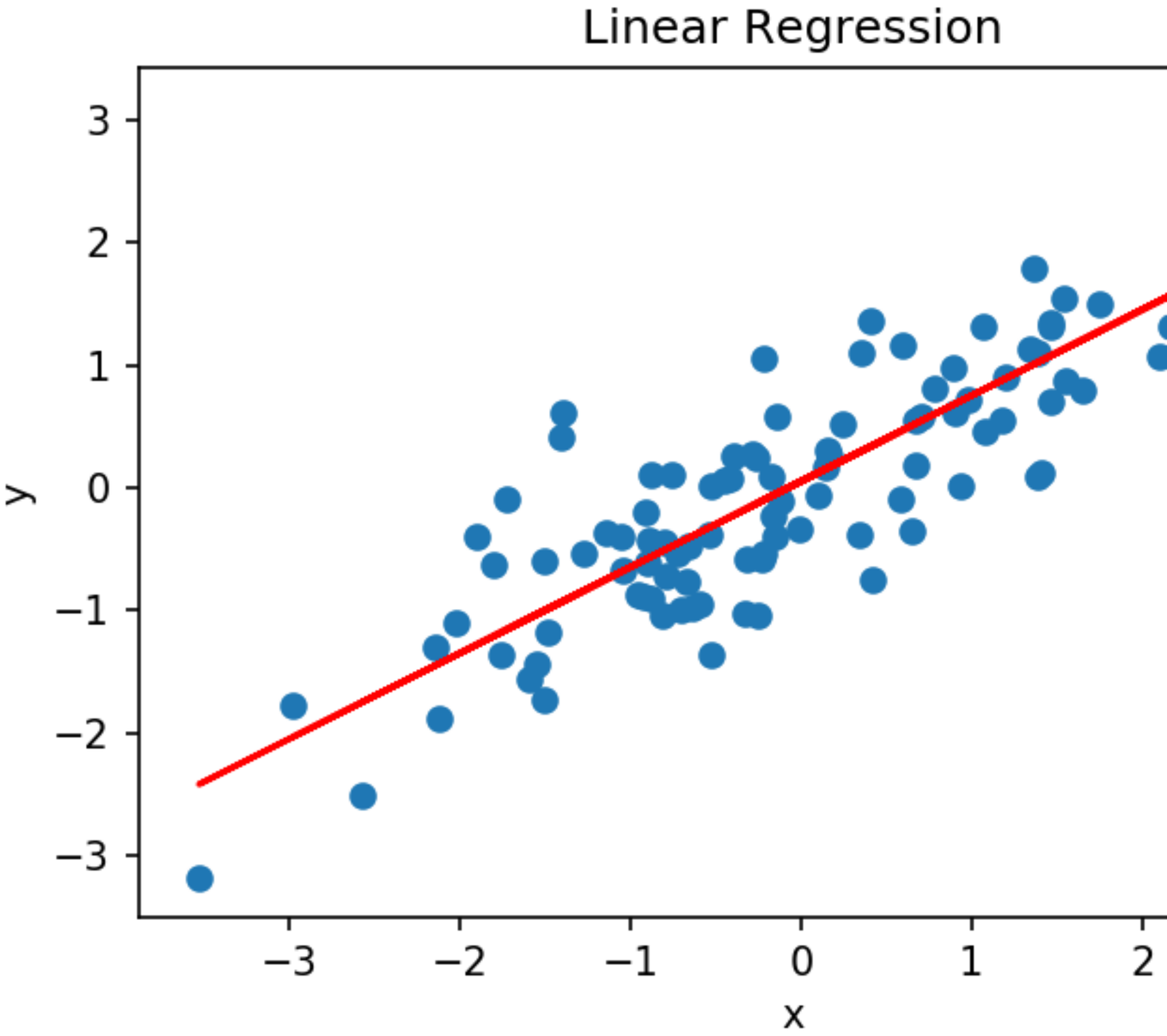
Математика алгоритма

Лямбда, МАИ

2021

Постановка задачи

- ДАНО: набор значений (y) в некоторых точках (x) - таблично заданная функция
- ТРЕБУЕТСЯ: найти такую функцию, которая наиболее точнс приблизит табличную функцию
- БОЛЕЕ СТРОГО: из семейства функций выбрать ту, которая обладает необходимым свойством



Подходы к решению

Не регрессия

- **Интерполяция** — способ выбрать из семейства функций ту, которая проходит через заданные точки. Часто функцию затем используют для вычисления в промежуточных точках. Например, мы вручную задаем цвет нескольким точкам и хотим чтобы цвета остальных точек образовали плавные переходы между заданными.
- **Аппроксимация** — способ выбрать из семейства «простых» функций приближение для «сложной» функции на отрезке, при этом ошибка не должна превышать определенного предела. Аппроксимацию используют, когда нужно получить функцию, похожую на данную, но более удобную для вычислений и манипуляций (дифференцирования, интегрирования и т.п).

Подходы к решению

Регрессия

Регрессия — способ выбрать из семейства функций ту, которая минимизирует функцию потерь (*loss function,* или *cost function*). Последняя характеризует насколько сильно пробная функция отклоняется от значений в заданных точках.

В частности, линейная регрессия выбирает функцию f из линейной комбинации наперед заданных базисных функций f_i

$$f = \sum_i w_i f_i$$

Метод наименьших квадратов

Простейший двумерный случай

- Пусть нам даны точки на плоскости $\{(x_1, y_1), \ldots (x_N, y_N)\}$ и мы ищем такую функцию $f(x) = a + bx$,чтобы ее график находился ближе всего к данным точкам. Таким образом, наш базис состоит из константной функции и линейной $(1,x)$.
- Расстояние можно измерять по-разному. Простейший - абсолютное значение разниц $|f(x_i) - y_i|$, тогда функция потерь - $\sum_i |f(x_i) - y_i|$. Это - *Least Absolute Distance* (LAD) регрессия.

Метод наименьших квадратов

Простейший двумерный случай

- Более популярная функция потерь - сумма квадратов отклонений регрессанта от модели - *Sum of Squared Errors* (SSE)

$$SSE(a, b) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

- МНК - линейная регрессия с функцией потерь SSE.
- Удобно, потому что производная квадратичной функции - линейная функция.
- Простейший способ минимизировать *SSE(a, b)* - вычислить производные по a и b, приравнять их к нулю и решить систему линейных уравнений.

$$\frac{\partial}{\partial a}\mathrm{SSE}(a,b) = -2\sum_{i=1}^{N}(y_i - a - bx_i),$$

$$\frac{\partial}{\partial b}\mathrm{SSE}(a,b) = -2\sum_{i=1}^{N}(y_i - a - bx_i)x_i.$$

Приравняем нулю:

$$0 = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i),$$

$$0 = -2\sum_{i=1}^{N}(y_i - \hat{a} - \hat{b}x_i)x_i,$$

И легко решим, если правильно запишем формулы:

$$\hat{a} = \frac{\sum_i y_i}{N} - \hat{b}\frac{\sum_i x_i}{N},$$

$$\hat{b} = \frac{\frac{\sum_i x_i y_i}{N} - \frac{\sum_i x_i \sum_i y_i}{N^2}}{\frac{\sum_i x_i^2}{N} - \left(\frac{\sum_i x_i^2}{N}\right)^2}.$$

Статистика

Важные понятия

Полученные формулы можно компактно записать с помощью статистических эстиматоров: среднего $\langle\cdot\rangle$, вариации $\sigma_{\cdot}$ (стандартного отклонения), ковариации $\sigma(\cdot,\cdot)$ и корреляции $\rho(\cdot,\cdot)$

$$\hat{a} = \langle y\rangle - \hat{b}\langle x\rangle,$$

$$\hat{b} = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\langle x^2\rangle - \langle x\rangle^2}.$$

Статистика

Важные понятия

Перепишем b как:

$$\hat{b} = \frac{\sigma(x, y)}{\sigma_x^2}$$

Статистика

Важные понятия

Введем коэффициент корреляции Пирсона:

$$\rho(x, y) = \frac{\sigma(x, y)}{\sigma_x \sigma_y}$$

$$\hat{b} = \rho(x, y) \frac{\sigma_y}{\sigma_x}.$$

Статистика

Важные понятия

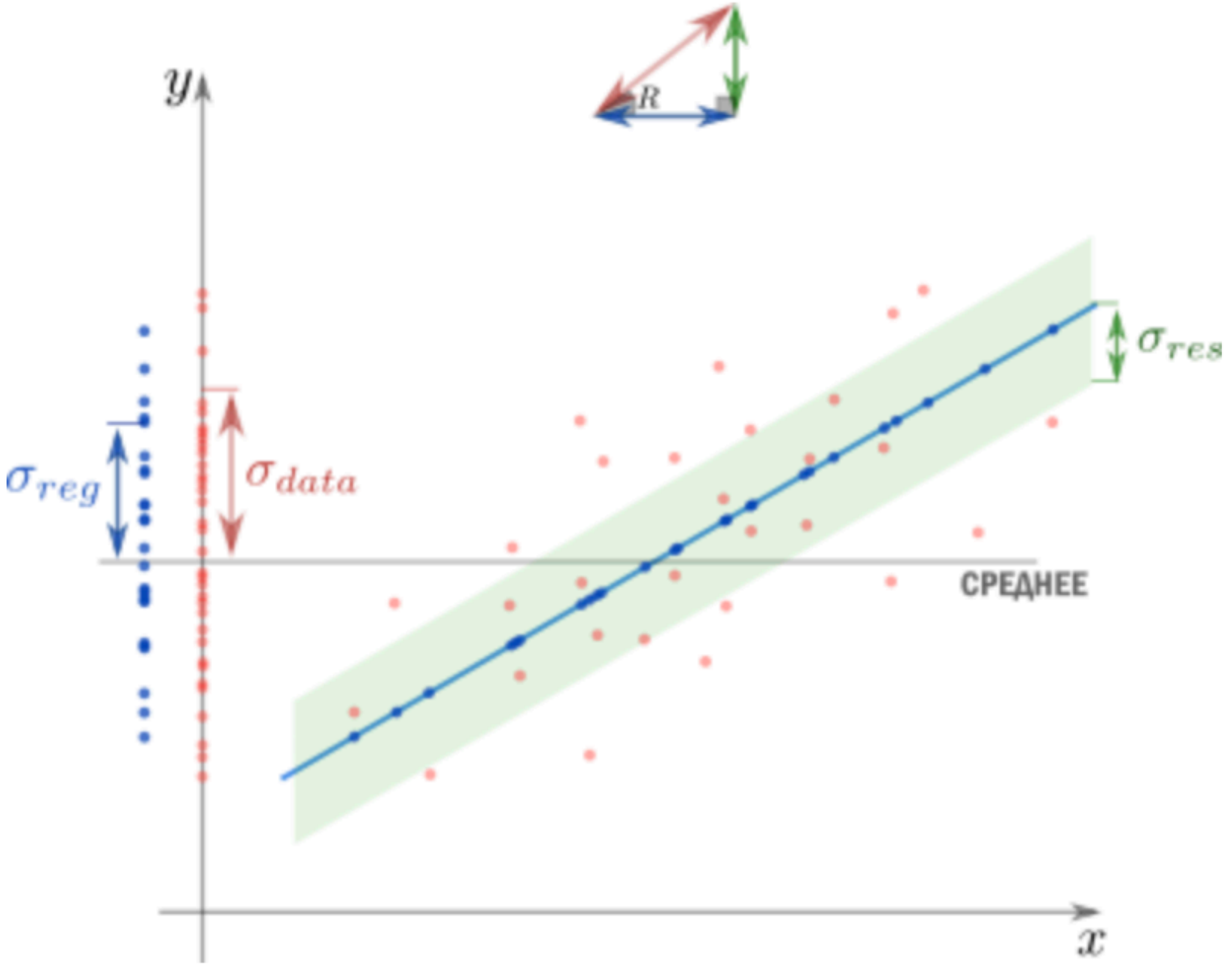
Квадрат коэффициента корреляции Пирсона - коэффициент детерминации

$$R = \rho^2$$

$$\mathrm{Var}_{data} = \frac{1}{N}\sum_i (y_i - \langle y \rangle)^2,$$

$$\mathrm{Var}_{res} = \frac{1}{N}\sum_i (y_i - \text{модель}(x_i))^2,$$

$$\mathrm{Var}_{reg} = \frac{1}{N}\sum_i (\text{модель}(x_i) - \langle y \rangle)^2.$$



$$\mathrm{Var}_{data} = \mathrm{Var}_{res} + \mathrm{Var}_{reg}.$$

$$\sigma^2_{data} = \sigma^2_{res} + \sigma^2_{reg}.$$

Статистика

Важные понятия

$$R^2 = \frac{\mathrm{Var}_{data} - \mathrm{Var}_{res}}{\mathrm{Var}_{data}} = 1 - \frac{\mathrm{Var}_{res}}{\mathrm{Var}_{data}}$$

Так мы получили *долю объясненной вариации*:

$$R^2 = \frac{\mathrm{Var}_{reg}}{\mathrm{Var}_{data}}.$$

Мультилинейная регрессия

А что, собственно, делать

До сих пор мы рассматривали задачу регрессии для одного скалярного признака x, однако обычно регрессор — это n-мерный вектор $\mathbf{x}$. Другими словами, для каждого измерения мы регистрируем n фичей, объединяя их в вектор. В этом случае логично принять модель с $n+1$ независимыми базисными функциями векторного аргумента — n степеней свободы соответствуют n фичам и еще одна — регрессанту y. Простейший выбор — линейные базисные функции $(1, x_1, \ldots, x_n)$. При $n = 1$ получим уже знакомый нам базис $(1, x)$.

Мультилинейная регрессия

А что, собственно, делать

Итак, мы хотим найти такой набор коэффициентов (вектор) **w,** что

$$\sum_{j=0}^{n} w_j x_j^{(i)} = \mathbf{w}^T x^{(i)} \simeq y_i, \quad i = 1, \ldots, N$$

Знак "$\simeq$" означает, что мы ищем решение, которое минимизирует сумму квадратов ошибок

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^\top \mathbf{x}^{(i)} \right)^2$$

Мультилинейная регрессия

А что, собственно, делать

Итого, мы должны решить уравнение:

$$X\mathbf{w} \simeq \mathbf{y},$$

Которое необязательно решается. Поэтому мы будем решать его приближенно.

Мультилинейная регрессия

А что, собственно, делать

И тут снова встает вопрос о приближении. Как искать самое близкое решение?

Вспомним занятие по оптимизации. Мы уже умеем минимизировать функции. Остается только выбрать функцию, которую мы будем минимизировать. Это и должна быть функция, показывающая приближенность нашего решения к идеальному.

Мультилинейная регрессия

А что, собственно, делать

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)})^2 = \frac{1}{N}\sum_{i=1}^{N}((\mathbf{w}^T\mathbf{x}^{(i)})^2 - 2(\mathbf{w}^T\mathbf{x}^{(i)}y^{(i)}) + (y^{(i)})^2)$$

$$\frac{d}{dw_j}MSE = \frac{1}{N}\sum_{i=1}^{N}2\mathbf{w}^T\mathbf{x}^{(i)}x_{ij} - 2x_{ij}y^{(i)} = \frac{2}{N}\sum_{i=1}^{N}(\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)})x_{ij}$$