Линейная регрессия

Корреляция — статистическая взаимосвязь двух или более случайных величин, при этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

$$r_{xy} = \frac{\sum (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \cdot \sum (y_i - \overline{y})^2}}$$

Коэффициент корреляции:

- может принимать значения от -1 до +1
- знак коэффициента показывает направление связи (прямая или обратная)
- абсолютная величина показывает силу связи

Линейная регрессия

- Строит линейную поверхность в качестве решения
- Ответ – взвешенная сумма значений признаков

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

Метод наименьших квадратов

- Минимизация $MSE = \sum_{k=0}^{n} (y_k - \hat{y}_k)^2 \ /\mathrm{n}$

- Условие минимума – равенство нулю производной, откуда
- $\mathrm{w} = (X^T X)^{-1} X^T Y$

Линейная регрессия

- Хорошо работает, если зависимость между целевой переменной и признаками имеет линейный характер
- Хорошо интерпретируется
- Перед использованием нужно провести нормализацию данных

Регуляризация

- Дисперсия – это способность модели реагировать на незначительные колебания в тренировочных данных
- Высокая дисперсия приводит к переобучению – ситуации, когда построенная модель хорошо работает на обучающей выборке, но относительно плохо работает на примерах, не участвовавших в обучении
- Решение – использование регуляризации – подавления весов

L2 – регуляризация (Ridge-регрессия)

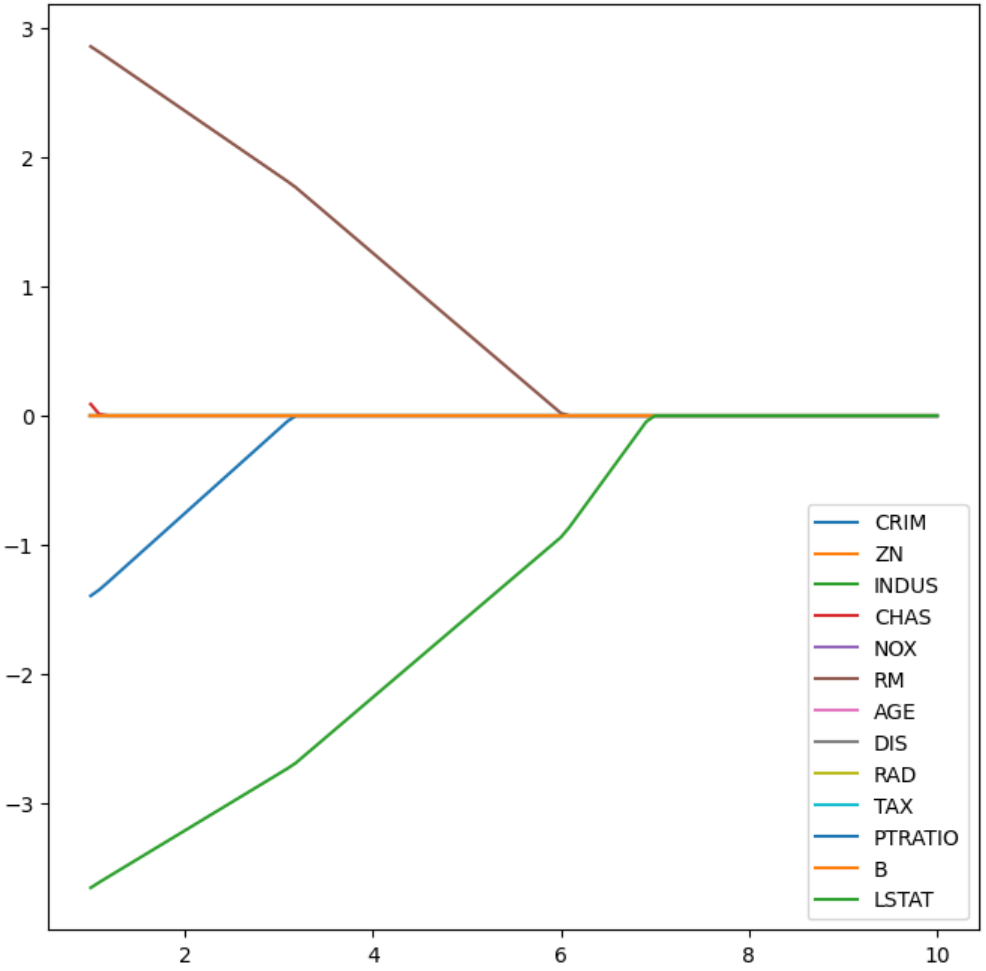
- $L(w)$ – функция потерь
- $L_2 = L(w) + \lambda \sum_{j=0}^{s} {w_j}^2$

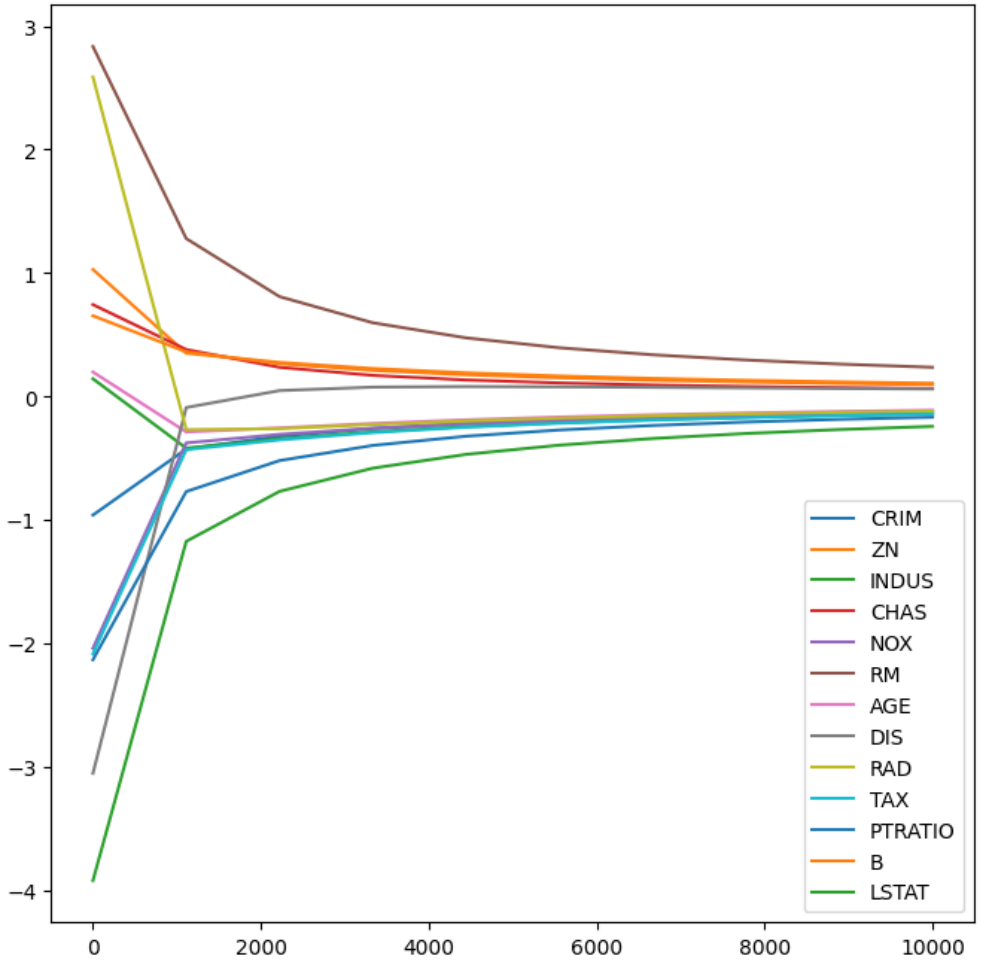
- λ – параметр регуляризации, коэффициент штрафа модели
- чем крупнее вес, тем он крупнее штрафуется

L1 – регуляризация (Lasso)

- $L(w)$ – функция потерь
- $L_2 = L(w) + \lambda \sum_{j=0}^{s} |w_j|$

Сравнение L1 и L2





Регуляризация

- Повышает стабильность модели
- Помогает в борьбе с переобучением
- L2 регуляризация имеет более «мягкий» отбор признаков
- Иногда применяется комбинация L1 и L2 – Elastic Net
- Может быть применена к линейной регрессии, логистической регрессии и не только
- Лучше работает с нормализованными признаками