Метод градиентного спуска. Метрики. Нормализация. Lasso и Ridge

Занятие 3

Как мы решаем задачу линейной регрессии?

Как мы решаем задачу линейной регрессии?

- Находим веса в уравнении $w_1x_1 + w_2x_2 + \dots + w_{n-1}x_{n-1} + w_n$
- Какую функцию минимизируем?

Как мы решаем задачу линейной регрессии?

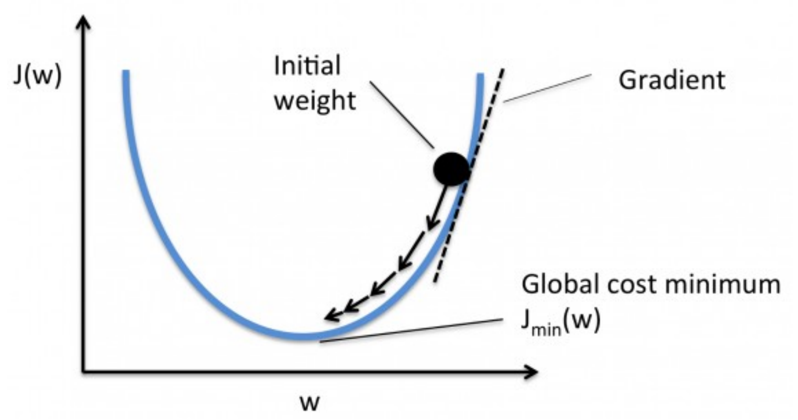
- Находим веса в уравнении $w_1x_1 + w_2x_2 + \ldots + w_{n-1}x_{n-1} + w_n$
- Какую функцию минимизируем?
- Минимизируем функционал ошибки – MSE (но можно и что-то другое)
- Формула: $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- Есть точное аналитическое решение: $\boxed{w = (X^T X)^{-1} X^T y}$

Недостатки аналитического решения

- Высокая вычислительная сложность нахождения обратной матрицы ($O(n^3)$ от числа признаков)
- Обратная матрица может быть вырожденной или плохо обусловленной
- Если заменить среднеквадратичный функционал ошибки на другой, то аналитическое решение может не найтись

Выход – **метод градиентного спуска**

- Вам о нем рассказывал Лагутин…
- Главная идея – в том, что у функции есть градиент: вектор, в направлении которого функция быстрее всего растет.
- Чтобы найти минимум функции, нужно двигаться в противоположную сторону.
- Тут придется считать производные…



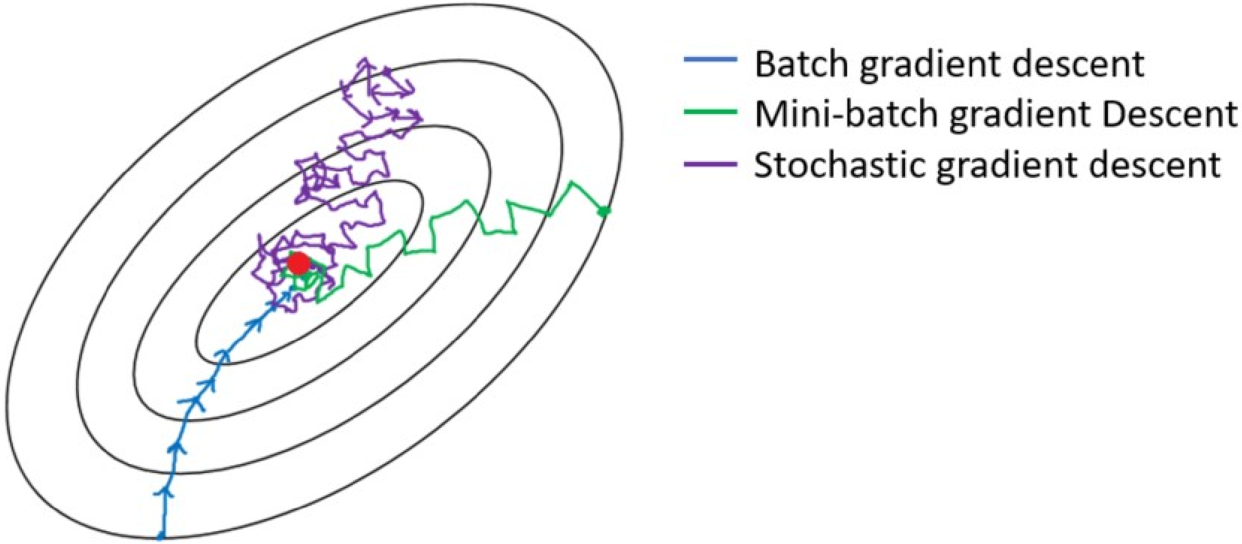
Как это работает

- Инициализируем веса рандомом
- Считаем частные производные для каждого веса
- Если вычесть из веса его частную производную, то вес сдвинется в нужную сторону
- Добиваемся того, что изменения становятся слишком незначительными или вообще перестают происходить – вуаля! Мы в минимуме (возможно, локальном…)

Градиентный спуск бывает:

- Обычный
- Стохастический:
  - на каждом шаге выбираем случайный элемент выборки, вычисляем веса на нем и двигаемся в его сторону
- Mini-batch:
  - выбираем партию случайных элементов (батч) и вычисляем производные только для них

Как это все выглядит:



Градиентный спуск не единственный.

- Модификации градиентного спуска:
  - ADAGRAD
  - Adam
  - Adadelta
  - SGDm – модификация SGD
- RMSPROP
- Метод моментов

Подробнее тут:
http://www.machinelearning.ru/wiki/images/a/a0/2016_417_ChabanenkoVD.pdf

Как оценить качество модели?

- Два способа оценивать качество (любой) модели:
  - смотреть глазками
  - считать циферки
- Для регрессии метрики одни, для классификации - другие

Функционал ошибки и метрики качества

- Функционал (функция) ошибки – функция, которую мы минимизируем, когда обучаем модель
- Метрика качества – функция, которая позволяет оценить уже обученную модель

Функции могут совпадать, но вы не путайте!

Метрики качества. Регрессия

- Какую уже знаете?

Метрики качества. Регрессия

- MSE

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Плюсы:
- подходит для сравнения моделей
- можно контролировать качество во время обучения

Минусы:
- плохо интерпретируется (нужно корень извлекать, чтобы вернуться к тем же единицам измерения)
- не ограничена сверху, как понять, когда качество идеальное?

Метрики качества. Регрессия

- MSE
- RMSE = root mean squared error

$$RMSE(a, X) = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2}$$

Плюсы:
- подходит для сравнения моделей
- можно контролировать качество во время обучения
- сохраняет единицы измерения

Минусы:
- не ограничена сверху, как понять, когда качество идеальное?

Метрики качества. Регрессия

- MSE
- RMSE
- MAE (mean absolute error)

$$MAE(a, X) = \frac{1}{l}\sum_{i=1}^{l} |a(x_i) - y_i|$$

Плюсы:
- менее чувствителен к выбросам, чем MSE

Минусы:
- нельзя брать производные
- тоже не ограничен сверху

Метрики качества. Регрессия

- MSE
- RMSE
- MAE
- R2 (коэффициент детерминации)

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^{l}(a(x_i) - y_i)^2}{\sum_{i=1}^{l}(y_i - \overline{y})^2}$$

Коэффициент детерминации – это квадрат коэффициента корреляции выборки. Коэффициент детерминации оценивает долю дисперсии (изменчивости) Y, которая объясняется с помощью X в простой линейной регрессионной модели.

Чем ближе к 1, тем лучше; если близко к 0, то плохо, если отрицательный – вс чн плх

Метрики качества. Регрессия

- MSE
- RMSE
- MAE
- R2
- MSLE = логарифмический

Подходит для задач с неотрицательной целевой переменной
Штрафует за отклонения в порядке величин
Штрафует заниженные прогнозы сильнее, чем завышенные

$$MSLE(a, X) = \frac{1}{l}\sum_{i=1}^{l}\left(\log(a(x_i) + 1) - \log(y + 1)\right)^2$$

Метрики качества. Регрессия

- MSE
- RMSE
- MAE
- R2
- MSLE
- MAPE (mean absolute percentage error)

$$MAPE(a, X) = \frac{1}{l}\sum_{i=1}^{l}\frac{|y_i - a(x_i)|}{|y_i|}$$

Плюсы:
- ограничена от 0 до 1
- хорошо объясняет (в процентах ошибки)

Минусы:
- по-разному относится к недо- и перепрогнозу, больше котирует недопрогнозы

Метрики качества. Регрессия

- MSE
- RMSE
- MAE
- R2
- MSLE
- MAPE
- SMAPE

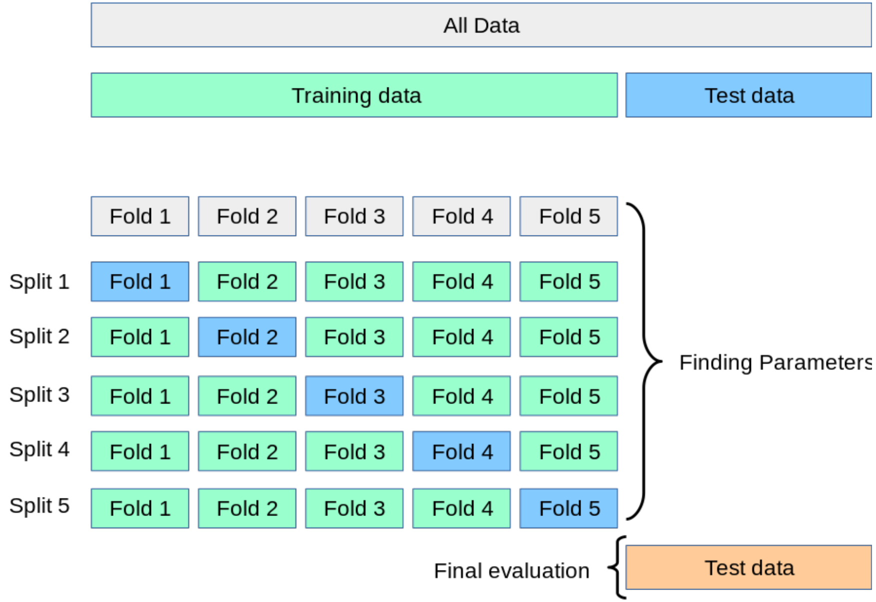
$$SMAPE(a, X) = \frac{1}{l}\sum_{i=1}^{l}\frac{|y_i - a(x_i)|}{(|y_i| + |a(x_i)|)/2}$$

Симметричный мапе: попытка уравновесить недо- и перепрогнозы. Хотя теперь они более близки, но все-таки не равны. Есть мнение, что смапе (как и мапе) – так себе метрика.

Как думаете, какая метрика самая клевая?

Кросс-валидация
(проверка на понимание)

- Разбиваем выборку на трейн и тест **несколько раз**.
- Сколько раз разбили, столько и фолдов.



Регуляризация

- Какие знаете признаки переобучения модели?

Регуляризация

- Признаки переобучения модели:
  - слишком разное качество на train и test
  - слишком большие значения весов
    - причиной больших значений весов могут быть линейно-зависимые фичи
- Как бороться? Регуляризацией!
- Регуляризация штрафует за слишком большие веса

Регуляризация

$$Q_{alpha}(w) = Q(w) + \alpha \cdot R(w) \to \min_{w}$$

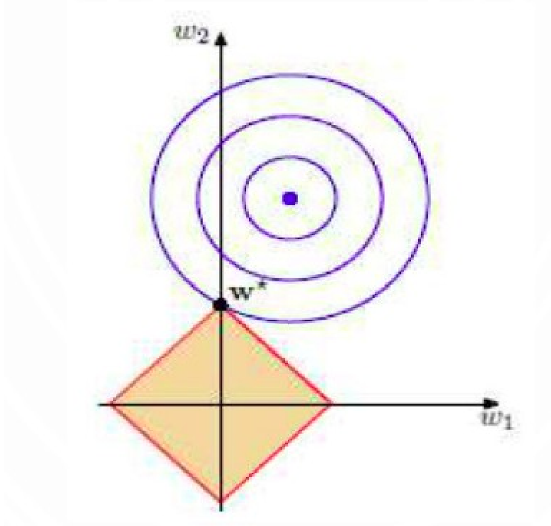
R(w) – это регуляризатор (тоже какая-то функция), α – параметр регуляризации (большое α сильно учитывает регуляризацию, маленькое нет)

- L1-регуляризатор: $R(w) = ||w||_1 = \sum_{i=1}^{d} |w_i|$
- L2-регуляризатор: $R(w) = ||w||_2 = \sum_{i=1}^{d} w_i^2$

- (то есть, L1 – сумма модулей весов, а L2 – сумма их квадратов)

Как это работает?

- L1 зануляет незначительные признаки => естественный отбор фич! (Не всегда хорошо)
- Картинка для ситуации, когда у нас всего 2 веса ----->



Как это работает?

- L2 не зануляет незначительные признаки, но делает их близкими нулю
- Картинка для ситуации, когда у нас всего 2 веса ----->