Лекция 9

Линейная регрессия с точки зрения МО. Задача классификации

Курс: Введение в DS на УБ и МиРА (весна, 2022)
Преподаватель: Владимир Омелюсик

30 мая 2022 г.

В предыдущих сериях

- Тестирование гипотез в линейной регрессии.
- Основные понятия машинного обучения.
- Виды задач машинного обучения.

Линейная регрессия

- Всё то же самое, что обсуждали до этого.
- Важно только качество предсказаний.
- Проблемы с обучением по формулам:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

  - Если матрица $(X^T X)$ необратима, то будут проблемы с вычислениями.
  - Произведение матриц – долгая операция.

Обучение: градиентный спуск

Обучение: градиентный спуск

- В многомерном случае рассчитываем градиент:

$$\nabla_x f(x) = (\frac{df}{dx_1}, \dots, \frac{df}{dx_d})$$

- Например, градиент MSE:

$$\nabla_\beta MSE = \frac{2}{N} X^T (Xw - y)$$

- Градиентный спуск для обучения:

$$\hat{\beta}_{t+1} = \hat{\beta}_t - \alpha \nabla_\beta MSE(\hat{\beta}_t),$$

где $\alpha > 0$ – длина шага.

Алгоритм градиентного спуска

1. Выбираем начальное приближение β_0.
2. Повторяем
$$\hat{\beta}_{t+1} = \hat{\beta}_t - \alpha \nabla_\beta MSE(\hat{\beta}_t),$$
3. Останавливаемся, если
$$\|\hat{\beta}_t - \hat{\beta}_{t-1}\|_2 \leqslant \varepsilon$$

Проблема градиентного спуска

- Градиентный спуск находит только локальные минимумы.
- Решение: мультистарт

Длина шага

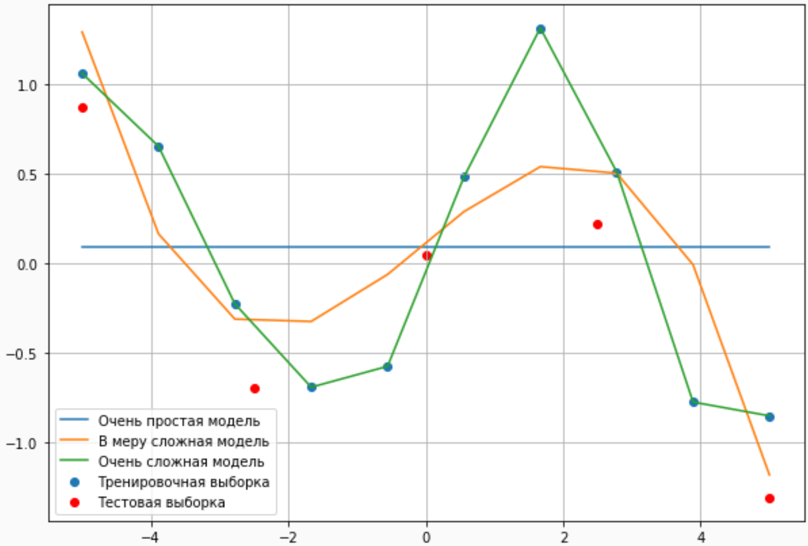
$$\beta_{t+1} = \beta_t - \alpha \nabla_\beta \mathit{MSE}(\beta_t),$$

- Позволяет контролировать скорость обучения.
- Если сделать слишком большой, можно «перепрыгнуть» минимум.
- Гиперпараметр, нужно подбирать.

Обобщающая способность модели

- Обобщающая способность – способность модели давать корректные предсказания на новых данных, не участвовавших при обучении.
- Недообучение – ситуация, когда модели не удалось правильно «запомнить» зависимости в данных. В этом случае качество будет низким как на обучающей выборке, так и на тестовой.
- Переобучение – ситуация, когда модель идеально «запомнила» свойства обучающей выборки, но не общие зависимости в ней. В этом случае качество будет высоким на обучающей выборке, но низким на новых данных.

Обобщающая способность модели



Переобучение в линейной регрессии

- Наблюдение: большие веса могут свидетельствовать о переобучении.

$$\hat{y}_i = 0.2 + 1495.23x_i + \dots,$$

  если x_i – вес человека в кг, а y_i – рост человека в см – странно.
- Идея – штрафовать большие веса.

Регуляризация

- Добавим к функции потерь регуляризатор. Например,

$$\|\beta\|_2^2 = \sum_{j=1}^{k} \beta_j^2$$

- Новая функция потерь:

$$\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

  где λ – коэффициент регуляризации.

- Запускаем градиентный спуск на новой функции потерь.
- Важно не включать в регуляризатор w_0.

Метрики качества на тестовой выборке

Всё те же, что были для статистики: MSE, MAE, R^2, ...

Метод k ближайших соседей

- Дано: X и y.
- Решаем задачу многоклассовой классификации: каждое наблюдение может относиться к одному из K классов: $y_i \in \{1, 2, \dots, K\}$.

Гипотеза компактности
У «похожих» друг на друга объектов будут «похожие» ответы.

- Как определить похожесть? Для числовых признаков, например, так:

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{k} (x_1^j - x_2^j)^2}$$

Метод k ближайших соседей

1. **Обучение.** В kNN отсутствует. На этапе обучения происходит запоминание обучающей выборки X, y.
2. **Предсказание.**
   2.1 Пусть нужно сделать предсказание для нового объекта x_i. Отсортируем объекты обучающей выборки по расстоянию до этого объекта.

   $$d(x_i, x_{(1)}) \leqslant d(x_i, x_{(2)} \leqslant \dots$$

   2.2 Предсказываем самый популярный класс среди k ближайших соседей.

   $$\hat{y}_i = \arg\max_C \sum_{i=1}^{k} [y_{(i)} = C]$$

Пример: kNN

Расстояния

- Числовые признаки.
  - Евклидово расстояние.

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{k} (x_1^j - x_2^j)^2}$$

  - Манхэттэнское расстояние.

$$d(x_1, x_2) = \sum_{j=1}^{k} |x_1^j - x_2^j|$$

- Категориальные признаки.
  - Считающее расстояние.

$$d(x_1, x_2) = \sum_{j=1}^{k} [x_1^j \neq x_2^j]$$

kNN: свойства

- Простой метод, основанный на расчётах расстояний.
- Гиперпараметры: число соседей k и функция расстояния.
- Проблема: поиск соседей может занимать долгое время.

Логистическая регрессия

- Дано: X и y.
- Решаем задачу бинарной классификации: $y_i \in \{-1, 1\}$.

Линейная классификация

Отступ

Обучение линейного классификатора

Верхняя оценка на функцию потерь

Логистическая функция потерь

Обучение

Мягкая и жёсткая классификация

Предсказания вероятностей