

Лекция 5

Выбор модели. Кросс-валидация. Отбор признаков

Макаренко В.А., Габдуллин Р.А.

МГУ им. М.В. Ломоносова

7 февраля 2023

Проблемы недообучения и переобучения

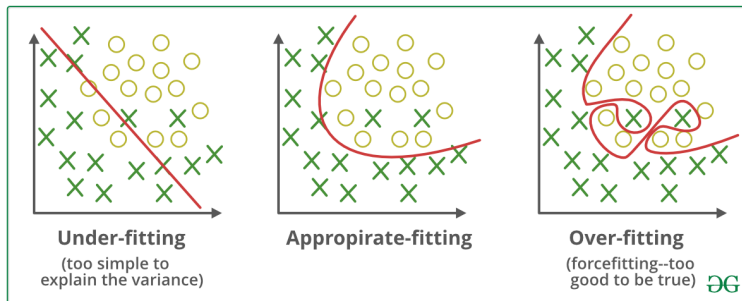


Рис.: Источник: geeksforgeeks.org

- **Недообучение (Underfitting).** Модель не обеспечивает достаточно малой величины средней ошибки на обучающей выборке.
- **Переобучение (Overfitting).** Средняя ошибка на новых данных (тестовой выборке) существенно выше, чем на обучающей выборке.

Причины недообучения и переобучения

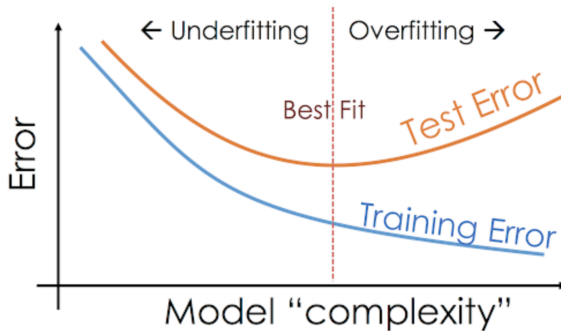


Рис.: Источник: analyticsvidhya.com

- **Недообучение.** Возникает из-за использования недостаточно гибкого семейства моделей.
- **Переобучение.** Возникает при использовании избыточно сложных моделей.

Оценка качества на отложенной выборке

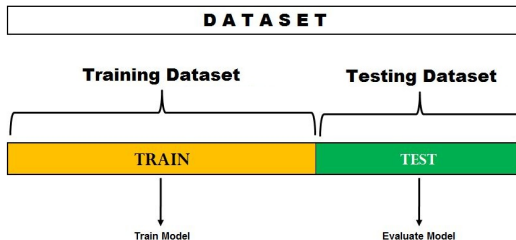


Рис.: Источник: datavedas.com

Оценка качества на отложенной выборке:

- Разбиваем обучающую выборку на две части: train и test.
- Обучаем модель на train.
- Оцениваем качество на test.

Проблема:

- Результат обучения и тестирования может сильно зависеть от выбранного разбиения.

Кросс-валидация

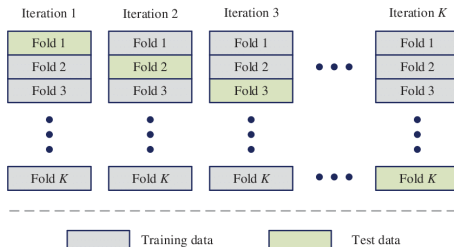


Рис.: Источник: [researchgate.net](https://www.researchgate.net)

Оценка качества на кросс-валидации:

- Разбиваем выборку на K частей
- По очереди используем j -ую часть ($1 \leq j \leq K$) для тестирования, а все остальные части для обучения.
- Усредняем K полученных оценок.

Оценка на кросс-валидации получается более надежной по сравнению с качеством на отложенной выборке, но при этом требуется обучить K моделей вместо одной.

Выбор модели с помощью кросс-валидации

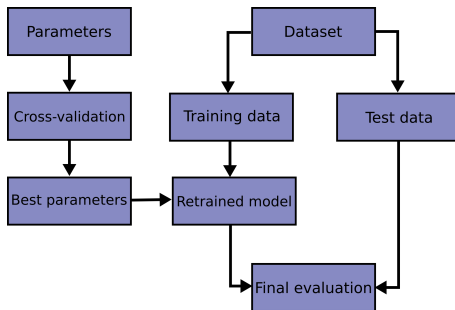


Рис.: Источник: scikit-learn.org

- Для каждой модели получаем оценку качества на кросс-валидации
- Выбираем модель с наилучшей оценкой качества.
- Обучаем лучшую модель на всей обучающей выборке.
- Проверяем итоговое качество на отложенной выборке.

Стратифицированная кросс-валидация

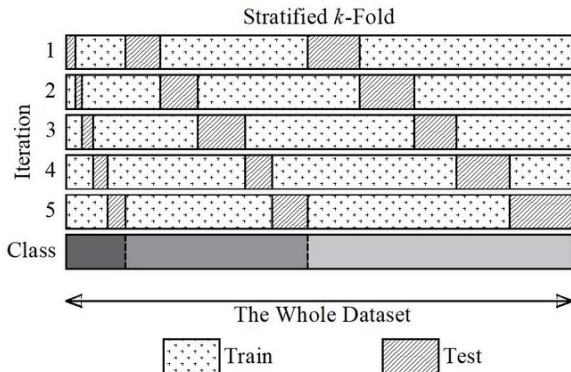


Рис.: Источник: [researchgate.net](https://www.researchgate.net)

Каждый набор содержит примерно ту же долю каждого целевого класса, что и вся обучающая выборка.

Кросс-валидация для групп зависимых наблюдений

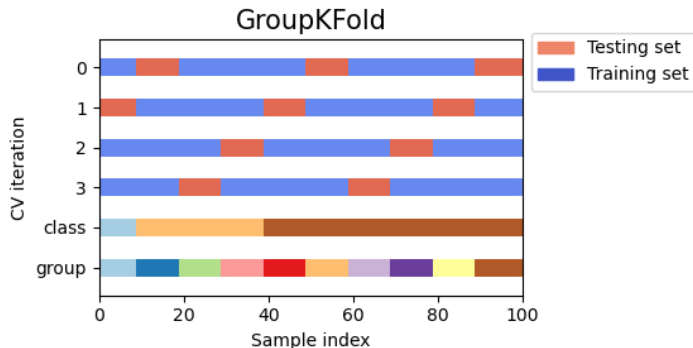


Рис.: Источник: scikit-learn.org

Все примеры каждой группы попадают либо в обучение, либо в валидацию.

Кросс-валидация для временных рядов

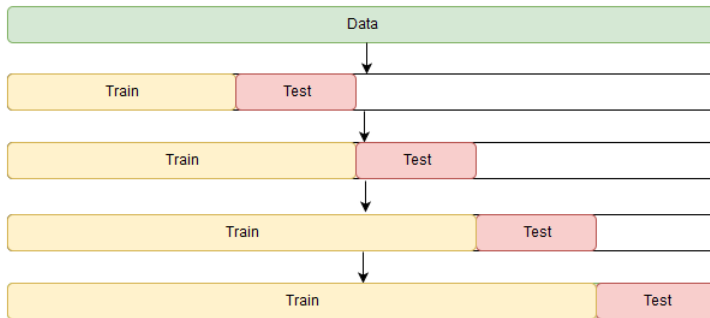


Рис.: Источник: stats.stackexchange.com

В обучение попадают примеры из прошлого, а в валидацию — из будущего.

Оптимизация гиперпараметров. Grid search.

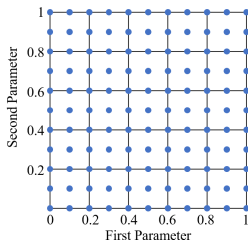


Рис.: Источник: [researchgate.net](https://www.researchgate.net)

- Перебираем гиперпараметры по сетке.
- Проводим исчерпывающий поиск по всему пространству гиперпараметров.
- Экспоненциальное по числу гиперпараметров время работы, опасность переобучения, можем проверить только фиксированное множество значений у каждого из параметров.

Оптимизация гиперпараметров. Random search.

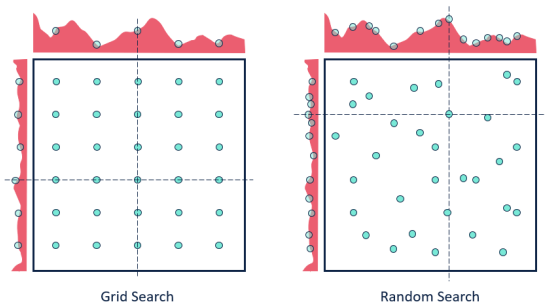


Рис.: Источник: alteryx.com

- Выбираем случайную точку из пространства гиперпараметров.
- Обычно требует сильно меньше времени, чем Grid search.
- Для каждого параметра просматриваем сильно больше значений.

Оптимизация гиперпараметров. Bayesian search.

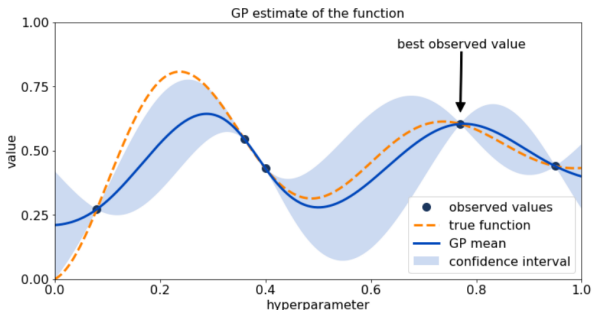


Рис.: Источник: cern.ch

- На основе предыдущих просмотренных точек прогнозируем, в какой следующей точке с наибольшей вероятностью лучшим оценку.

Инструменты для выбора моделей и оптимизации гиперпараметров

Библиотеки:

- sklearn: модуль *model_selection*
- hyperopt
- bayes_opt
- optuna

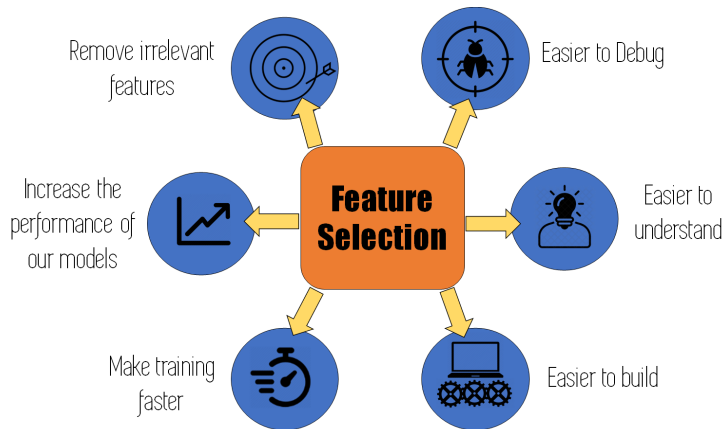


Рис.: Источник: towardsdatascience.com

Удаление признаков с малой дисперсией

- Предположение: чем меньше дисперсия у признака, тем меньше он содержит информации (если дисперсия равна нулю, то признак является константным).
- Удаляем признаки, у которых дисперсия ниже заданного порога.

Отбор признаков по степени зависимости от целевой переменной

- Для каждого признака оцениваем степень зависимости от целевой переменной.
- Выбираем признаки с наиболее выраженной зависимостью.

Способы измерения зависимости:

- Статистические тесты о независимости (например, критерий χ^2).
- Оценка взаимной информации.

$$I(X, Y) = \sum \sum P(x, y) \log \frac{P(x, y)}{P(x)P(y)}.$$

- Оценка корреляции.

Отбор признаков с помощью L_1 -регуляризации

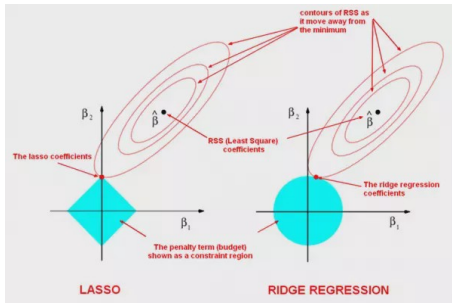


Рис.: Источник: medium.com

- Обучаем линейную модель (например, линейную регрессию или логистическую регрессию) с L_1 -регуляризацией.
- Отбираем признаки с ненулевыми весами.
- Варьируя силу регуляризации, можно отобрать желаемое количество признаков.

Рекурсивное удаление признаков

Пусть имеется вспомогательная модель, с помощью которой будем оценивать важность признаков (например, линейная модель).

Процесс отбора признаков:

- Обучаем вспомогательную модель на всех признаках.
- Отбрасываем наименее важные признаки (например, признаки с очень маленькими по модулю весами, если обучаем линейную модель).
- Повторяем процесс для выбранных признаков.

Последовательный отбор признаков

Три основных способа отбора признаков:

- Forward selection. Жадный итеративный процесс добавления признака на каждом шаге.
 - Начинаем без признаков.
 - Находим признак, добавление которого минимизирует ошибку.
 - Повторяем процесс, добавив этот признак в модель.
- Backward elimination. Жадный итеративный процесс удаления признака на каждом шаге.
 - Сначала рассматриваем все множество признаков.
 - Находим признак, удаление которого минимизирует ошибку на кросс-валидации.
 - Повторяем процесс, удалив этот признак из модели.
- Bidirectional elimination. Гибрид forward selection и backward selection.
 - На каждом шаге решаем, удалить или добавить какой-либо признак.

- Недообучение и переобучение.
- Оценка качества на отложенной выборке.
- Оценка качества на кросс-валидации.
 - Стратифицированная кросс-валидация.
 - Кросс-валидация для групп зависимых наблюдений.
 - Кросс-валидация для временных рядов.
 - Выбор модели с помощью кросс-валидации.
- Отбор признаков.
 - Удаление признаков с малой дисперсией.
 - Отбор признаков по степени зависимости от целевой переменной.
 - Отбор признаков с помощью L_1 -регуляризации.
 - Рекурсивное удаление признаков.
 - Последовательный отбор признаков.
 - Forward selection.
 - Backward selection.
 - Bidirectional selection.