

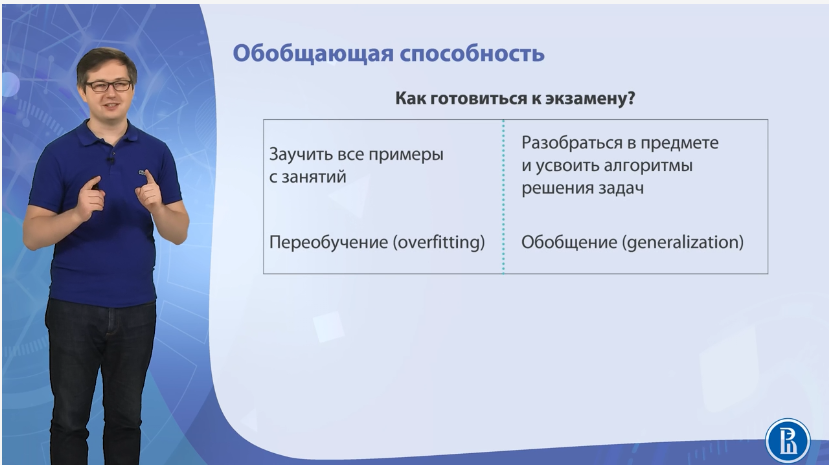
Проблема подбора k



k=1 – приводит к очень плохой модели, которая дает низкое качество на новых данных

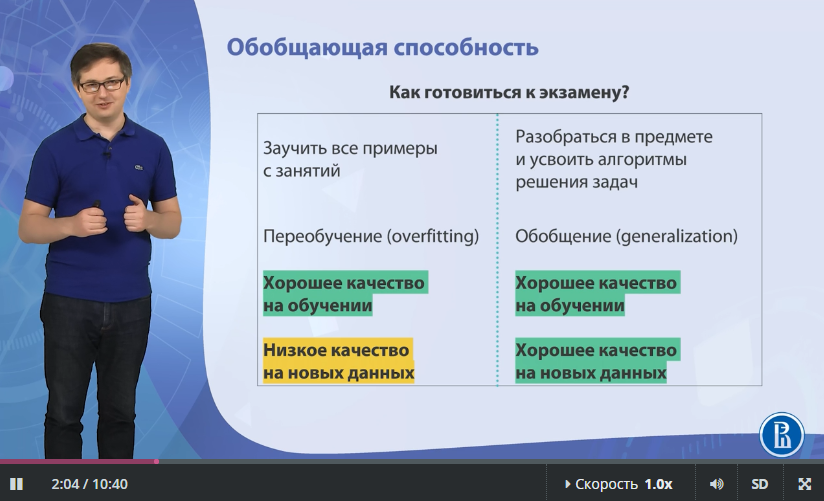
Большее значение k дает хорошее качество





Overfitting – слишком сильно подгоняемся под выборку

Generalization – обобщение и способность решать похожие задачи



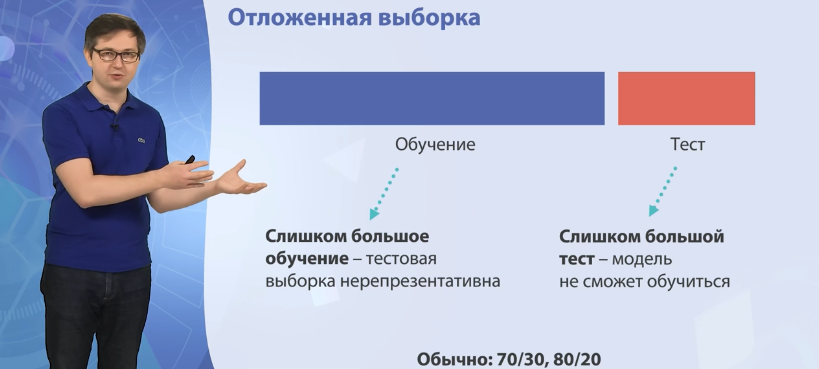
Это легко поймать, если использовать несколько наборов данных

Для измерения понадобятся две выборки:



Тестовую изначально модели не показываем. На тестовой будем измерять качество модели

Как разбивать выборку?



Обычно пользуются опытом или здравым смыслом

У разбиения данных на две выборки есть недостатки:

Если в выборке есть особенные объекты, на которых модель всегда будет ошибаться, и они все попадут в тестовую выборку, то качество на этой тестовой выборке будет плохим из-за этих объектов. Мы этого не узнаем, т.к. мы зафиксировали разбиение и не будем пытаться по другому поделить выборку.

Способ, который этой проблемы избегает, называется кросс-валидация



Мы берем все размеченные данные, всю обучающую выборку (синий блок сверху), и разбиваем на несколько блоков примерно равного размера. Сначала объявляем тестовым третий блок, строим модель на обучающей выборке (1 и 2 блок), проверяем качество на тестовой выборке.

Затем объявляем тестовым второй блок, а 1 и 3 – обучающая выборка, и т.д.

В итоге получаем три значения качества, мы их усредняем и получаем оценку качества по кросс-валидации. В таком случае каждый объект побывает в тесте и в обучении





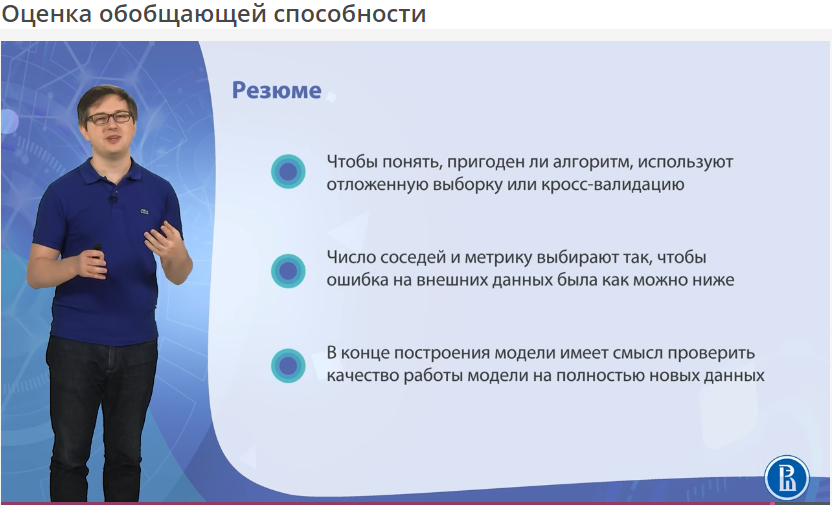
Синяя – ошибка на обучающих данных, красная – на тестовых данных.

При k=1 ошибка на обучающей выборке нулевая

При k слишком большом ошибка начинает расти



Валидационная- пробуем разные варианты к. К валидационной выборке мы все равно обращаемся много раз и подгоняем под нее. Тестовая – набор новых для модели данных



Давайте проверим, насколько хорошо вы поняли материал прослушанной лекции.

2 из 2 баллов (не оценивается)

Мы проводим кросс-валидацию по пяти блокам. Сколько раз нужно обучить модель, чтобы получить оценку ошибки модели?

10

5 верно

1

Чем кросс-валидация лучше оценивания ошибки по отложенной выборке? Выберите все подходящие ответы.

Отложенная выборка подходит не для всех моделей.

Каждый объект хотя бы раз побывает в обучающей выборке, и мы поймём, не приводит ли он к каким-то проблемам с моделью. верно

Каждый объект хотя бы раз побывает в тестовой выборке, и мы узнаем, если его отсутствие в обучении как-то серьёзно портит модель. верно

Кросс-валидация быстрее, чем оценка по отложенной выборке