Лекция №4
по машинному обучению

Типы данных при линейной регрессии

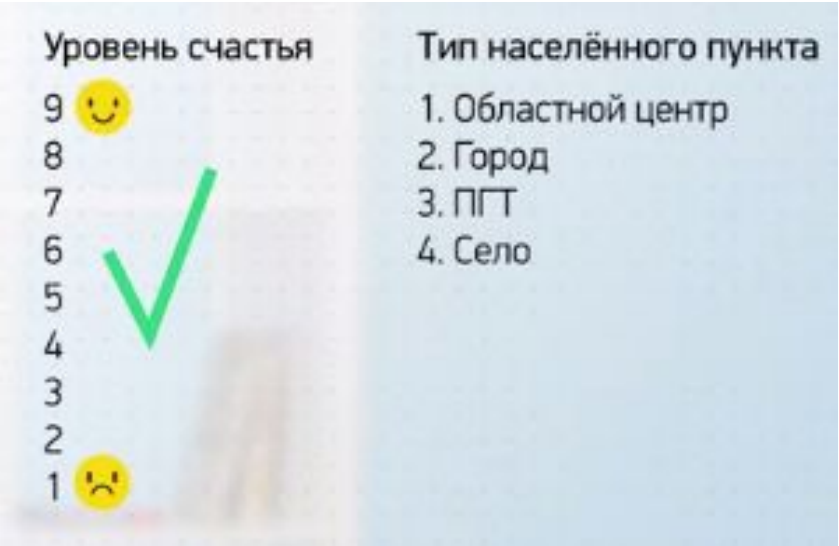
Можно использовать:

- Метрические (количественные) переменные.
- Порядковые (ранговые) переменные.

ИСПОЛЬЗОВАНИЕ ПОРЯДКОВЫХ ДАННЫХ: ОГРАНИЧЕНИЯ

1. Порядковые переменные используются в линейной регрессии только в качестве фактора (отклик может быть только **количественным**).
2. Количество рангов должно быть достаточно велико.
3. Расчёт «среднего» должен иметь содержательный смысл.

Типы данных при линейной регрессии



Бинарные переменные можно включать, если такого что большая часть значений это 0 или 1

Не должно быть выбросов и других аномалий

Типы данных при линейной регрессии

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + \varepsilon$$

y — количество переходов на сайт

x_1 — количество кликов

x_2 — количество звонков

x_3 — день недели

x_4 — есть ли у фирмы аккаунты в социальных сетях

x_5 — медианная позиция фирмы в поисковой выдаче

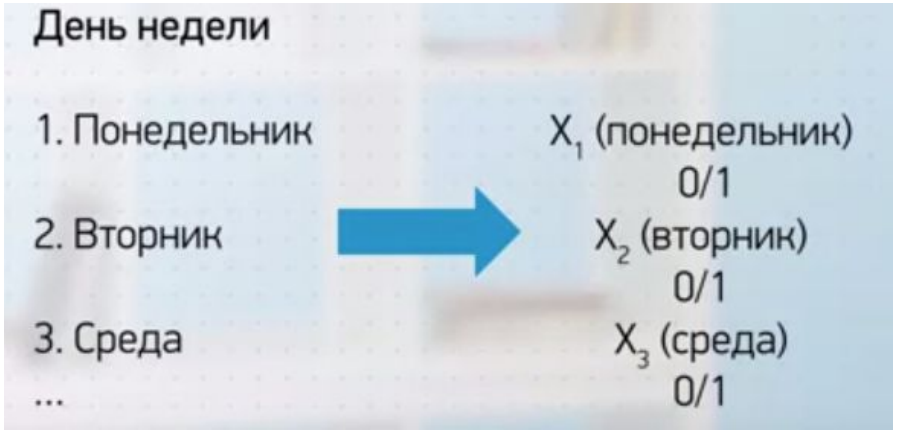
Типы данных при линейной регрессии

ИСПОЛЬЗОВАНИЕ НОМИНАЛЬНЫХ ДАННЫХ: ОГРАНИЧЕНИЯ

В чистом виде не используются.

Можно включать, если перекодировать в набор фиктивных переменных.

Если признак принимает K значений (есть K категорий), то создается K – 1 переменных, принимающих значения 0 – 1.



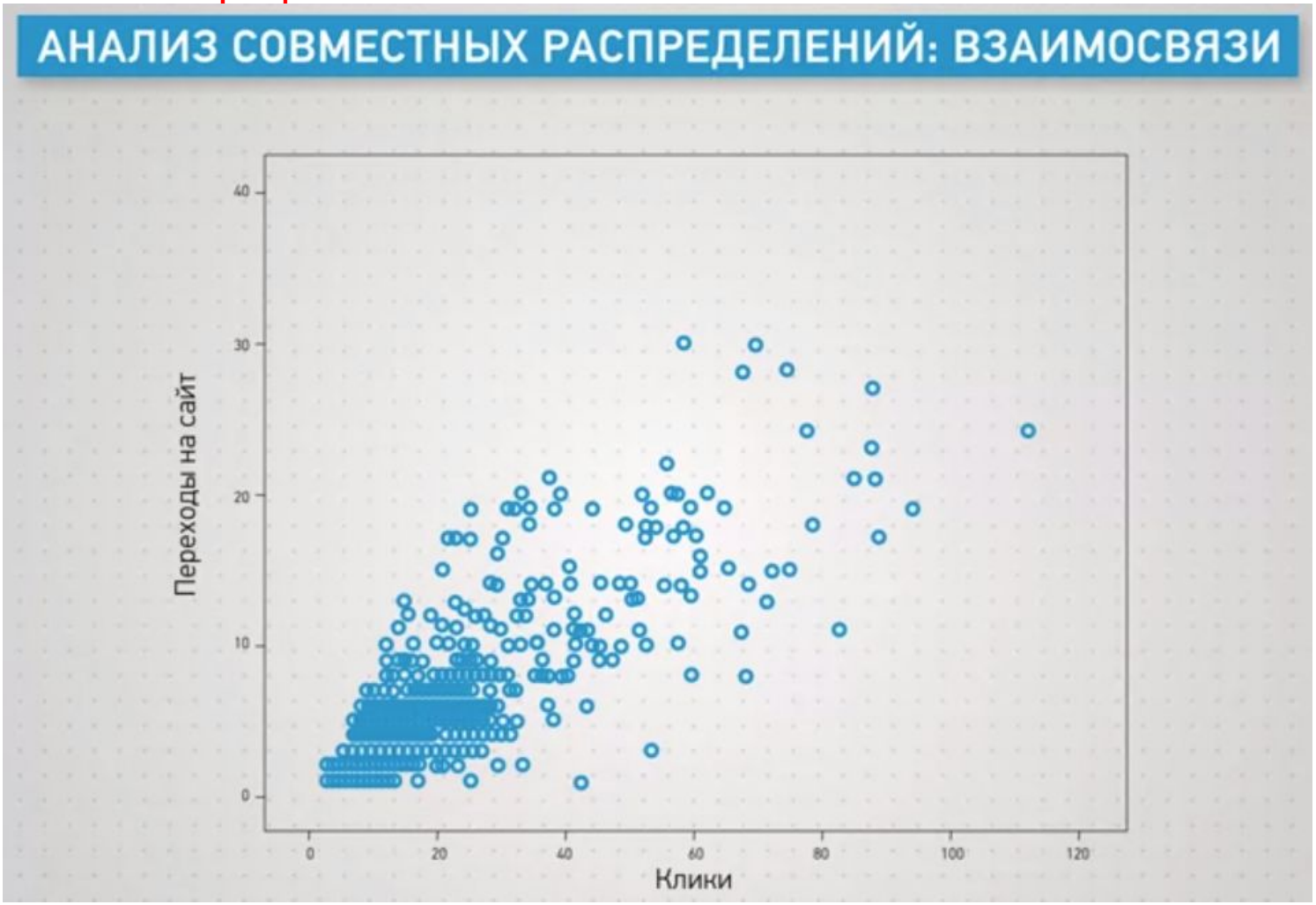
Типы данных при линейной регрессии

ФАКТОРЫ ДОЛЖНЫ БЫТЬ

1. Все связаны с откликом.
2. Не связаны (или слабо связаны) между собой.

АНАЛИЗ ОДНОМЕРНЫХ РАСПРЕДЕЛЕНИЙ: ВЫБРОСЫ

Подготовка данных при линейной регрессии



Подготовка данных при линейной регрессии

ПРЕЖДЕ, ЧЕМ СТРОИТЬ МОДЕЛЬ, УБЕДИТЕСЬ В ТОМ, ЧТО:

1. Все интервальные признаки очищены от выбросов и неопределённых значений.
2. Все факторы связаны с откликом (и слабо связаны между собой).
3. Все признаки нужного типа, если нет — что они преобразованы.

СПИСОК ПЕРЕМЕННЫХ, ПОДГОТОВЛЕННЫХ ДЛЯ ВКЛЮЧЕНИЯ В МОДЕЛЬ

x_1 — количество кликов (очищены от выбросов)

x_2 — количество звонков (очищены от выбросов)

$x_3 - x_8$ — фиктивные переменные для дней недели (7 дней, 6 переменных)

x_9 — есть ли у фирмы аккаунты в социальных сетях

x_{10} — позиция в выдаче: корреляционный анализ показал, что этот фактор значимо не связан с откликом. В модель включать бессмысленно.

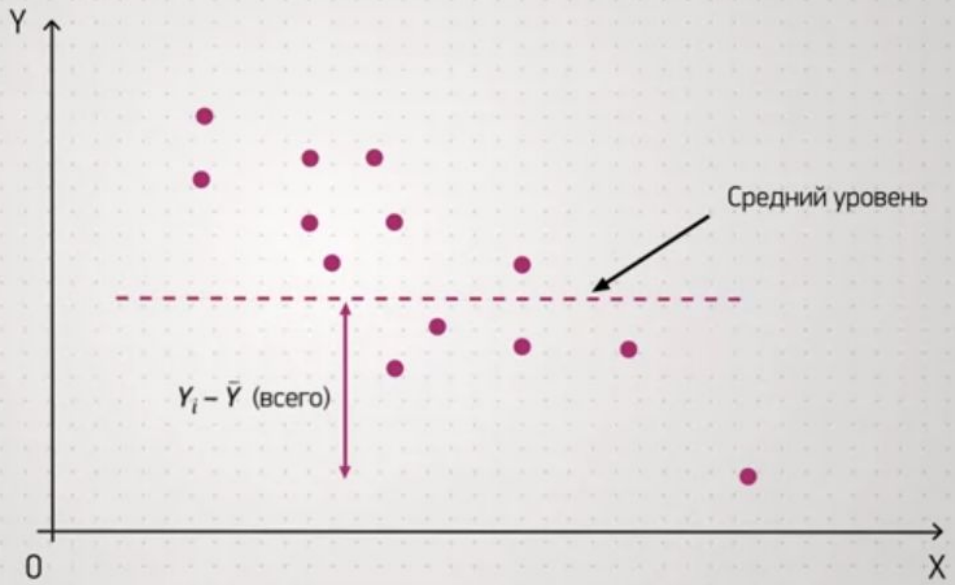
Оценка качества линейной регрессии

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Коэффициент детерминации R^2 — это доля дисперсии отклика, которая объясняется рассматриваемой моделью зависимости.

Общая дисперсия отклика

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$



Оценка качества линейной регрессии

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Коэффициент детерминации R^2 — это доля дисперсии отклика, которая объясняется рассматриваемой моделью зависимости.

КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Общая дисперсия отклика

$$TSS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

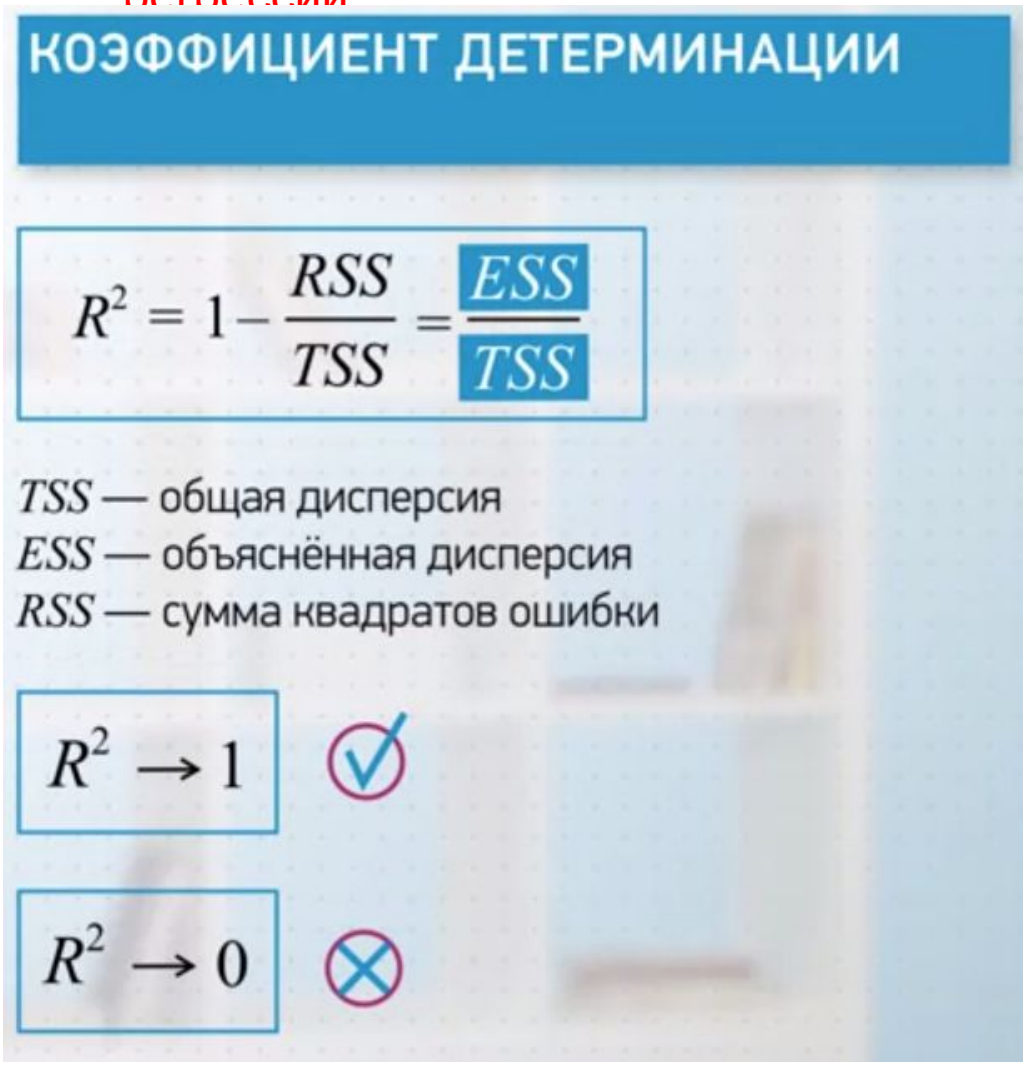
Объяснённая дисперсия

$$ESS = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

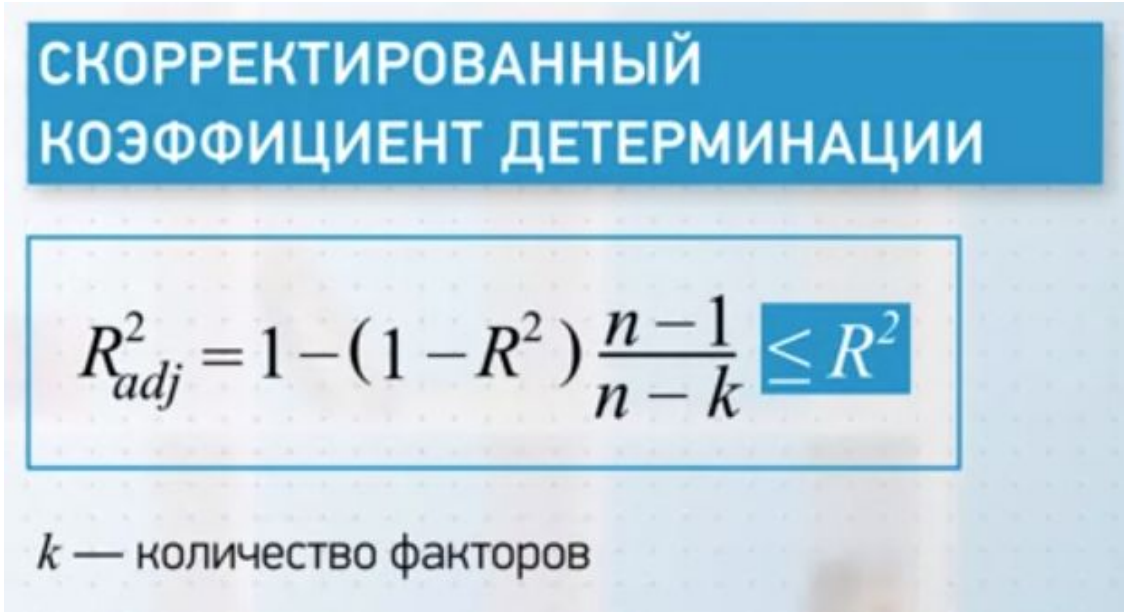
Сумма квадратов ошибки

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Оценка качества линейной регрессии



Оценка качества линейной регрессии



Оценка качества линейной регрессии

ПОЧЕМУ ВАЖНО НЕ ВКЛЮЧАТЬ В МОДЕЛЬ МНОГО ПРИЗНАКОВ?

1. Модель с большим количеством признаков трудно интерпретировать.

2. Модель с большим количеством признаков становится менее устойчивой.

Модель с большим количеством признаков становится неустойчивой, потому что один незначимый для неё признак может оказаться с большим значением и повлиять на отклик.

Оценка качества линейной регрессии

Оценка качества линейной регрессии

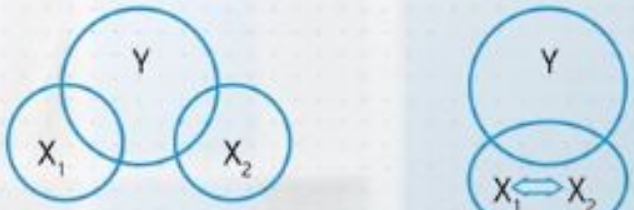
1. Нет единственной оценки МНК (уравнения с совсем разными коэффициентами могут давать результат одного качества).

$X_2 = 10 \times X_1$

$Y = 10 + 10 \times X_1 + 10 \times X_2$

$Y = 10 - 10 \times X_1 + 12 \times X_2$

ПОЧЕМУ ЭТО ПЛОХО?



2. Невозможно отделить эффекты каждого фактора в отдельности.

Оценка качества линейной регрессии

ПОЛНАЯ МУЛЬТИКОЛЛИНЕАРНОСТЬ

Причина — ошибка исследователя
(на этапе отбора признаков что-то пошло не так).

Например, распространённая ошибка —
включение лишней фиктивной переменной.

X_1 — понедельник
X_2 — вторник
X_3 — среда
X_4 — четверг
X_5 — пятница
X_6 — суббота

Оценка качества линейной регрессии

Оценка качества линейной регрессии

ЧАСТИЧНАЯ МУЛЬТИКОЛЛИНЕАРНОСТЬ — ЭТО ПЛОХО?

1. Можно не делать ничего.
2. В отличие от полной (строгой) частичная мультиколлинеарность не нарушает предположений о регрессионной модели.
3. Но может стать проблемой, когда коэффициенты корреляции большие (>0,8-0,9).

Оценка качества линейной

Оценка качества линейной регрессии

КАК ОБНАРУЖИТЬ МУЛЬТИКОЛЛИНЕАРНОСТЬ: НЕКОТОРЫЕ СИМПТОМЫ

4. Коэффициент VIF (variance inflation factor) рассчитывается для каждого фактора и показывает степень увеличения дисперсии за счёт коррелированности каждого фактора с остальными.

ЕСЛИ МУЛЬТИКОЛЛИНЕАРНОСТЬ ПРОБРАЛАСЬ В МОДЕЛЬ

1. Исключить одну из переменных.
2. Преобразовать коррелированные переменные.
3. Ничего не делать.

Смещение, разброс, переобучение и недообучение.

Переобучение (overfitting) – явление, когда ошибка на тестовой выборке заметно больше ошибки на обучающей. Это главная проблема машинного обучения: если бы такого эффекта не было (ошибка на тесте примерно совпадала с ошибкой на обучении), то всё обучение сводилось бы к минимизации ошибки на тесте (т.н. эмпирическому риску)

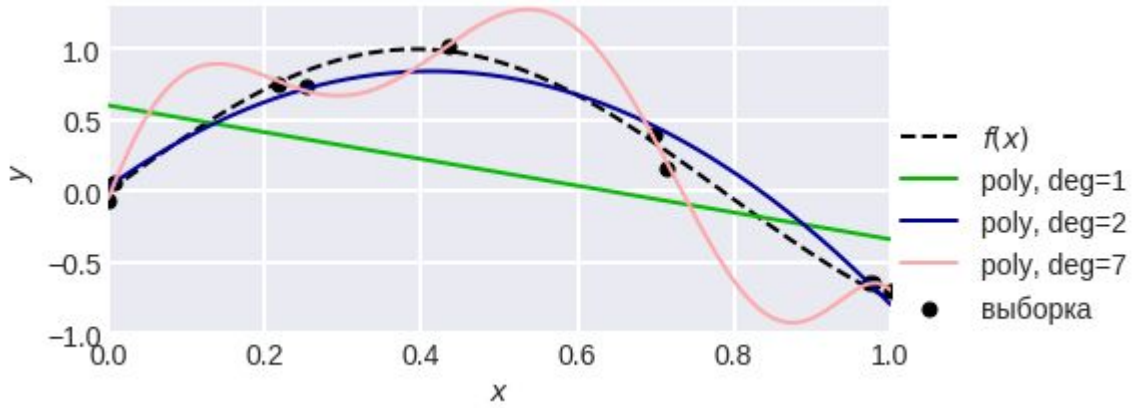
Недообучение (underfitting) – явление, когда ошибка на обучающей выборке достаточно большая, часто говорят «не удаётся настроиться на выборку». Такой странный термин объясняется тем, что недообучение при настройке алгоритмов итерационными методами (например, нейронных сетей методом обратного распространения) можно наблюдать, когда сделано слишком маленькое число итераций, т.е. «не успели обучиться»

Смещение, разброс, переобучение и недообучение.

Сложность (complexity) модели алгоритмов (допускает множество формализаций) – оценивает, насколько разнообразно семейство алгоритмов в модели с точки зрения их функциональных свойств (например, способности настраиваться на выборки). Повышение сложности (т.е. использование более сложных моделей) решает проблему недообучения и вызывает переобучение.

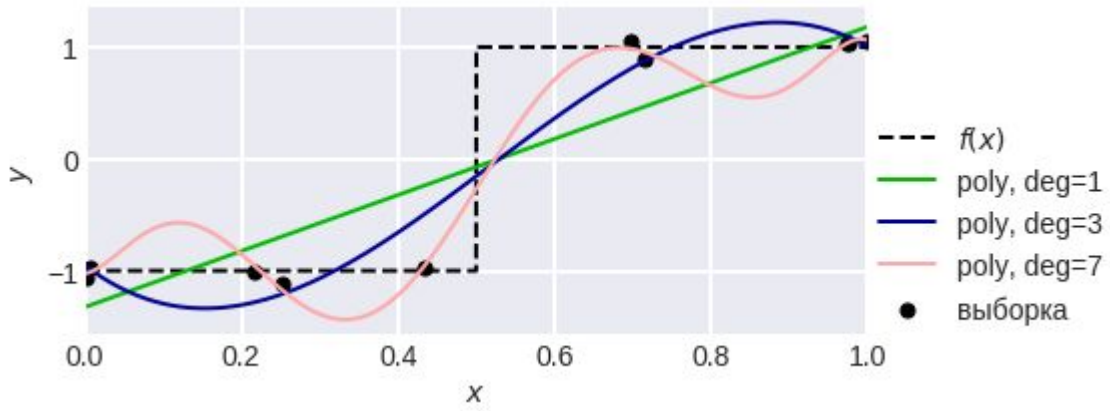
Пример переобучения.

$y = \sin(4x) + \text{шум}$

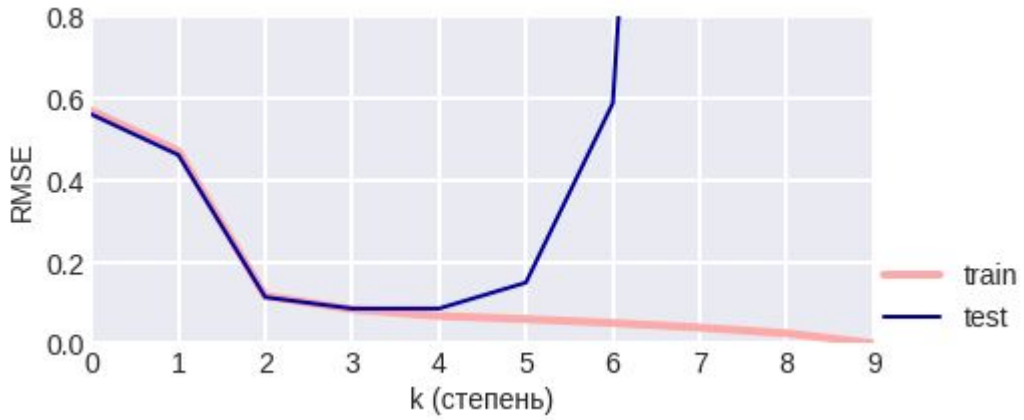


Смещение, разброс, переобучение и недообучение.

Пример переобучения.



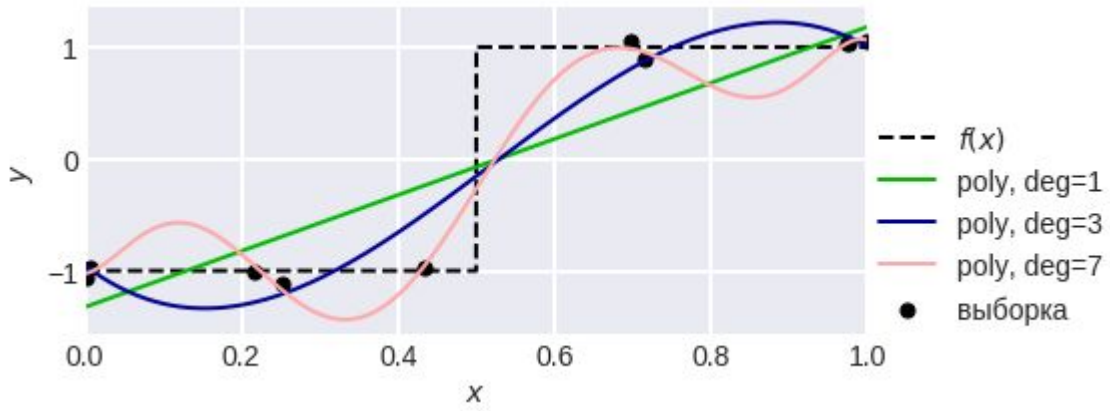
зашумлённой пороговой зависимости



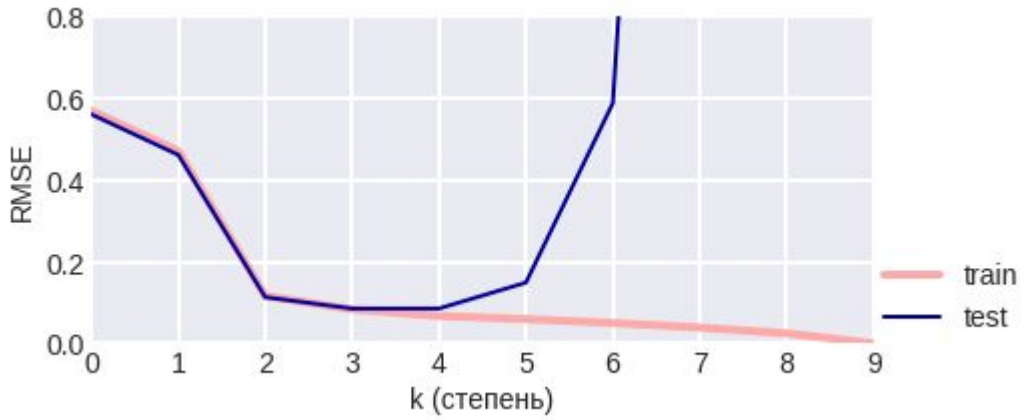
Видно, что с увеличением степени ошибка на обучающей выборке падает, а на тестовой (мы взяли очень мелкую сетку отрезка [0, 1]) – сначала падает, потом возрастает.

Смещение, разброс, переобучение и недообучение.

Пример переобучения.



зашумлённой пороговой зависимости



Видно, что с увеличением степени ошибка на обучающей выборке падает, а на тестовой (мы взяли очень мелкую сетку отрезка [0, 1]) – сначала падает, потом возрастает.

Смещение, разброс, переобучение и недообучение.

Дисперсия случайной величины связана с математическим ожиданием квадрата этой случайной величины следующим соотношением:

$$D_x = \sigma^2 = M_{x^2} - (M_x)^2.$$

$$y \equiv y(x) = f(x) + \varepsilon, \varepsilon \sim \text{norm}(0, \sigma^2)$$

Мы строим алгоритм (в нашем случае полином фиксированной степени) $a=a(x)$, посмотрим чему равно математическое ожидание квадрата отклонения ответа алгоритма от истинного значения:

$$\mathrm{E}(y-a)^2 = \mathrm{E}(y^2 + a^2 - 2ya) =$$

$$= \mathrm{E}\,y^2 - (\mathrm{E}\,y)^2 + (\mathrm{E}\,y)^2 + \mathrm{E}\,a^2 - (\mathrm{E}\,a)^2 + (\mathrm{E}\,a)^2 - 2f\,\mathrm{E}\,a =$$

$$= \mathrm{D}\,y + \mathrm{D}\,a + (\mathrm{E}\,y)^2 + (\mathrm{E}\,a)^2 - 2\,\mathrm{E}\,ya =$$

$$= \mathrm{D}\,y + \mathrm{D}\,a + f^2 + (\mathrm{E}\,a)^2 - 2f\,\mathrm{E}\,a =$$

$$= \mathrm{D}\,y + \mathrm{D}\,a + (\mathrm{E}(f-a))^2 \equiv \sigma^2 + \text{variance}(a) + \text{bias}^2(f,a)$$

Смещение, разброс, переобучение и недообучение.

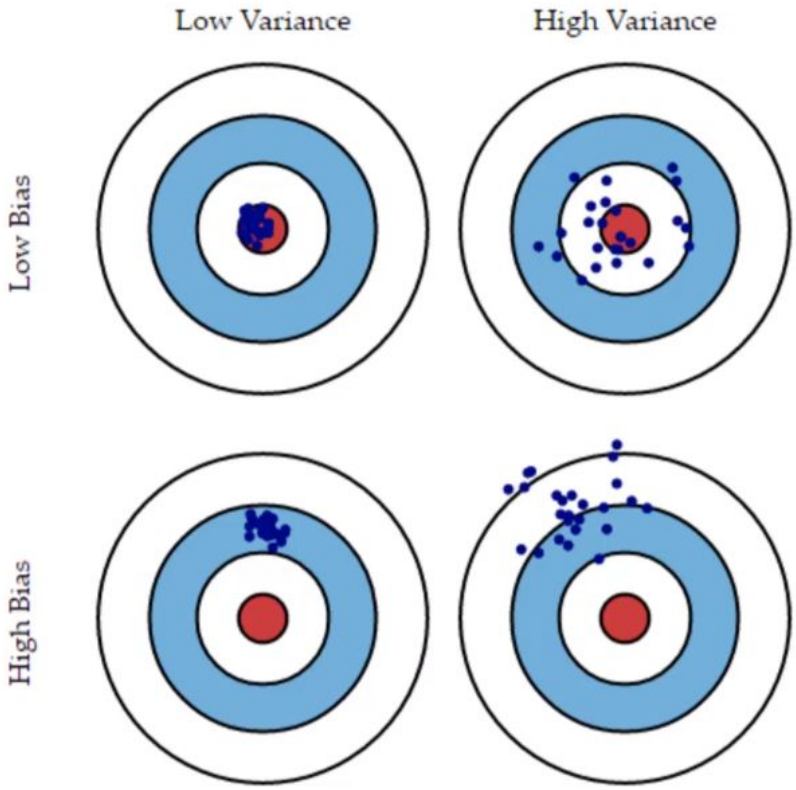
обучающая выборка выбирается случайно из некоторого распределения, настроенный алгоритм тоже случаен, поскольку зависит от выборки, настройка алгоритма также может быть стохастической. Таким образом, матожидание берётся по всем данным (обучающим выборкам) и настройкам алгоритма, а сами формулы записываются в конкретной точке x:

$$\mathrm{E}(y-a)^2 \equiv \mathrm{E}_{(x_i, f(x_i)+\varepsilon_i)_{i=1}^m} (y(x)-a(x))^2$$

Разбросом (variance) мы назвали дисперсию ответов алгоритмов $\mathrm{D}a$, а **смещением (bias)** – матожидание разности между истинным ответом и выданным алгоритмом: $\mathrm{E}(f-a)$. Мы получили, что ошибка раскладывается на три составляющие. Первая связана с шумом в самих данных, а вот две остальные связаны с используемой моделью алгоритмов. Понятно, что **разброс характеризует разнообразие алгоритмов** (из-за случайности обучающей выборки, в том числе шума, и стохастической природы настройки), а **смещение – способность модели алгоритмов настраиваться на целевую зависимость**. Проиллюстрируем это. На

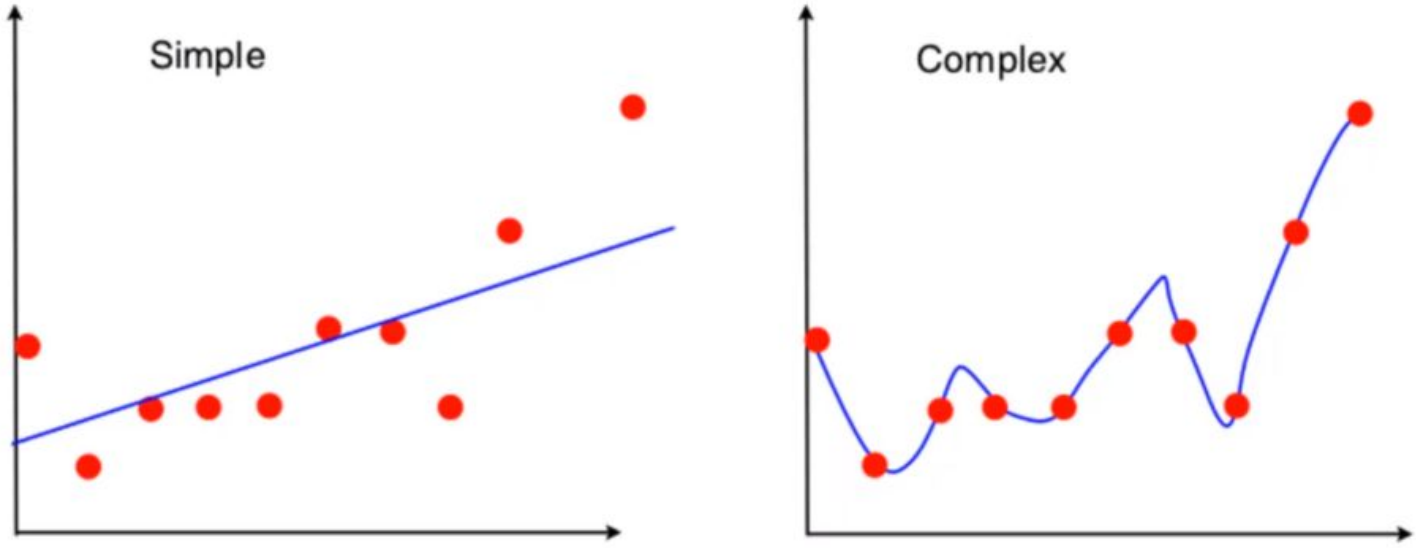
Смещение, разброс, переобучение и недообучение.

Для простых моделей характерно недообучение (они слишком простые, не могут описать целевую зависимость и имеют большое смещение), для сложных – переобучение (алгоритмов в модели слишком много, при настройке мы выбираем ту, которая хорошо описывает обучающую выборку, но из-за сильного разброса она может допускать большую ошибку на тесте).

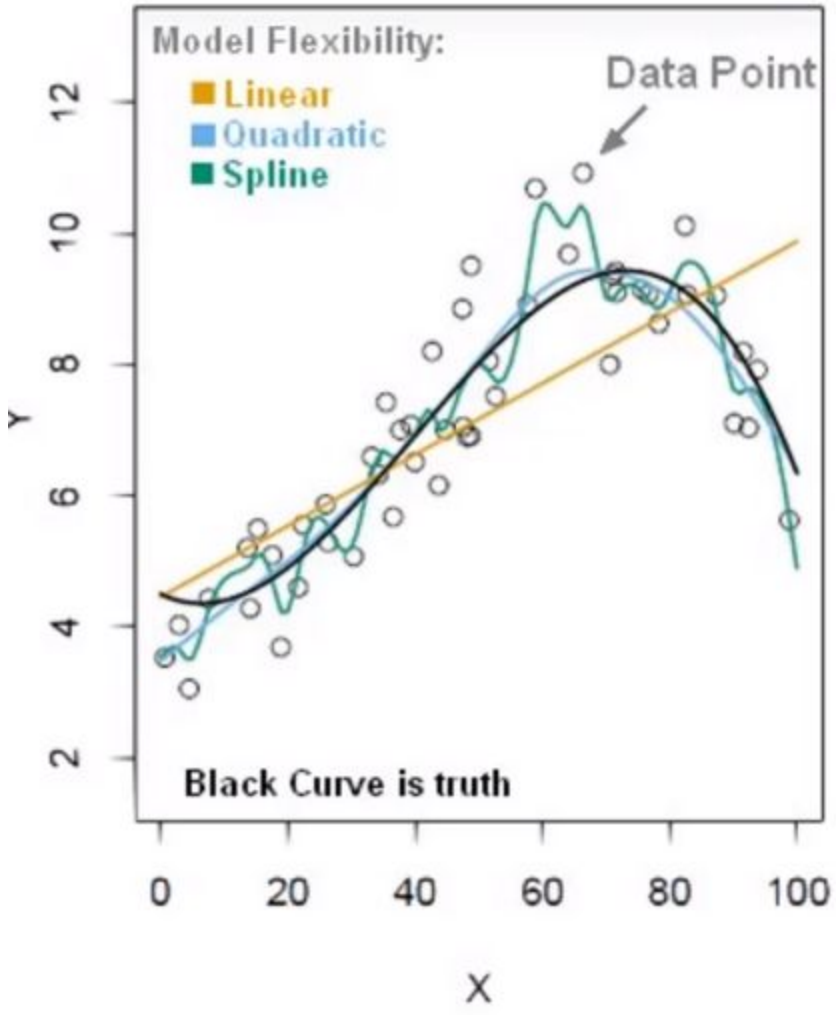


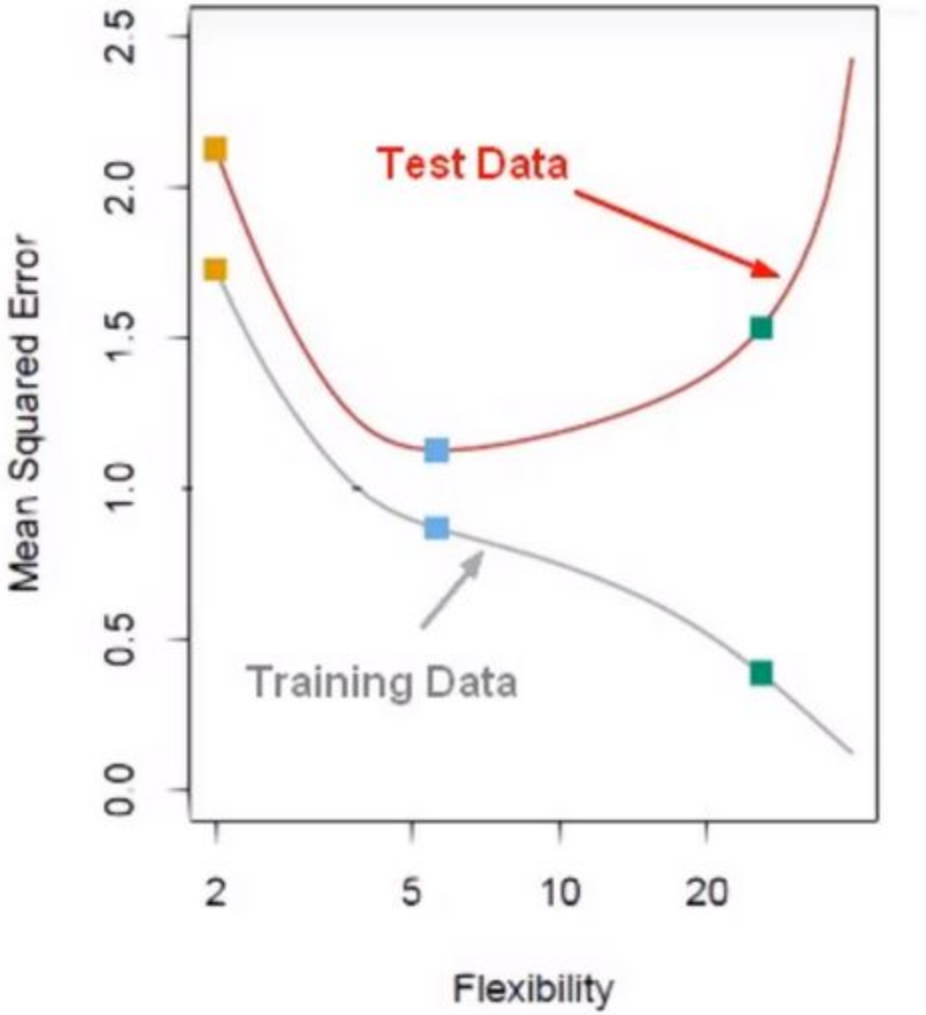
Смещение, разброс, переобучение и недообучение.

Для простых моделей характерно недообучение (они слишком простые, не могут описать целевую зависимость и имеют большое смещение), для сложных – переобучение (алгоритмов в модели слишком много, при настройке мы выбираем ту, которая хорошо описывает обучающую выборку, но из-за сильного разброса она может допускать большую ошибку на тесте).



Смещение, разброс, переобучение и недообучение.





Смещение, разброс, переобучение и недообучение.

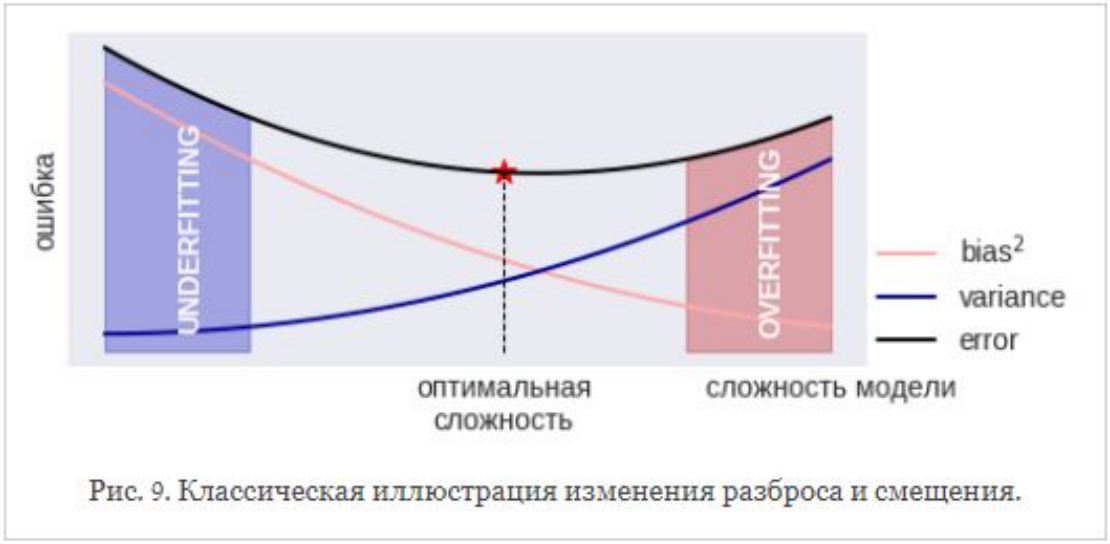


Рис. 9. Классическая иллюстрация изменения разброса и смещения.

Для простых моделей характерно недообучение (они слишком простые, не могут описать целевую зависимость и имеют большое смещение), для сложных – переобучение (алгоритмов в модели слишком много, при настройке мы выбираем ту, которая хорошо описывает обучающую выборку, но из-за сильного разброса она может допускать большую ошибку на тесте).

Смещение, разброс, переобучение и недообучение.

Литература

1. https://dyakonov.org/2018/04/25/%D1%81%D0%BC%D0%B5%D1%89%D0%B5%D0%BD%D0%B8%D0%B5-bias-%D0%B8-%D1%80%D0%B0%D0%B7%D0%B1%D1%80%D0%BE%D1%81-variance-%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D0%B8-%D0%B0%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82/