Машинное обучение

ФИО преподавателя: Оцоков Шамиль Алиевич

e-mail: shamil24@mail.ru

Типы машинного обучения

Индуктивное (по прецедентам) и дедуктивное. Некоторые методы индуктивного обучения были разработаны в качестве альтернативы классическим статистическим подходам. Индуктивное обучение основано на выявлении эмпирических закономерностей, дедуктивное — на формализации знаний экспертов и их использовании в качестве базы знаний. Первый тип характеризуется большим количеством данных и отсутствием или ненадобностью прошлого опыта. Второй тип обучения отличается малым массивом данных или выбором в пользу малых наборов данных, а также глубокими знаниями изучаемого вопроса.

Статистическая теория обучения

Статистическая теория обучения — это модель для обучения машин на основе статистики и функционального анализа. Статистическая теория обучения имеет дело с задачами нахождения функции предсказывания, основанной на данных..

В качестве более общего случая представьте, что мы наблюдаем некоторый количественный отклик Y и p отдельных предикторов $X_1, X_2, \ldots, X_p$. Мы делаем предположение о том, что существует определенная связь между Y и $X = (X_1, X_2, \ldots, X_p)$, которую в очень общей форме можно записать как

Статистическая теория

$$Y = f(X) + \epsilon.$$

ошибка

Здесь f — это некоторая фиксированная, но не известная функция от $X_1, \ldots, X_p$, а ϵ — *ошибка*, которая не зависит от X и имеет нулевое среднее значение. В таком представлении f выражает *систематическую* информацию о Y, содержащуюся в X.

Предсказание

Во многих ситуациях набор входных переменных X легко доступен, однако получить выходную переменную Y не так просто. Благодаря тому, что ошибки имеют нулевое среднее значение, при таком сценарии мы можем предсказать Y с помощью

$$\hat{Y} = \hat{f}(X),$$

где $\hat{f}$ представляет собой нашу оценку f, а $\hat{Y}$ — предсказанное значение Y.

E

устранимая и неустранимая ошибки

Точность $\hat{Y}$ в качестве предсказанного значения Y зависит от двух величин, которые мы будем называть *устранимой ошибкой* и *неустранимой ошибкой*. Обычно $\hat{f}$ не будет идеальной оценкой f, и эта неточность приведет к возникновению некоторой ошибки. Такая ошибка является *устранимой*, поскольку потенциально мы можем улучшить точность $\hat{f}$, используя более подходящий статистический метод для оценивания f. Но даже если бы имелась возможность достичь настолько идеальной оценки f, что $\hat{Y} = f(X)$, наше предсказанное значение все равно содержало бы в себе некоторую ошибку! Подобная ошибка известна как *неустранимая*, поскольку как бы хорошо мы ни оценили f, мы не сможем снизить ошибку, привнесенную за счет ϵ.

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{устранимая}} + \underbrace{\text{Var}(\epsilon)}_{\text{неустранимая}},$$

E

где $E(Y - \hat{Y})^2$ представляет собой среднее, или *ожидаемое*, значение квадрата разности между предсказанным и истинным значением Y, а $\mathrm{Var}(\epsilon)$ — *дисперсию*, связанную с ошибкой ϵ.

Ак

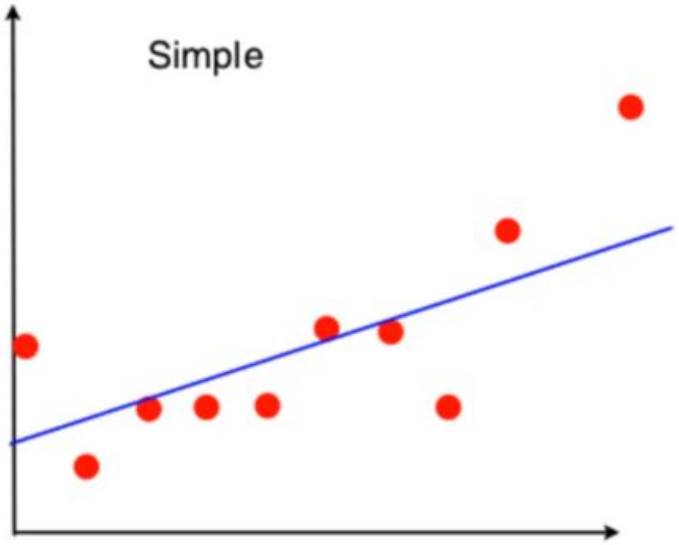
Разброс - характеризует разнообразие алгоритмов (из-за случайности обучающей выборки, в том числе шума, и стохастической природы настройки)
Смещение – способность модели алгоритмов настраиваться на целевую зависимость

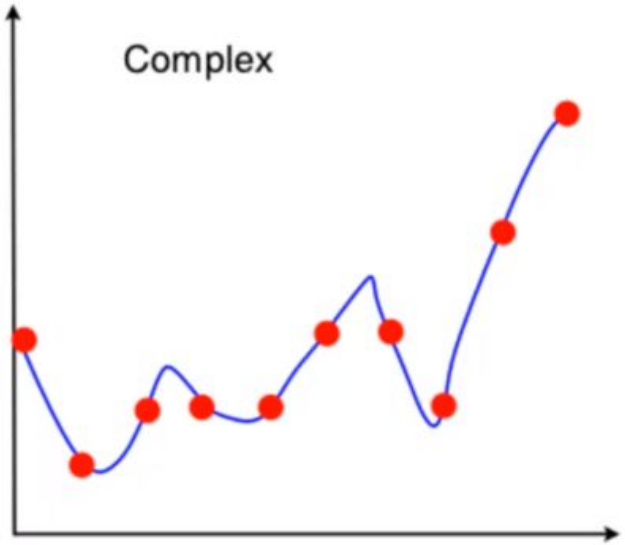
Смещение, разброс, переобучение и недообучение.

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.

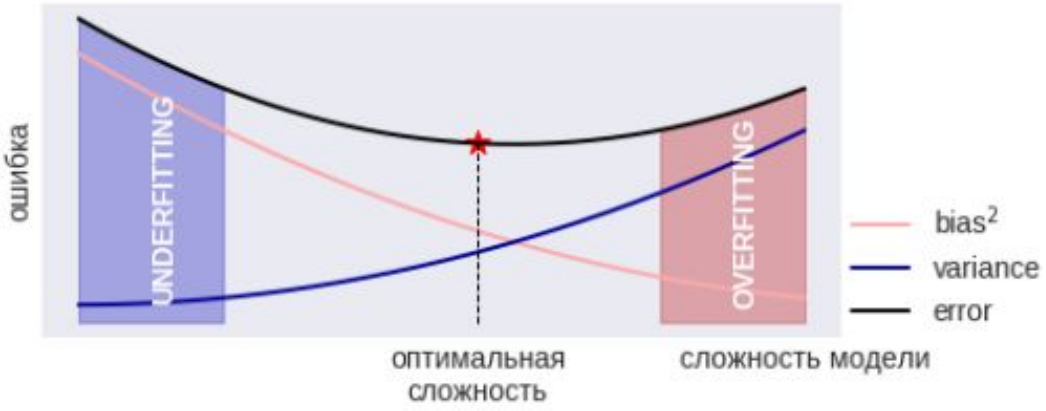


Рис. 9. Классическая иллюстрация изменения разброса и смещения.

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).

Статистический вывод

Часто мы заинтересованы в понимании того, как изменение $X_1, \ldots, X_p$ влияет на Y. В такой ситуации мы хотим оценить f, но наша цель не обязательно заключается в получении предсказаний для Y. Вместо этого мы хотим понять взаимоотношение между X и Y или, более конкретно, понять функциональную связь между Y и $X_1, \ldots, X_p$. В этом случае $\hat{f}$ нельзя рассматривать в качестве черного ящика, поскольку нам нужно знать ее точную форму. При таком сценарии мы можем быть заинтересованы в ответе на следующие вопросы:

- *Какие предикторы связаны с откликом?* Часто только небольшая часть имеющихся в распоряжении предикторов тесно связана с Y. В зависимости от стоящей задачи нахождение ограниченного числа *важных* предикторов в большом наборе возможных переменных может оказаться чрезвычайно полезным.

- *Какова связь между откликом и каждым предиктором?* Некоторые предикторы могут иметь положительную связь с Y в том смысле, что увеличение предиктора вызывает возрастание значений Y. Другие предикторы могут оказывать противоположный эффект. В зависимости от сложности f связь между откликом и некоторым предиктором может зависеть также от значений других предикторов.

Параметрические и непараметрические методы

пусть y_i обозначает отклик у i–го наблюдения. Обучающие данные тогда состоят из пар $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, где $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Наша цель заключается в применении некоторого метода статистического обучения к входным данным для нахождения неизвестной функции f. Другими словами, мы хотим найти такую функцию $\hat{f}$, что $Y \approx \hat{f}(X)$ для любого наблюдения (X, Y). В общих чертах большинство методов статистического обучения для решения этой задачи можно разделить на *параметрические* и *непараметрические*.

Параметрические и непараметрические методы

1. Во–первых, мы делаем некоторое предположение о функциональной форме f. Например, одно из простых предположений заключается в том, что f является линейной функцией от X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Вместо нахождения совершенно произвольной p–мерной функции $f(X)$ нам нужно будет оценить лишь $p + 1$ коэффициентов $\beta_0, \beta_1, \ldots, \beta_p$.

2. После выбора модели нам потребуется процедура, которая использует обучающие данные для *подгонки*, или *обучения*, модели. В случае линейной модели нам необходимо оценить параметры $\beta_0, \beta_1, \ldots, \beta_p$. Другими словами, мы хотим найти такие значения этих параметров, при которых

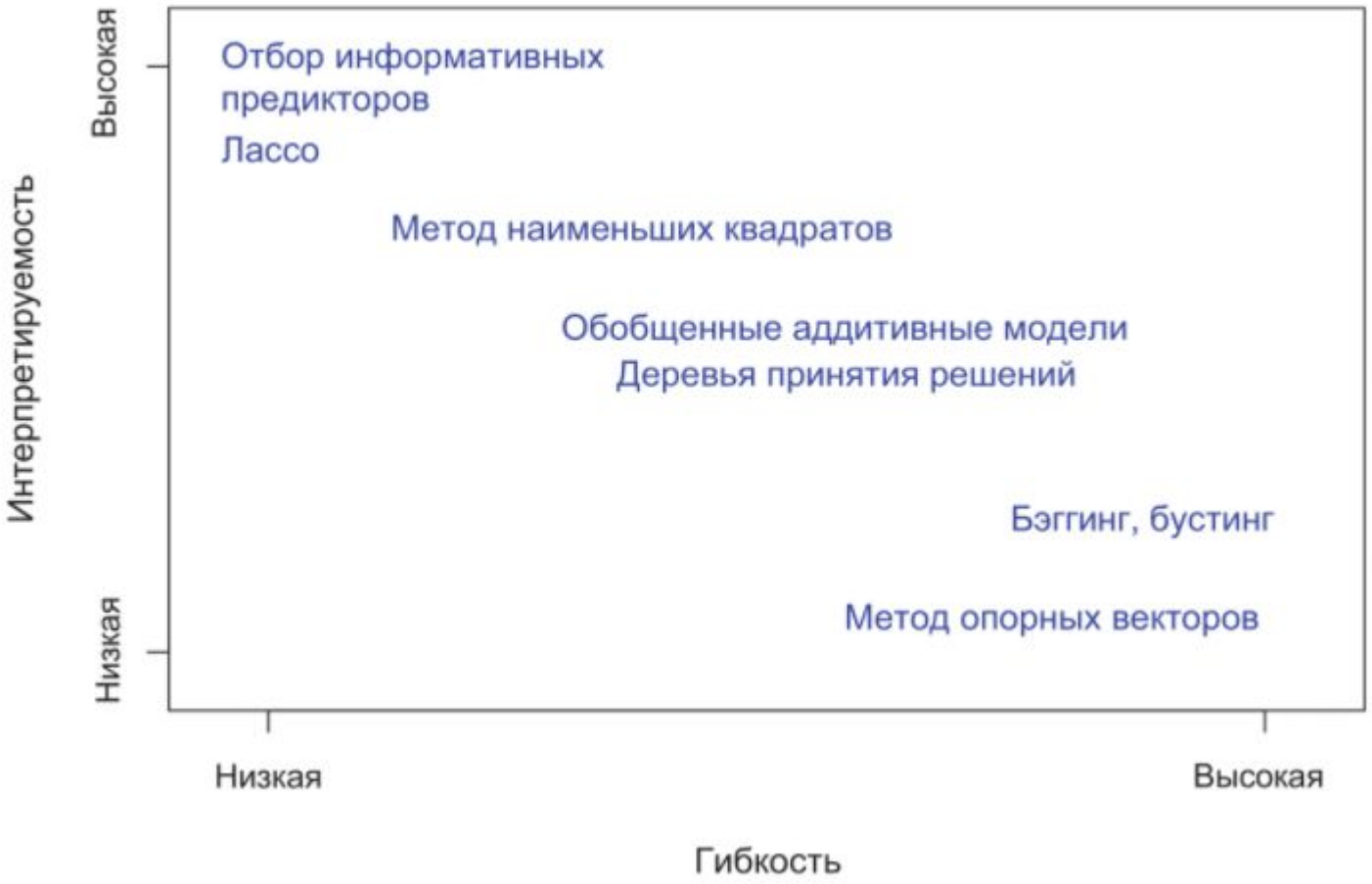
$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Параметрические и непараметрические методы

Непараметрические методы

Непараметрические методы не делают явных предположений в отношении функциональной формы f. Вместо этого они выполняют поиск такой оценки f, которая приближается к данным максимально близко, не будучи при этом слишком грубой или слишком извилистой. Такие подходы могут иметь существенное преимущество по сравнению с параметрическим методами: избегая предположения о конкретной функциональной форме f, они способны точно описать более широкий ряд возможных форм f. Любой параметрический подход несет с собой возможность того, что используемая для оценивания f функциональная форма значительно отличается от истинной функции, вследствие чего итоговая модель будет плохо описывать данные. В то же время непараметрические методы полностью уходят от этой опасности, поскольку, по сути, не делается никакого предположения о форме f. Однако непараметрические методы страдают существенным недостатком: поскольку они не сводят проблему оценивания f к ограниченному набору параметров, то для получения точной оценки f требуется очень большое число наблюдений (намного больше, чем обычно необходимо для какого–либо параметрического метода).

Параметрические и непараметрические методы

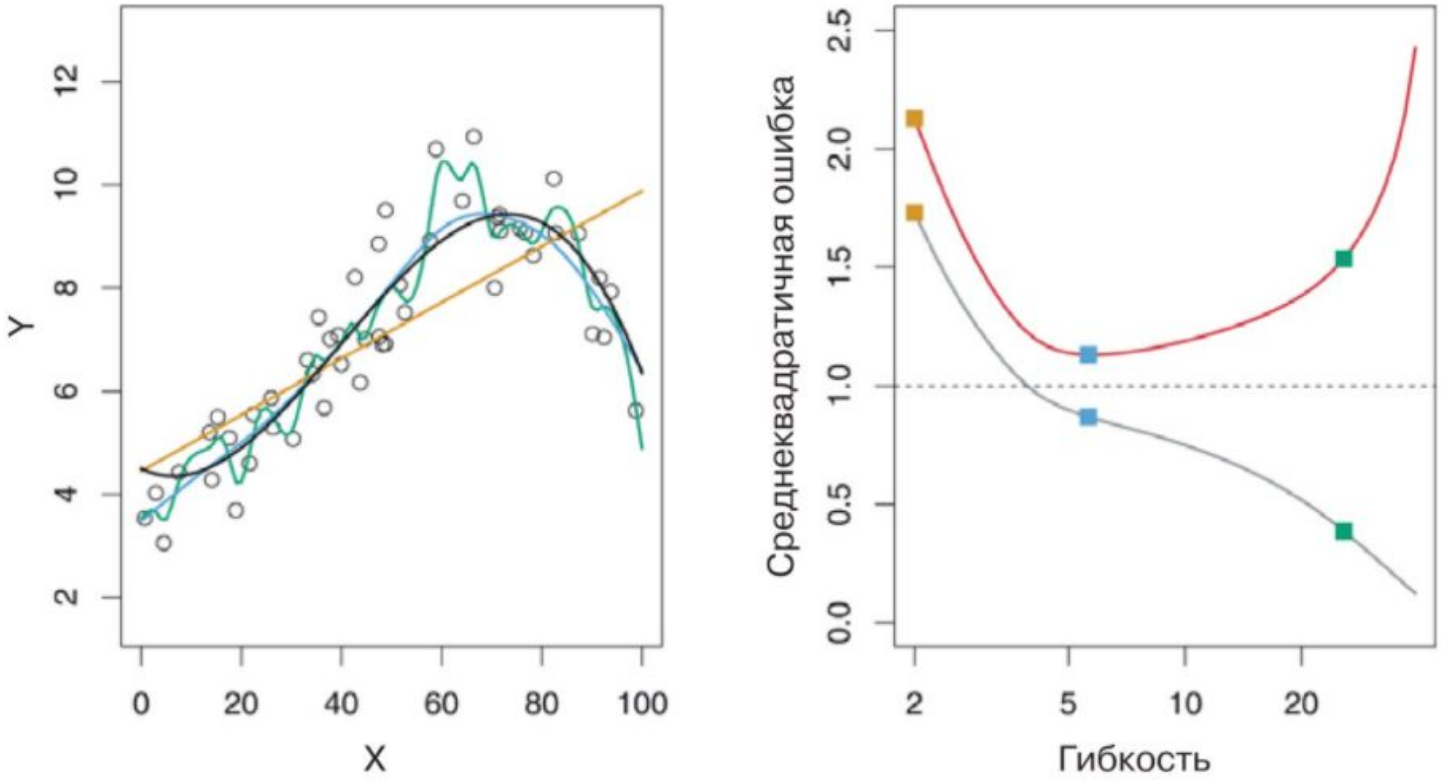


Компромисс между смещением и дисперсией

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + Var(\epsilon).$$

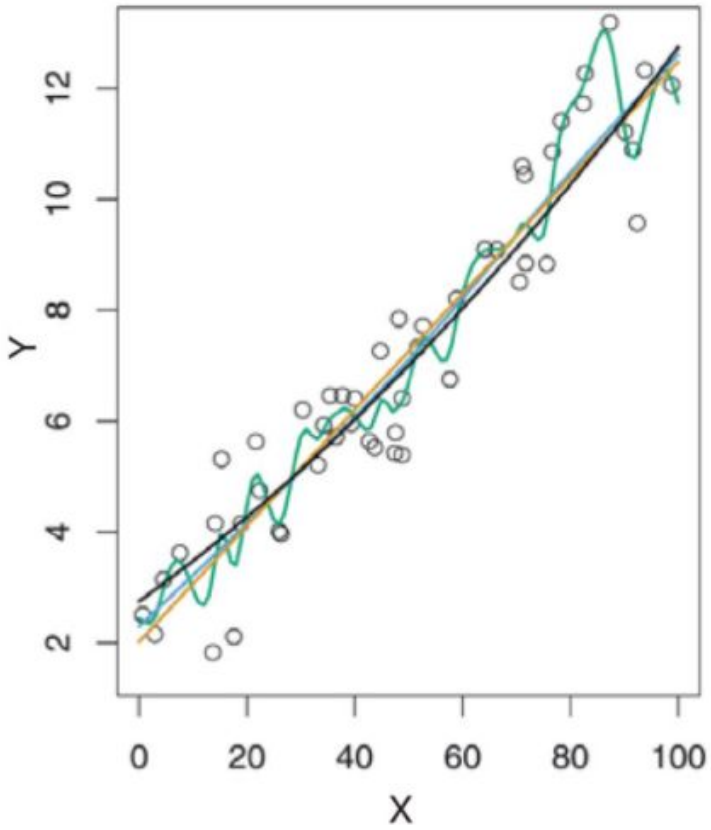
Здесь $E\left(y_0 - \hat{f}(x_0)\right)^2$ обозначает *математическое ожидание* ошибки на контрольной выборке и представляет собой среднее значение MSE, которое мы получили бы при многократном повторном оценивании f на основе большого числа обучающих выборок и вычислении ошибки для каждого контрольного значения x_0. Общую ожидаемую MSE на контрольной выборке можно вычислить путем усреднения $E\left(y_0 - \hat{f}(x_0)\right)^2$ для всех возможных проверочных значений x_0.

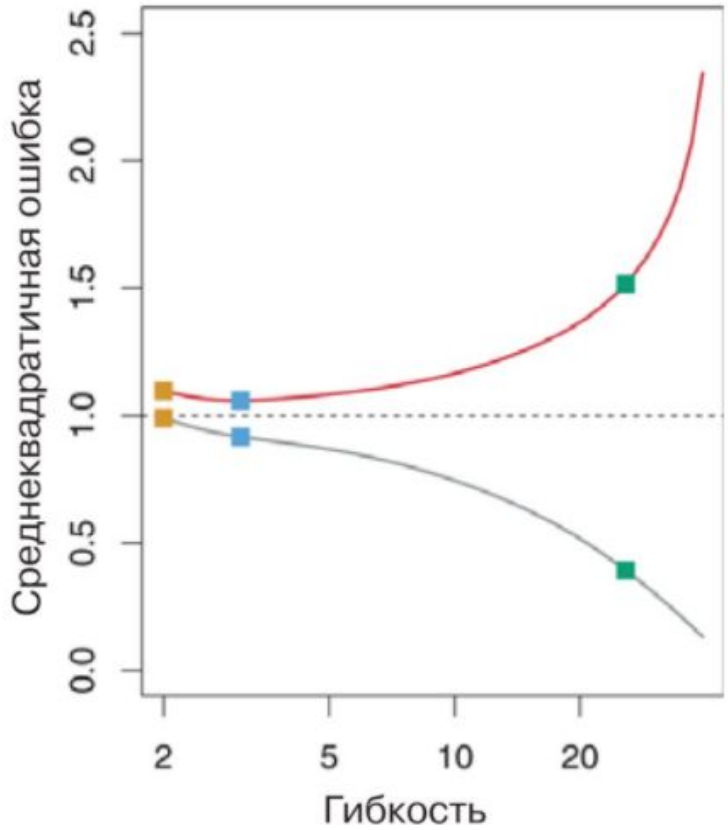
Компромисс между смещением и дi



Слева: *данные, имитированные на основе f, показаны полыми точками черного цвета. Представлены три способа подгонки f: линейная регрессия (оранжевая линия) и две модели гладких сплайнов (голубая и зеленая линии).* Справа: *среднеквадратичная ошибка на обучающих (серая линия) и контрольных данных (красная линия), а также минимально возможное значение ошибки для контрольных данных (пунктирная линия). Квадратные символы соответствуют ошибкам на обучающих и контрольных выборках, которые были получены для трех*

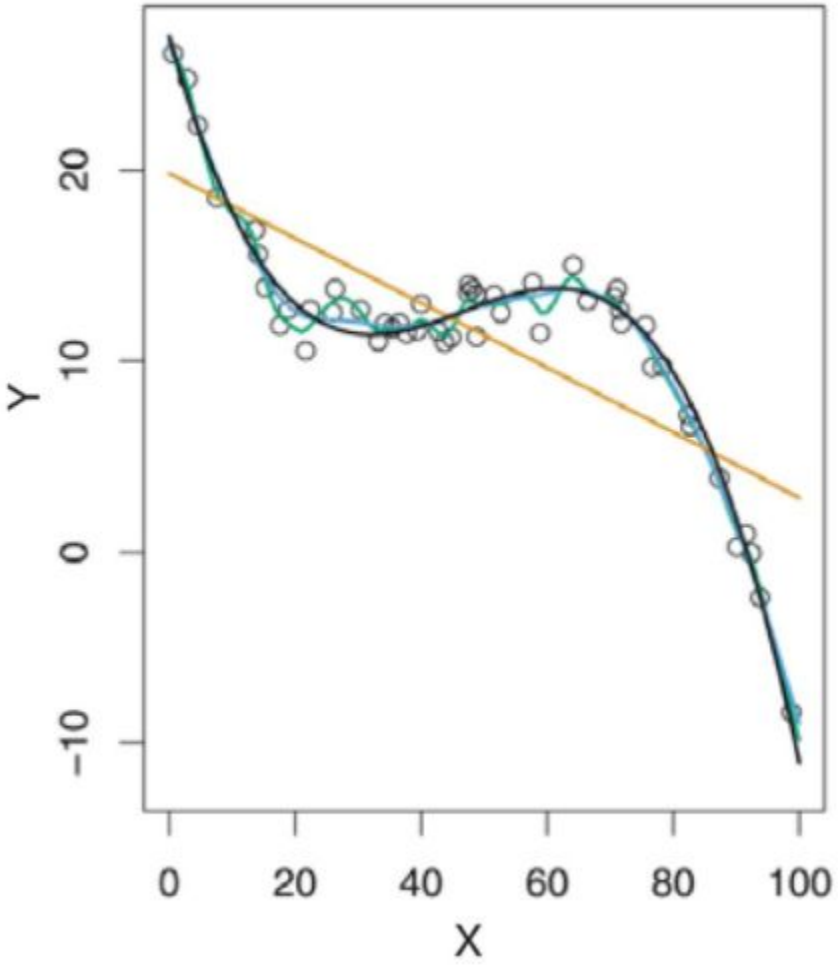
Компромисс между смещением и дисперсией

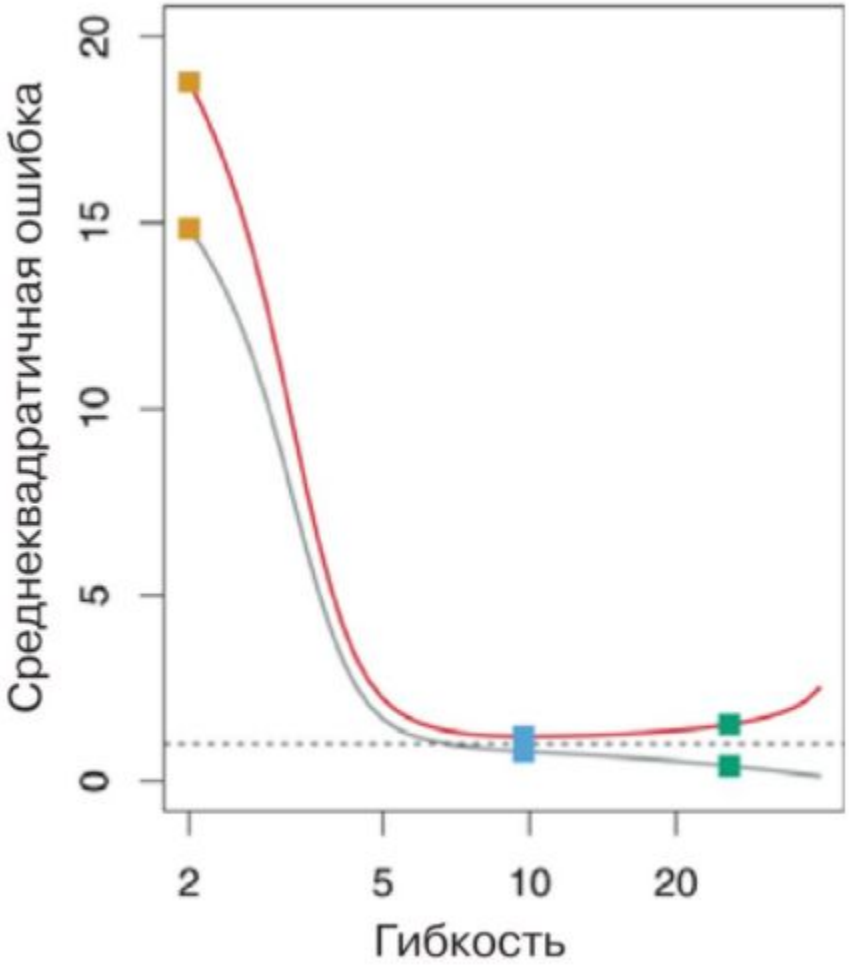




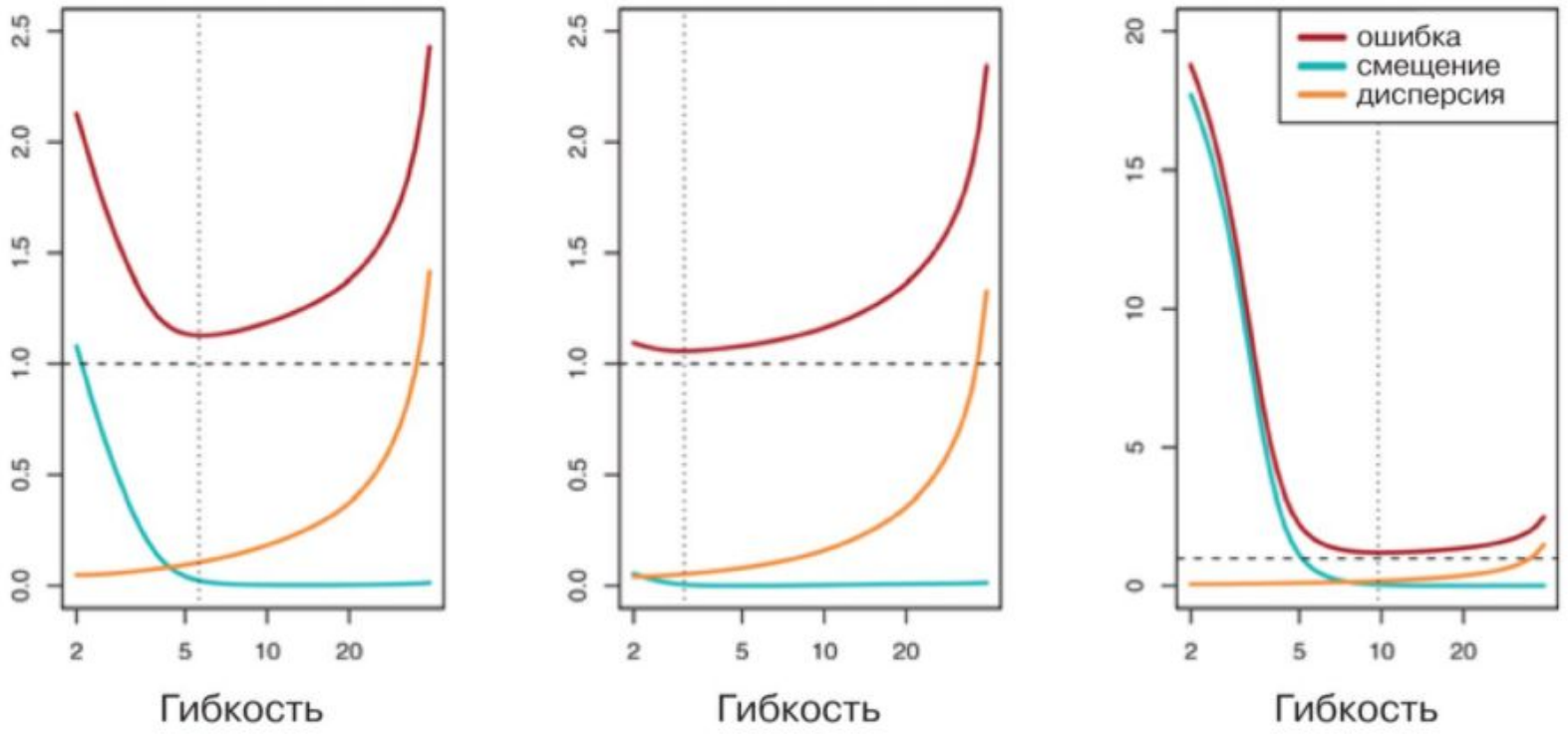
с истинной функцией f, которая намного ближе к линейной. В этой ситуации линейная регрессия очень хорошо описывает данные

Истинная функция существенно отличается от линейной





Истинная функция существенно отличается от линейной



Квадрат смещения (голубая кривая), дисперсия (оранжевая кривая), Var(ϵ) (пунктирная линия) и MSE на контрольной выборке (красная кривая) для трех наборов данных, показанных на рис. 2.9–2.11. Вертикальная пунктирная линия показывает уровень гибкости, соответствующий наименьшей MSE

Степени обученности модели

Недообученная модель

Модель, слишком сильно упрощающая закономерность $\mathcal{X} \to \mathcal{Y}$.

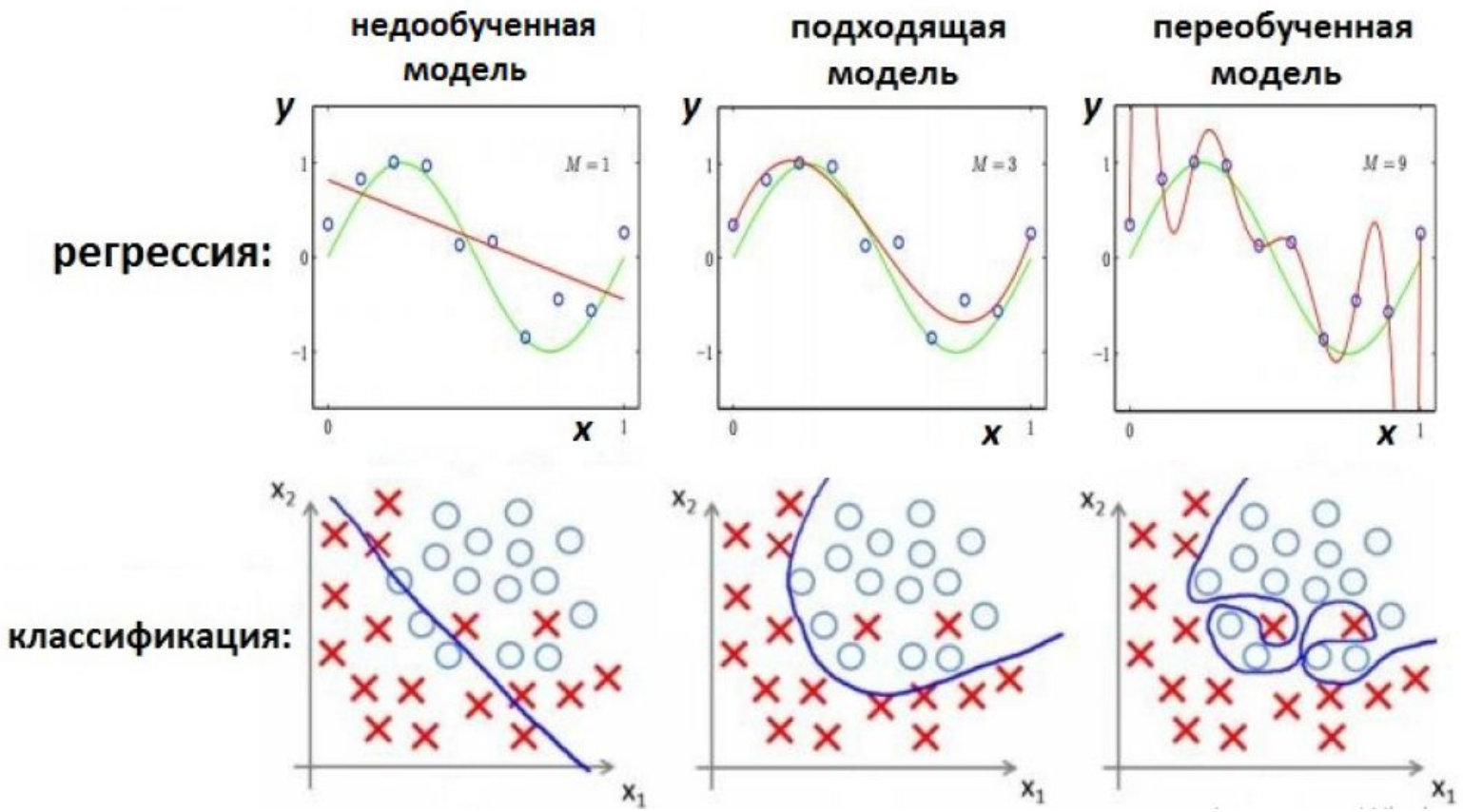
Переобученная модель

Модель, слишком сильно настроенная на особенности обучающей выборки (на шум в наблюдениях), а не на реальную закономерность $\mathcal{X} \to \mathcal{Y}$.

Ситуацию, когда некоторый метод обеспечивает небольшую MSE на обучающих данных и высокую MSE на проверочных данных, называют *переобучением* модели. Это происходит потому, что наша процедура статистического обучения слишком усердно пытается найти закономерности в обучающих данных и в результате может обнаружить некоторые закономерности, которые просто случайны и никак не связаны с истинными свойствами неизвестной функции f. При переобучении модели MSE на контрольной выборке будет большой потому, что предполагаемые закономерности, найденные нашим методом в обучающих данных, в проверочных данных просто не существуют.

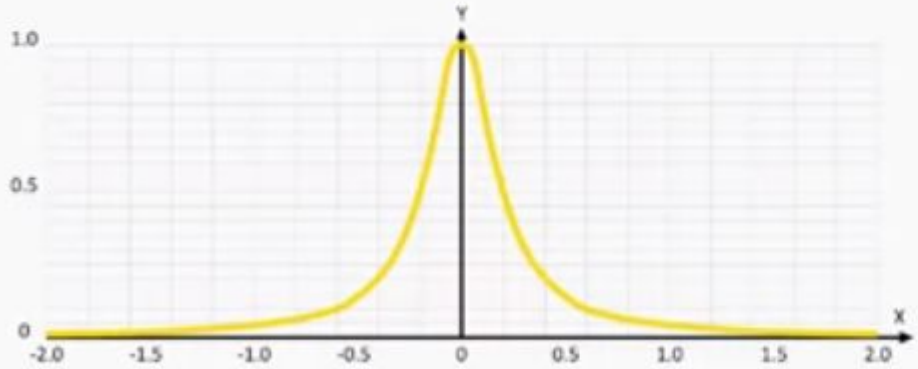
Примеры недообученных и переобученных моделей

истинная закономерность
оцененная закономерность полиномом степени M
объекты обучающей выборки



Пример. Переобучение полиномиальной регрессии

Зависимость $y(x) = \dfrac{1}{1+25x^2}$ на отрезке $x \in [-2, 2]$.



Признаковое описание $x \mapsto (1, x^1, x^2, \ldots, x^n)$.

Модель полиномиальной регрессии

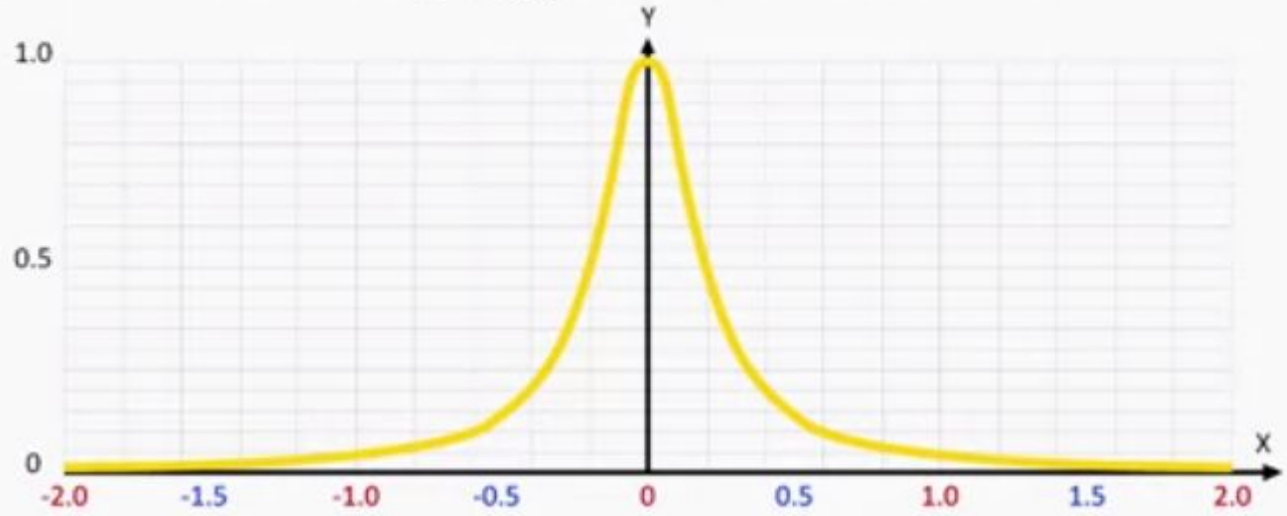
$$a(x, \theta) = \theta_0 + \theta_1 x + \cdots + \theta_n x^n \quad \text{— полином степени } n.$$

Обучение методом наименьших квадратов:

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \cdots + \theta_n x_i^n - y_i)^2 \to \min_{\theta_0, \ldots, \theta_n}.$$

Пример. Переобучение полиномиальной регрессии

Зависимость $y(x) = \dfrac{1}{1+25x^2}$ на отрезке $x \in [-2, 2]$.



Обучающая выборка:

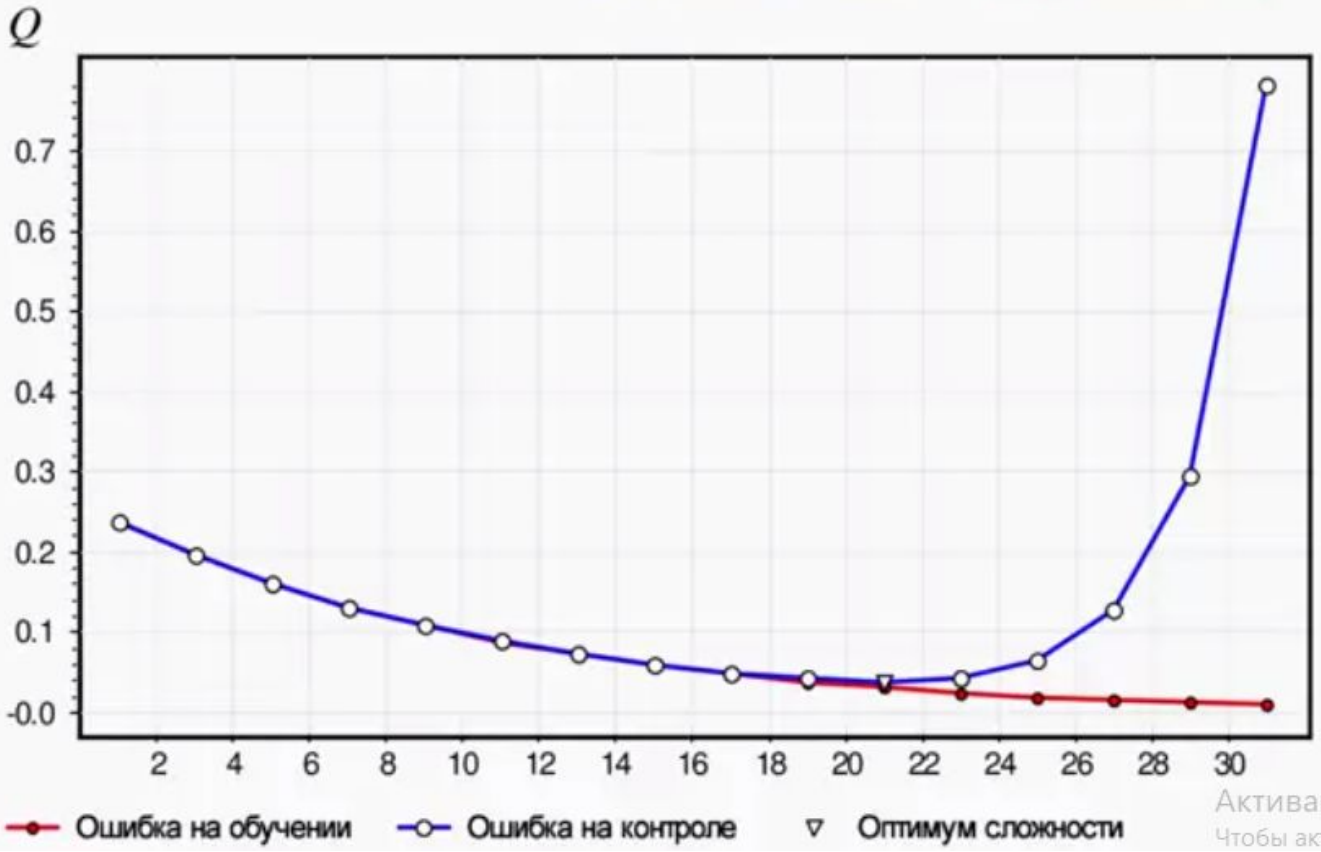
$$X^\ell = \left\{ x_i = 4\tfrac{i-1}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell \right\}.$$

Контрольная выборка:

$$X^k = \left\{ x_i = 4\tfrac{i-0.5}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell - 1 \right\}.$$

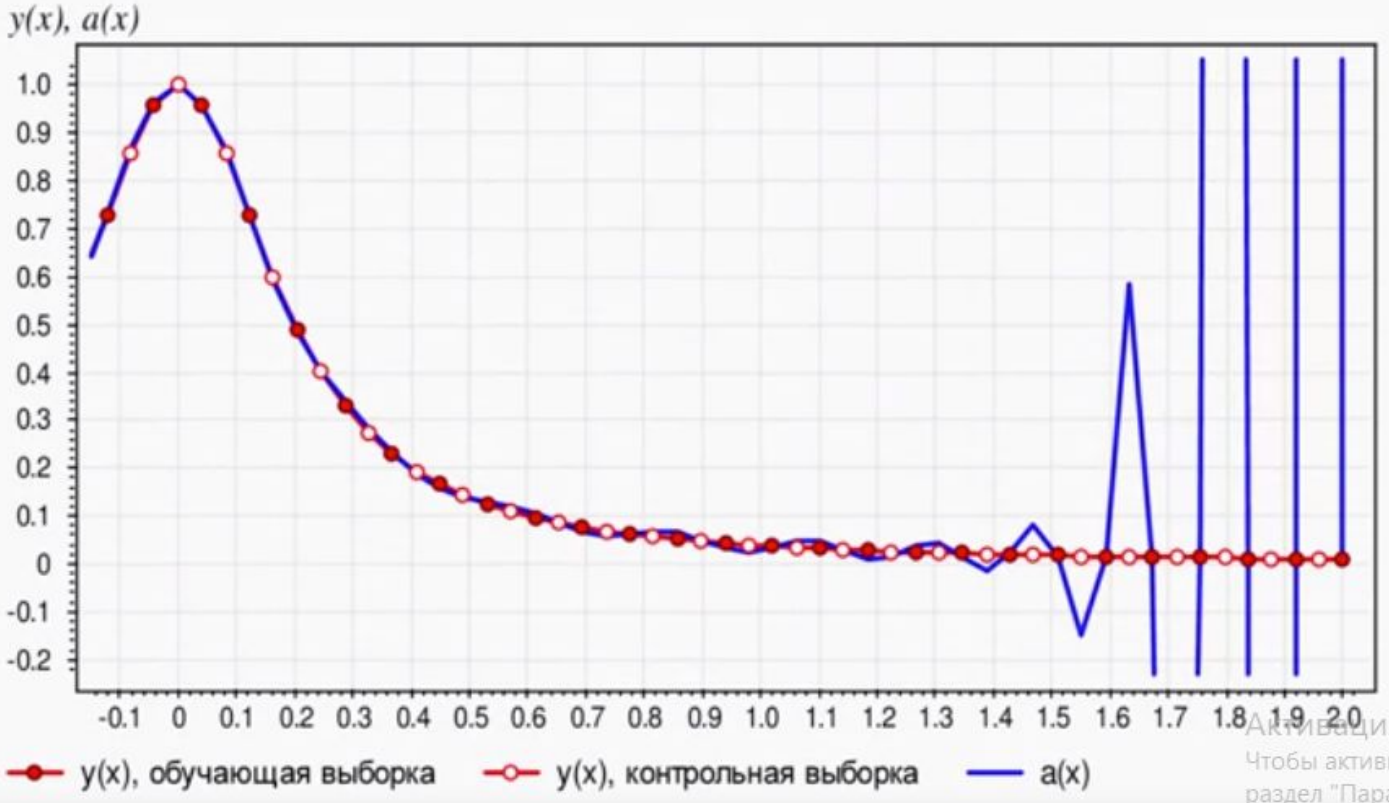
Пример переобучения: эксперимент при l = 50, n = 1,...,31

Переобучение — это когда $Q(\mu(X^\ell), X^k) \gg Q(\mu(X^\ell), X^\ell)$:



Пример переобучения: эксперимент при l = 50, n = 38

$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x) \text{ — полином степени } n = 38$$



Эмпирические оценки обобщающей способности

- Эмпирический риск на тестовых данных (hold-out):

$$\mathrm{HO}(\mu, X^\ell, X^k) = Q(\mu(X^\ell), X^k) \to \min$$

- Скользящий контроль (leave-one-out), $L = \ell + 1$:

$$\mathrm{LOO}(\mu, X^L) = \frac{1}{L}\sum_{i=1}^{L} \mathscr{L}(\mu(X^L \setminus \{x_i\}), x_i) \to \min$$

- Кросс-проверка (cross-validation) по N разбиениям, $X^L = X_n^\ell \sqcup X_n^k$, $L = \ell + k$:

$$\mathrm{CV}(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} Q(\mu(X_n^\ell), X_n^k) \to \min$$

Эксперименты на реальных данных

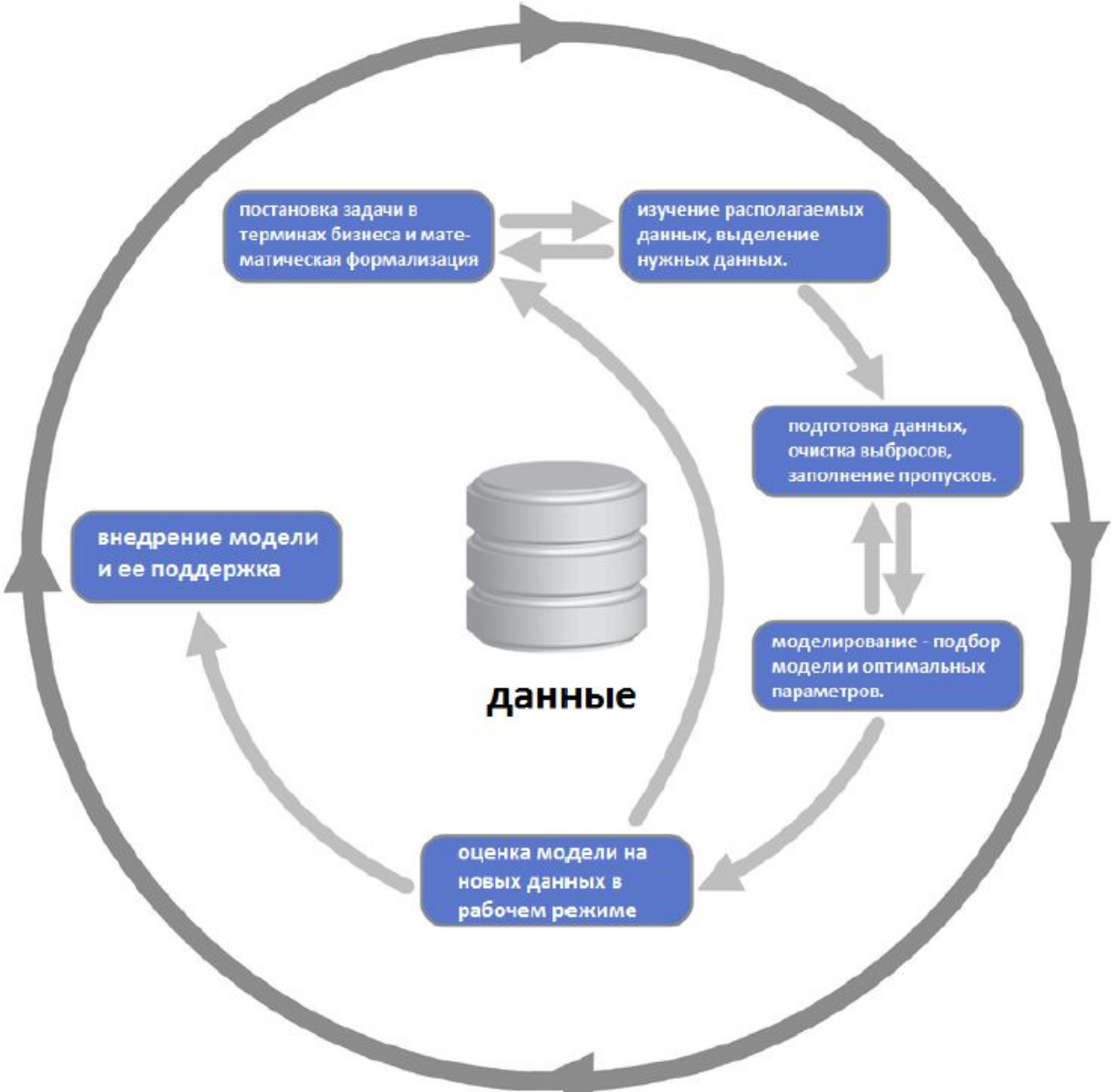
Эксперименты на конкретной прикладной задаче:

- цель — решить задачу как можно лучше
- важно понимание задачи и данных
- важно придумывать информативные признаки
- конкурсы по анализу данных: http://www.kaggle.com

Эксперименты на наборах прикладных задач:

- цель — протестировать метод в разнообразных условиях
- нет необходимости (и времени) разбираться в сути задач :(
- признаки, как правило, уже кем-то придуманы
- репозиторий UC Irvine Machine Learning Repository
  http://archive.ics.uci.edu/ml (308 задач, 09-02-2015)

Методология CrispDM



- **Прикладные задачи машинного обучения**
  встречаются во всех областях бизнеса, науки, производства

- **Особенности данных в прикладных задачах:**
  — разнородные (признаки измерены в разных шкалах);
  — неполные (измерены не все, имеются пропуски);
  — неточные (измерены с погрешностями);
  — противоречивые (объекты одинаковые, ответы разные);
  — избыточные (сверхбольшие, не помещаются в память);
  — недостаточные (объектов меньше, чем признаков);
  — неструктурированные (нет признаковых описаний);
  — нетривиальные критерии качества.

Этапы решения задач машинного обучения:

- понимание задачи и данных;
- предобработка данных и изобретение признаков;
- построение модели;
- сведение обучения к оптимизации;
- решение проблем оптимизации и переобучения;
- оценивание качества решения;
- внедрение и эксплуатация.

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)

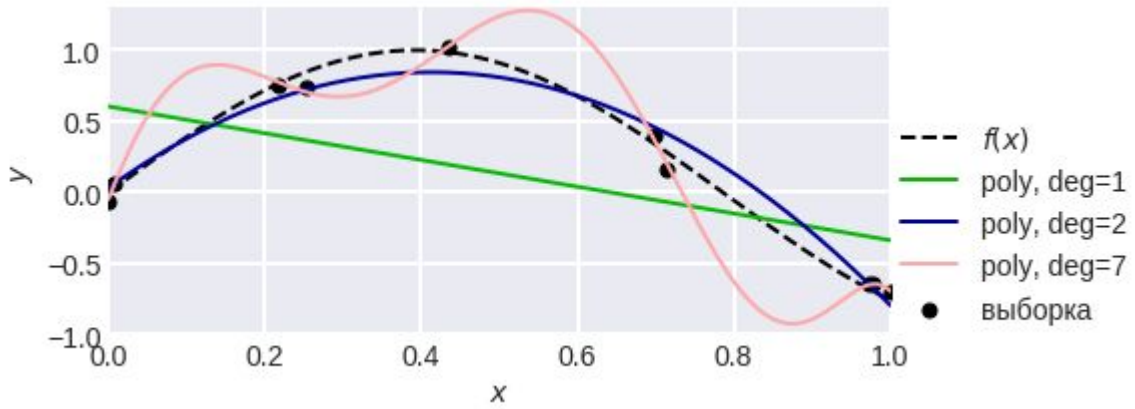
Недообучение (underfitting) – явление, когда ошибка на обучающей выборке достаточно большая, часто говорят «не удаётся настроиться на выборку». Такой странный термин объясняется тем, что недообучение при настройке алгоритмов итерационными методами (например, нейронных сетей методом обратного распространения) 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.

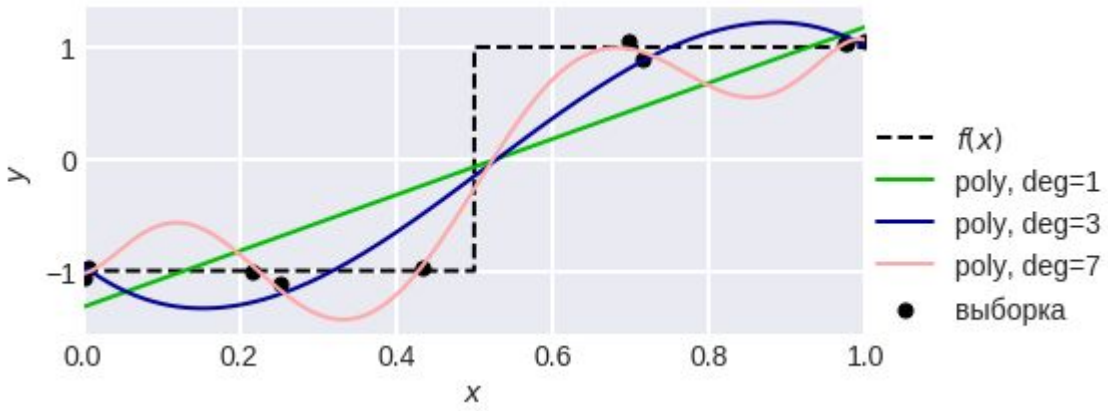
Пример переобучения.

$y = \sin(4x) + \text{шум}$

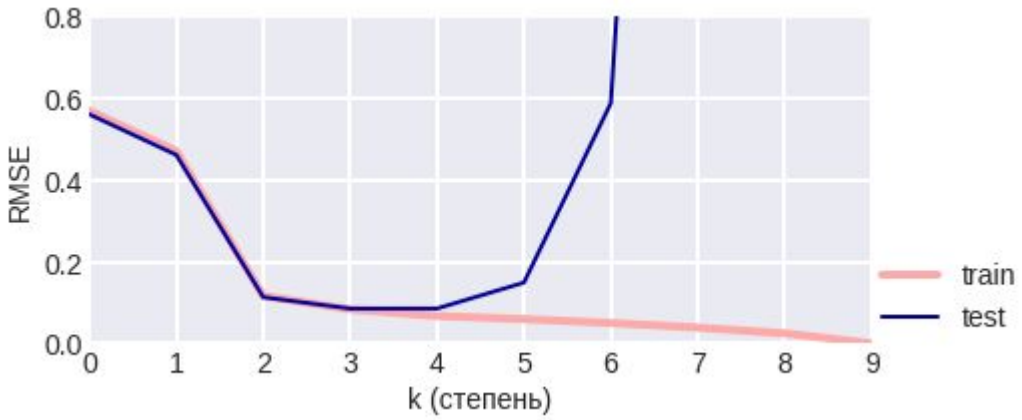


Смещение, разброс, переобучение и недообучение.

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.

Измерение качества модели через среднеквадратическое отклонение.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

где $\hat{f}(x_i)$ — это предсказанное значение, которое $\hat{f}$ дает для i–го наблюдения. MSE будет низкой, если предсказанные значения отклика очень близки к истинным значениям, и высокой, если для некоторых наблюдений предсказанные и истинные значения существенно разнятся.

Список литературы

1. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р . Введение в статистическое обучение с примерами на языке R

2. 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/

3. Грас Data Science. Наука о данных с нуля, 2017 г.

4. Введение в машинное обучение. https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie/lecture/CLOS0/formal-naia-postanovka-zadachi-mashinnogho-obuchieniia

Спасибо за внимание!