ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ УПРАВЛЕНИЯ

Основан в 1919 году

Методы и технологии машинного обучения

Лекция 1: Введение в статистическое обучение

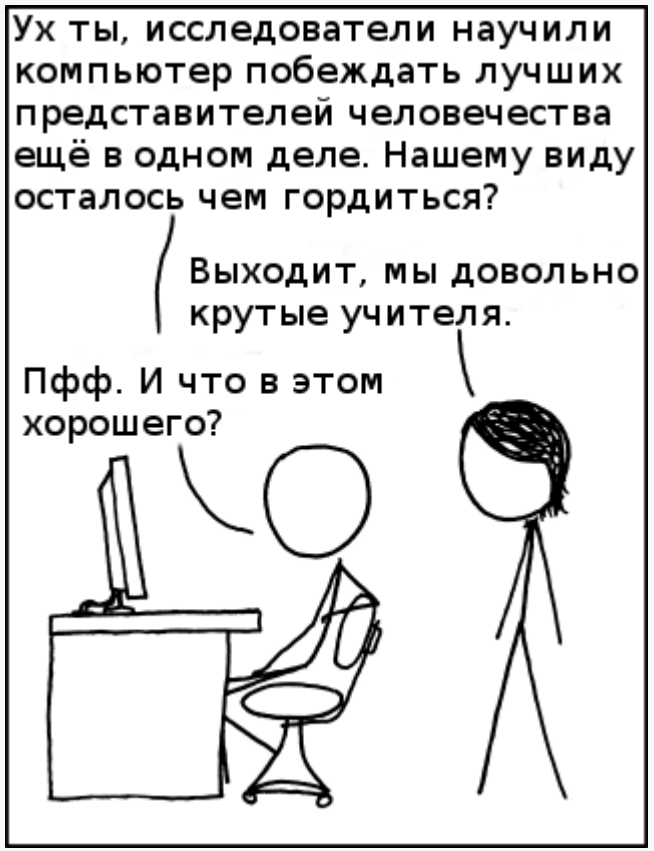
Светлана Андреевна Суязова (Аксюк)
s.aksuk@kiber-guu.ru

осенний семестр 2021 / 2022 учебного года

План лекции

- Что такое статистическое обучение

- Типы решаемых задач
- Непрерывный Y: понятия, связанные с точностью модели
- Дискретный Y: основные измерители точности
- Литература и ресурсы



xkcd.com/894

Под **статистическим обучением** понимают огромный набор инструментов, предназначенных для *понимания данных*.

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

Машинное обучение – область практической деятельности и сфера научных исследований алгоритмов, извлекающих смысл из данных.

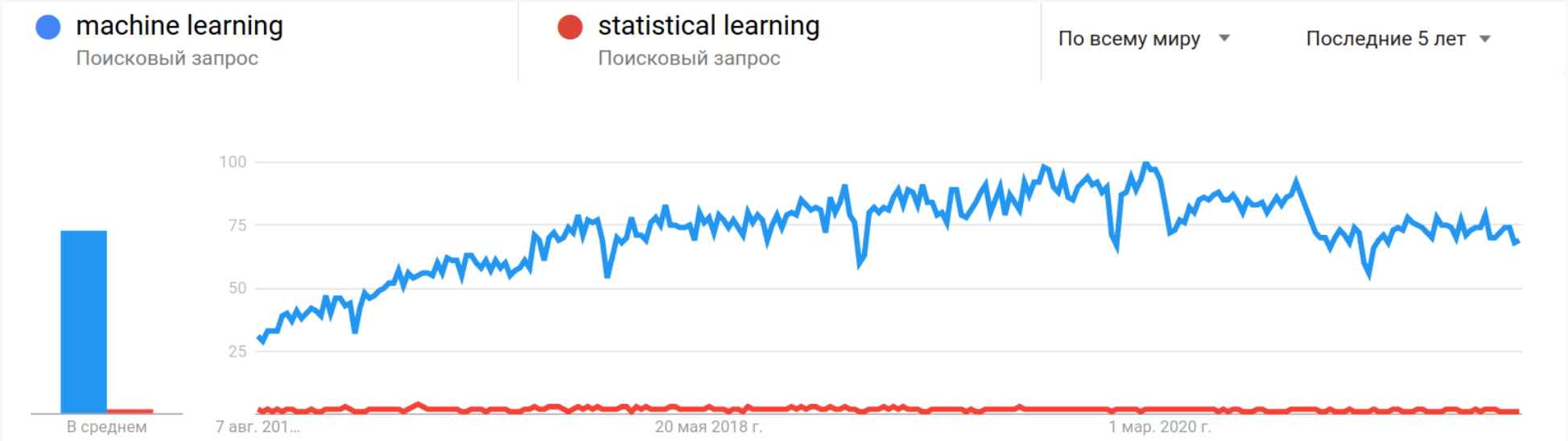
Рашка С., Python и машинное обучение

Машинное обучение – систематическое обучение алгоритмов и систем, в результате которого их знания или качество работы *возрастают с накоплением опыта*.

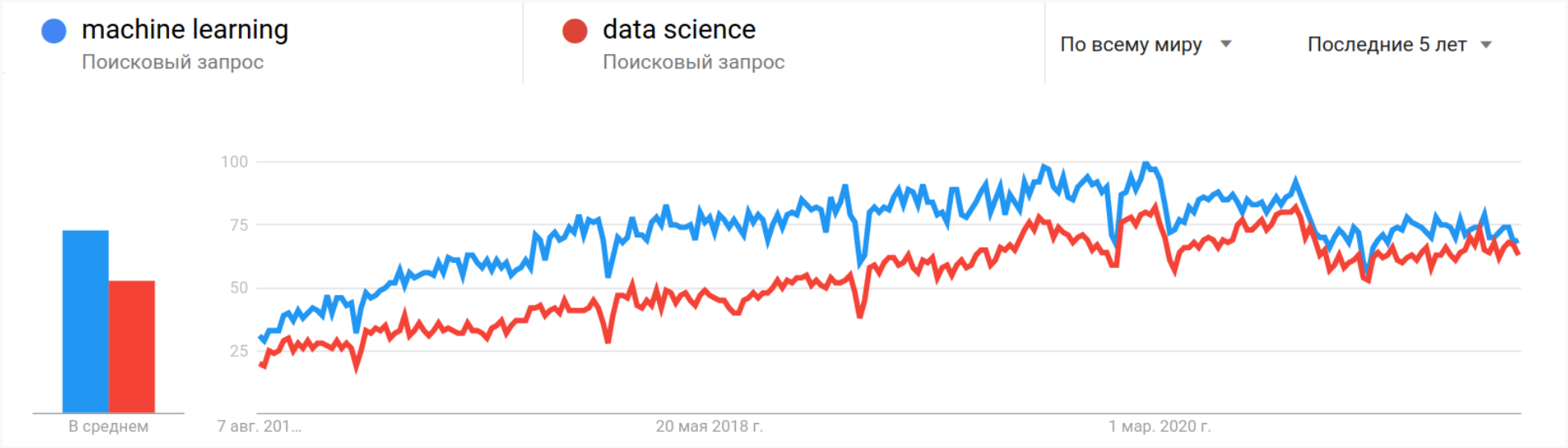
Флах П. Машинное обучение

Data Science (наука о данных) – междисциплинарная область, которая использует научные методы и алгоритмы для извлечения знаний из данных, представленных в различной форме, как структурированной, так и неструктурированной, аналогично *интеллектуальному анализу данных* (*Data Mining*).

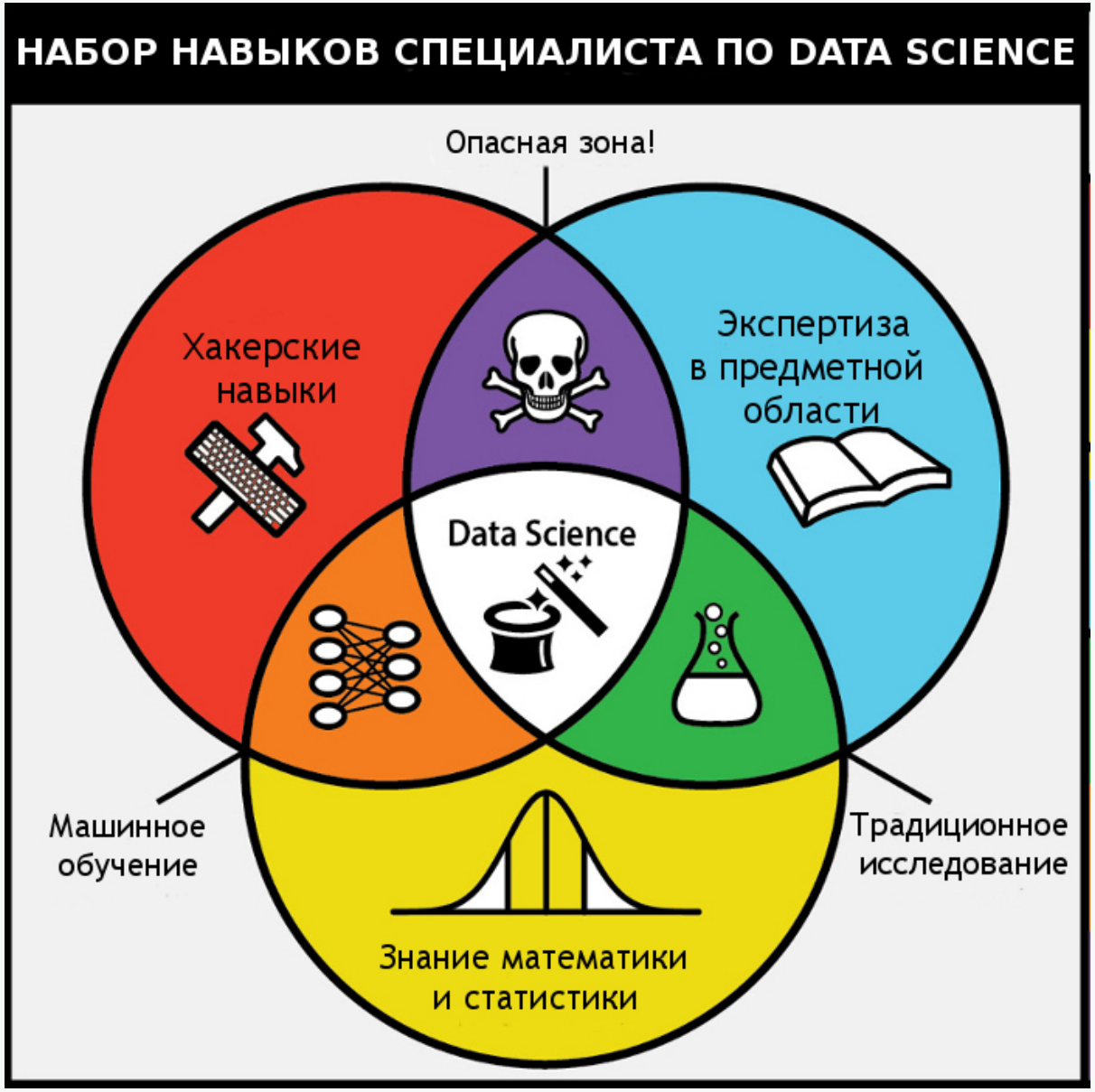
wiki



Google тренды: динамика запросов "machine learning" и "statistical learning"



Google тренды: динамика запросов "machine learning" и "data science"



Anna Schneider, How to become a data scientist before you graduate
drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Рабочий процесс машинного обучения

Предобработка

1. Сбор и подготовка данных. *Метки* – значения зависимой переменной, *признаки* – объясняющие переменные.

Обучение

1. Обучение модели на данных. *Модель* – отображение исходных данных на результаты.
2. Оценка качества и производительности модели. *Тестовая* выборка – часть данных с известными значениями меток, которая не использовалась при обучении модели.
3. Оптимизация модели. Подбор таких значений *параметров модели*, которые дают наименьшую ошибку на тестовой выборке.

Эксплуатация

1. Прогноз и/или интерпретация
2. Непрерывное совершенствование модели

Бринк Х., Ричардс Д., Феверолф М. Машинное обучение; Рашка С. Python и машинное обучение

Формальное описание данных

Число наблюдений: n. Число переменных: p.

X – матрица независимых переменных (признаков)

Y – вектор-столбец значений зависимой переменной (меток)

x_i – вектор i-ого наблюдения (строка X)

$\mathbf{x}_j$ – вектор j-ой переменной (столбец X)

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}, \; y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}.$$

Обучение и тестирование

Собственно построение модели:
обучающая выборка (`train`), n_{TRAIN} наблюдений

Оценка точности модели (и *ошибки вне выборки*):
тестовая выборка (`test`), n_{TEST} наблюдений

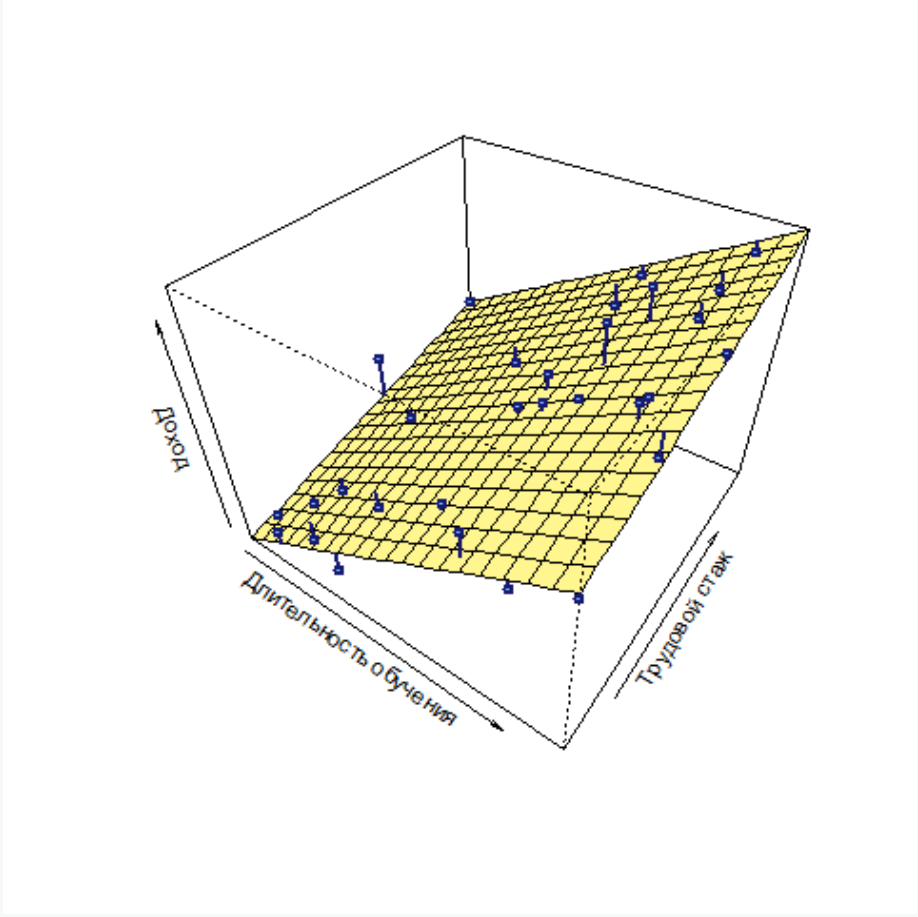
Модель: $Y = f(X) + \epsilon$,
где $f(X)$ – систематическая составляющая, ϵ – ошибка, не зависящая от X.

Параметрические методы рассматривают $f(X)$ как функцию и оценивают её параметры. **Непараметрические методы** не делают предположений о форме $f(X)$.

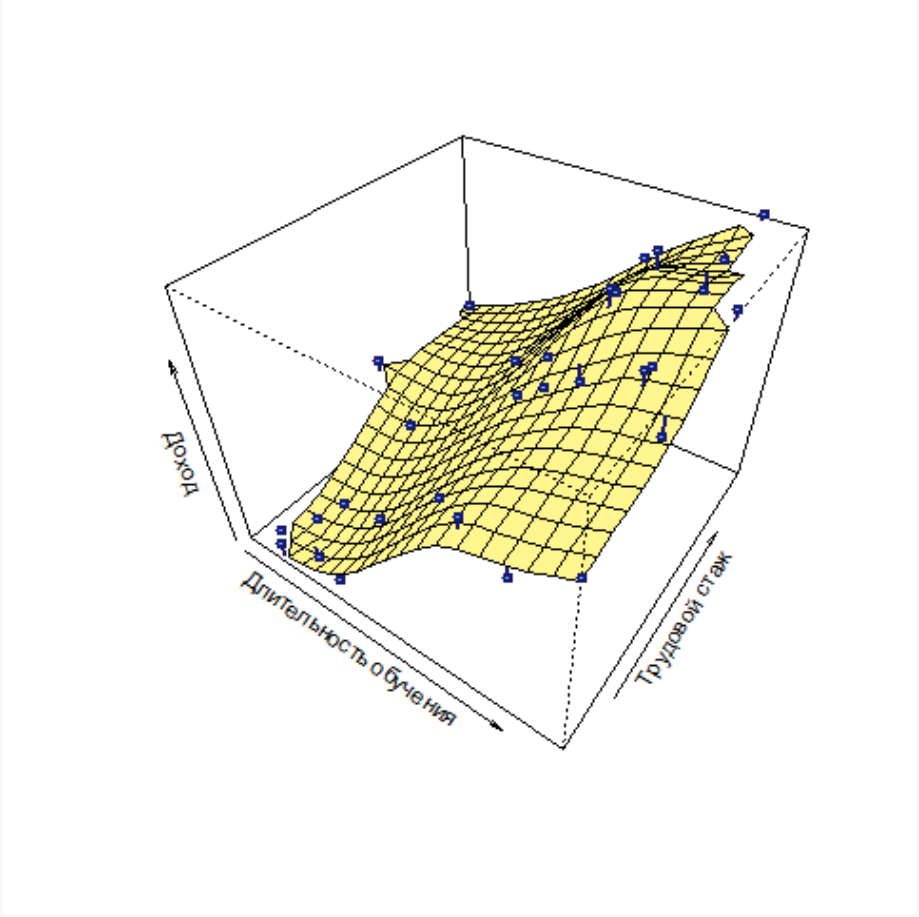
Сравнение подходов

Характеристика	Параметрические методы	Непараметрические методы
Возможность прогноза	есть	есть
Статистический вывод	есть	нет
Гибкость модели	регулируется формой функции	регулируется настроечными параметрами
Модель интерпретируема	почти всегда	почти никогда

Пример на данных `Income {ISLR}`



Параметрическая модель: линейная регрессия



Непараметрическая модель: сплайн "тонкая пластина"

План лекции

- Что такое статистическое обучение
- Типы решаемых задач
- Непрерывный Y: понятия, связанные с точностью модели
- Дискретный Y: основные измерители точности
- Литература и ресурсы

Классификация и регрессия

Y непрерывный (среднемесячный доход в рублях) – **задача регрессии**.

Y дискретный (уровень дохода: "высокий", "средний", "низкий") – **задача классификации**.

Большинство методов подходят для решения задач либо классификации, либо регрессии, но граница не жёсткая.

Регрессия (Р) или классификация (К)?

1. Прогноз и выявление факторов повторного инфаркта по показателям демографии, диеты, анализов. Зависимая переменная (Y): вероятность инфаркта.
2. Прогноз повторного инфаркта. Y: факт инфаркта.
3. Распознавание рукописных цифр. Y – шаблоны написания цифр.
4. Выявление в данных признаков, которые являются факторами риска возникновения определённого вида рака по клиническим и демографическим переменным. Y – вероятность болезни.
5. Прогноз и интерпретация рейтинга надёжности банка по финансовым показателям (методика рейтингования не задана, позиции шкалы рейтинга неизвестны).

Прогноз и статистический вывод

Предсказание Y для новых значений X – **задача прогноза**.

Интерпретация влияния признаков $\mathbf{x}_j$ на Y по модели Y – **задача статистического вывода**.

Статистический вывод может предшествовать построению модели – это применение техник преобразования пространства признаков на этапе предварительного анализа данных.

Прогноз (П) или статистический вывод (СВ)?

1. Прогноз и выявление факторов повторного инфаркта по показателям демографии, диеты, анализов. Зависимая переменная (Y): вероятность инфаркта.
2. Прогноз повторного инфаркта. Y: факт инфаркта.
3. Распознавание рукописных цифр. Y – шаблоны написания цифр.
4. Выявление в данных признаков, которые являются факторами риска возникновения определённого вида рака по клиническим и демографическим переменным. Y – вероятность болезни.
5. Прогноз и интерпретация рейтинга надёжности банка по финансовым показателям (методика рейтингования не задана, позиции шкалы рейтинга неизвестны).

Обучение с учителем и без учителя

Обучение с учителем (*supervised learning*): в обучающей выборке известны значения Y.

Обучение без учителя (*unsupervised learning*): значения Y неизвестны.

- кластерный анализ
- обобщение
- ассоциативные правила
- снижение размерности
- визуализация

Обучение с учителем (У) обучение или без учителя (БУ)?

1. Прогноз и выявление факторов повторного инфаркта по показателям демографии, диеты, анализов. Зависимая переменная (Y): вероятность инфаркта.
2. Прогноз повторного инфаркта. Y: факт инфаркта.
3. Распознавание рукописных цифр. Y – шаблоны написания цифр.
4. Выявление в данных признаков, которые являются факторами риска возникновения определённого вида рака по клиническим и демографическим переменным. Y – вероятность болезни.
5. Прогноз и интерпретация рейтинга надёжности банка по финансовым показателям (методика рейтингования не задана, позиции шкалы рейтинга неизвестны).

Выбор и применение метода обучения с учителем

- Нужен прогноз или статистический вывод?
- Зависимая переменная (Y) непрерывная или дискретная?

Этап	Параметрические	Непараметрические
1. Подготовка	выбрать форму функции	выбрать значения настроечных параметров
2. Подгонка	оценить параметры модели, проверить значимость	аппроксимировать обучающие данные
3. Валидация	оценить точность, проверить допущения	оценить точность

План лекции

- Что такое статистическое обучение
- Типы решаемых задач
- Непрерывный Y: понятия, связанные с точностью модели
- Дискретный Y: основные измерители точности
- Литература и ресурсы

Точность модели

Непрерывный Y

Ошибка на обучающей выборке:

$$MSE = \frac{1}{n_{TRAIN}} \sum_{i \in TRAIN} \left(y_i - \hat{f}\left(x_i\right) \right)^2$$

Ошибка на тестовой выборке:

$$MSE_{test} = \frac{1}{n_{TEST}} \sum_{i \in TEST} \left(y_i - \hat{f}\left(x_i\right) \right)^2$$

Данные для примера №1

Сгенерируем X и Y:

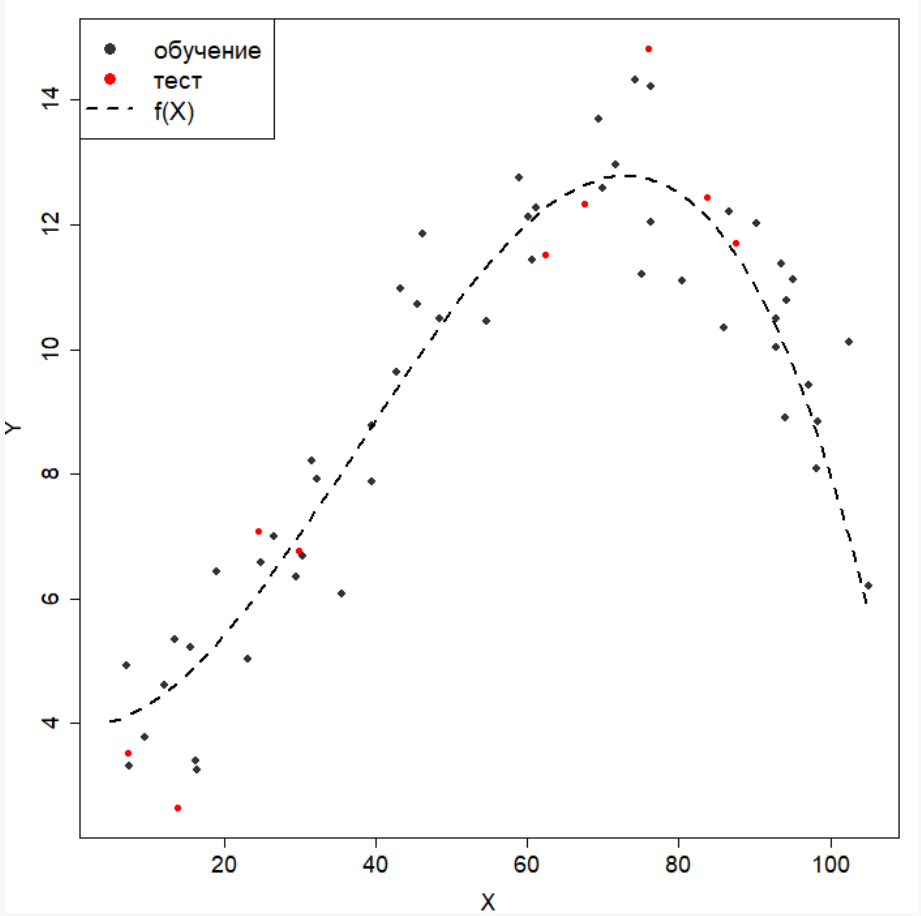
- $X \sim U(5, 105)$ (равномерное распределение)
- $Y = f(X) + \epsilon$, где

$$f(X) = 4 - 0.02X + 0.0055X^2 - \\ -4.9 \cdot 10^{-5} \cdot X^3$$

$\epsilon \sim N(0, 1)$ (нормальное распределение)

Количество наблюдений: 60.
Доля обучающей выборки: 85%.

Неустранимая ошибка: $Var(\epsilon) = 1$.

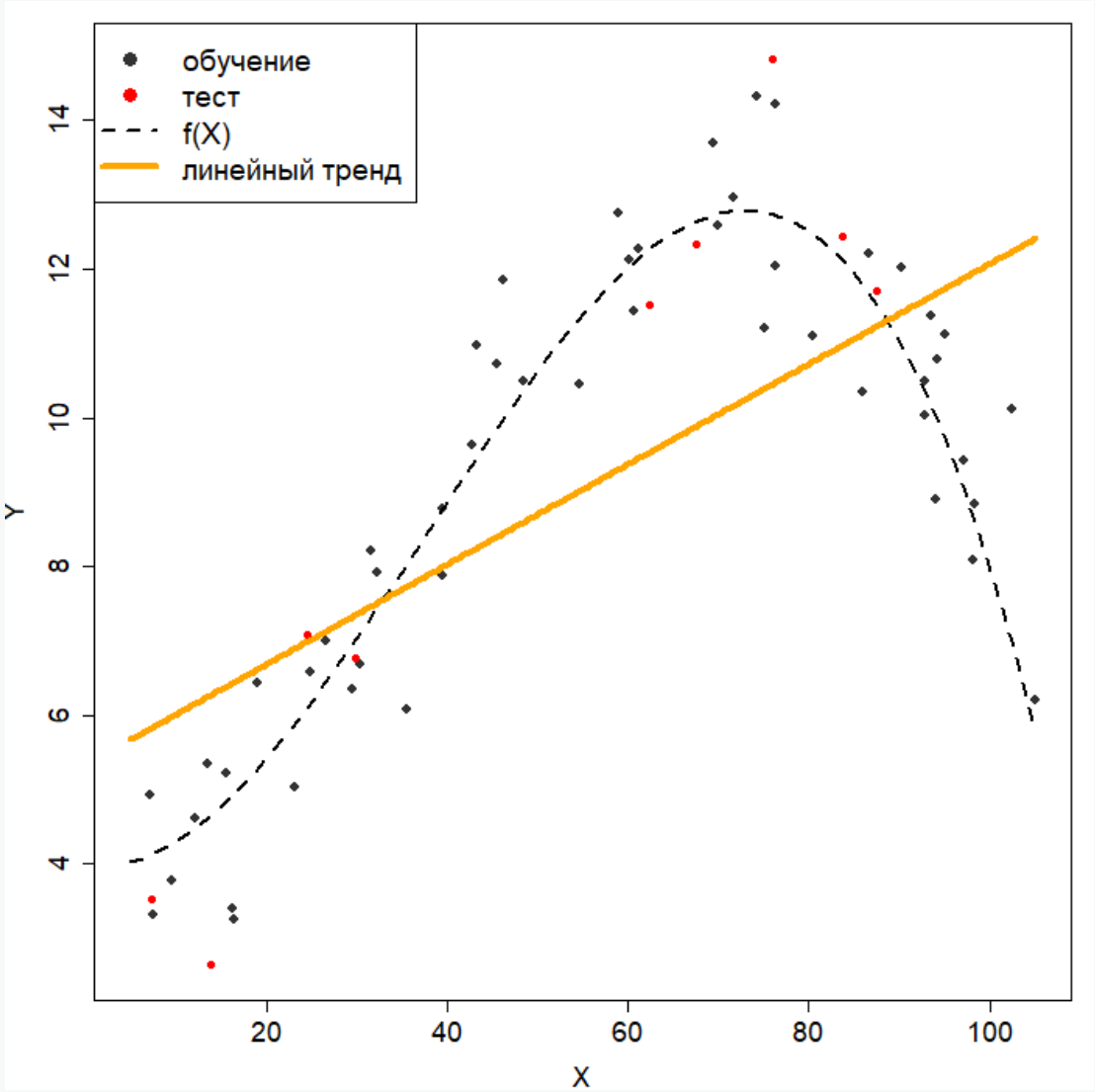


Гибкость модели ~ число степеней свободы

1. Линейная регрессия (2 степени свободы)
2. Сплайн с 6 степенями свободы
3. Сплайн с 38 степенями свободы

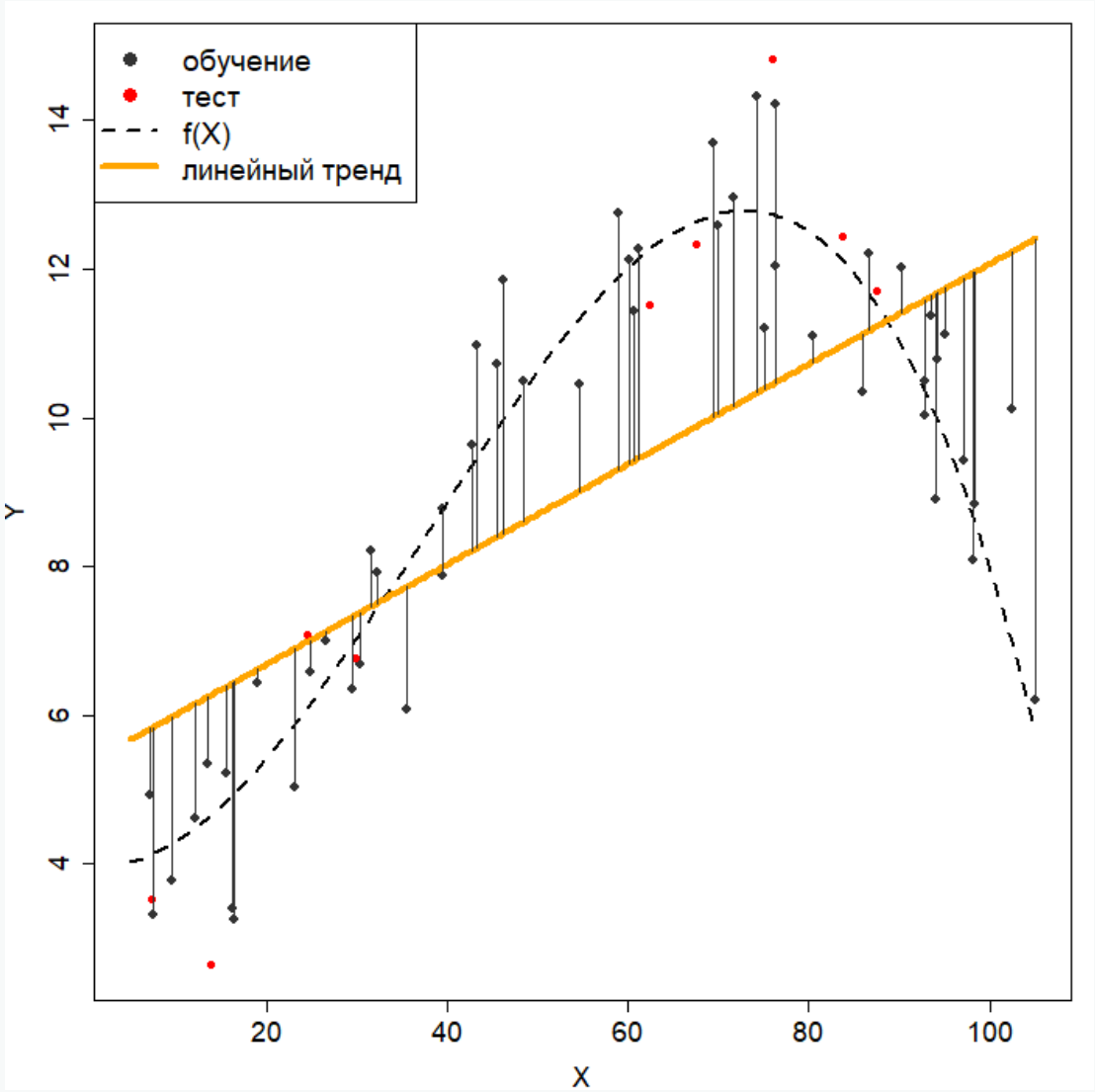
Рассмотрим практический пример №1.

1. Линейный тренд



1. Линейный тренд

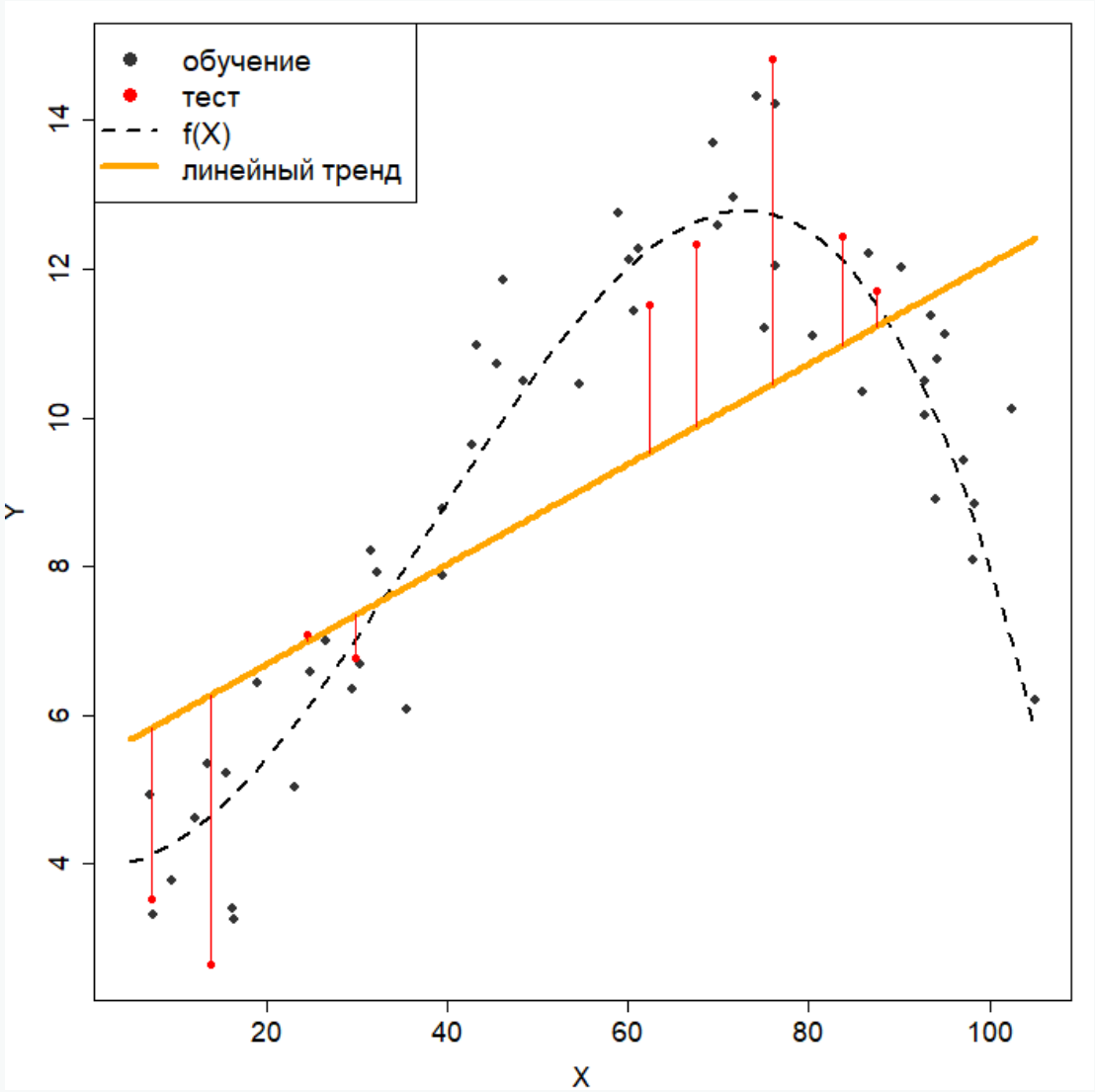
$MSE = 5.05$



1. Линейный тренд

$MSE = 5.05$

$MSE_{TEST} = 5.56$

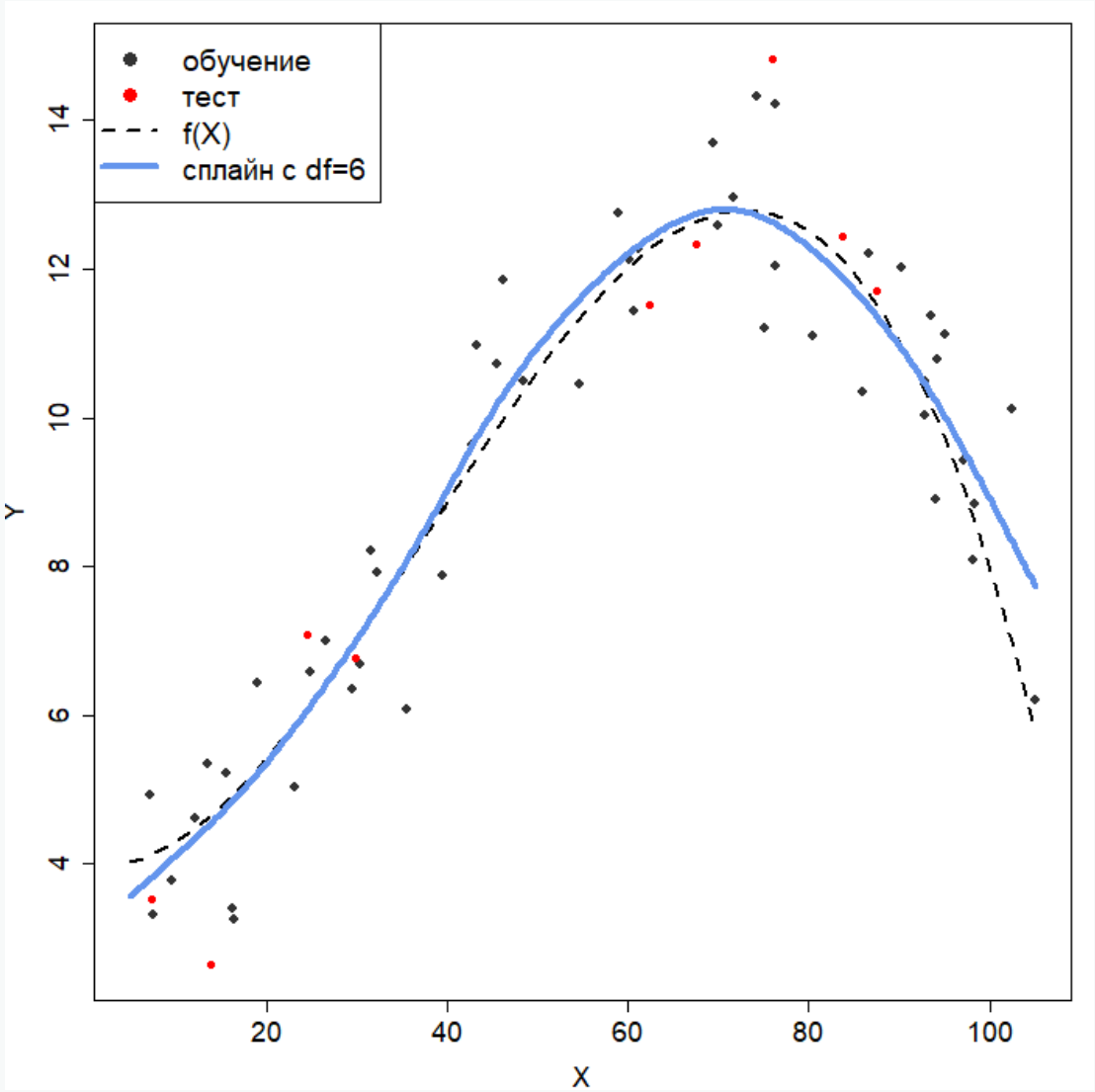


1. Линейный тренд

   $MSE = 5.05$

   $MSE_{TEST} = 5.56$

2. Сплайн с $df = 6$



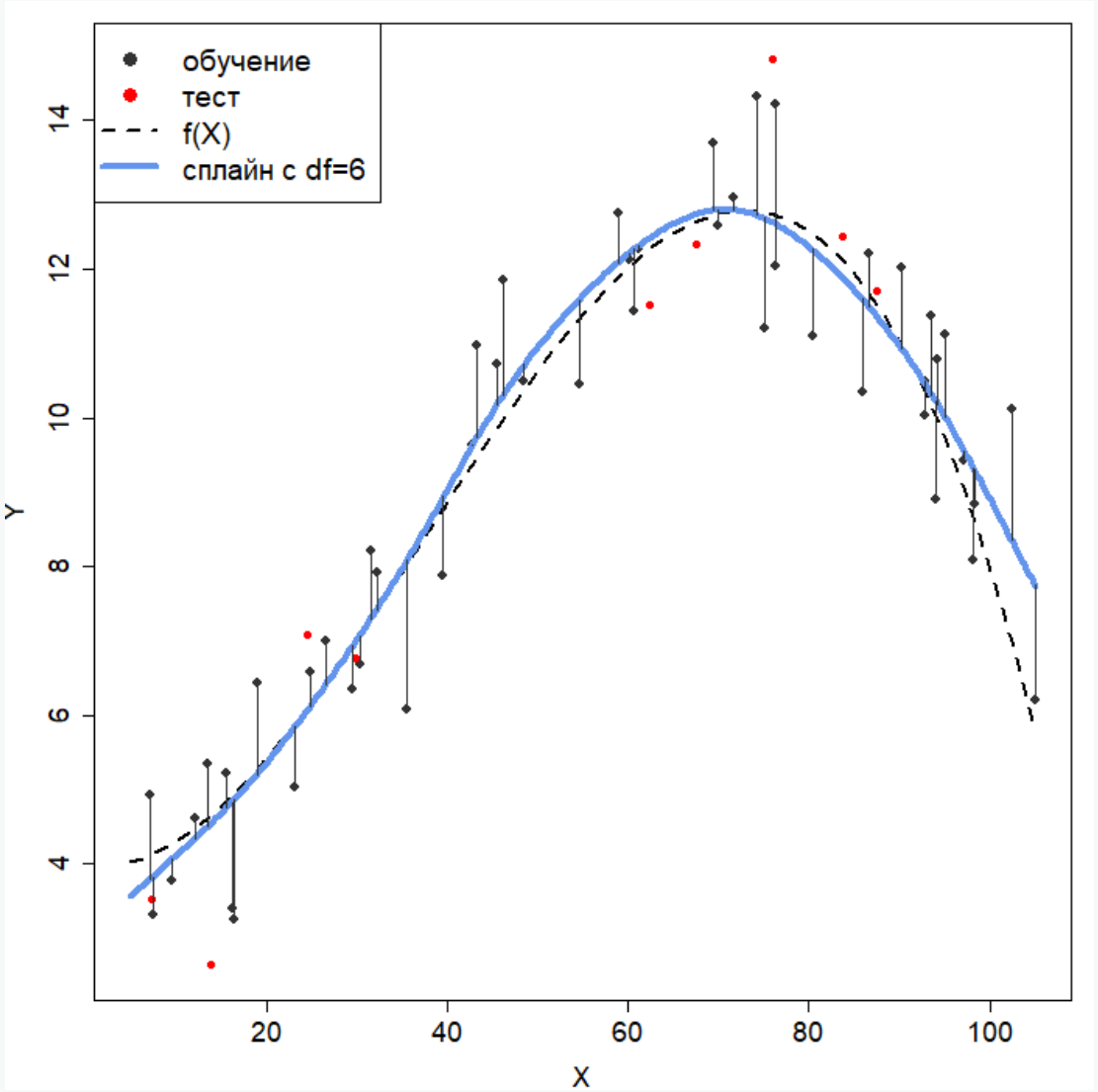
1. Линейный тренд

   $MSE = 5.05$

   $MSE_{TEST} = 5.56$

2. Сплайн с $df = 6$

   $MSE = 0.95$



1. Линейный тренд

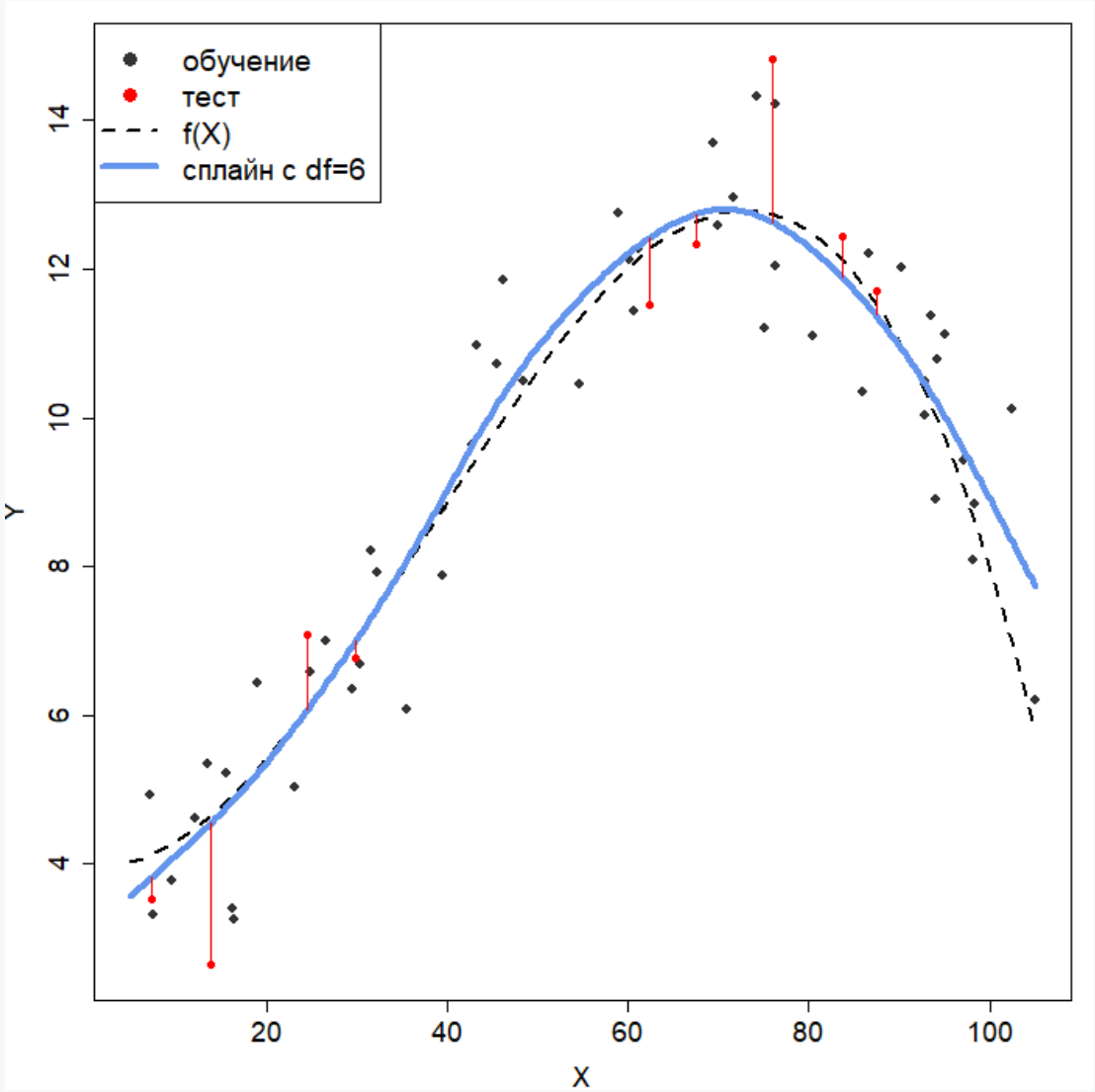
$MSE = 5.05$

$MSE_{TEST} = 5.56$

2. Сплайн с $df = 6$

$MSE = 0.95$

$MSE_{TEST} = 1.21$



1. Линейный тренд

   $MSE = 5.05$

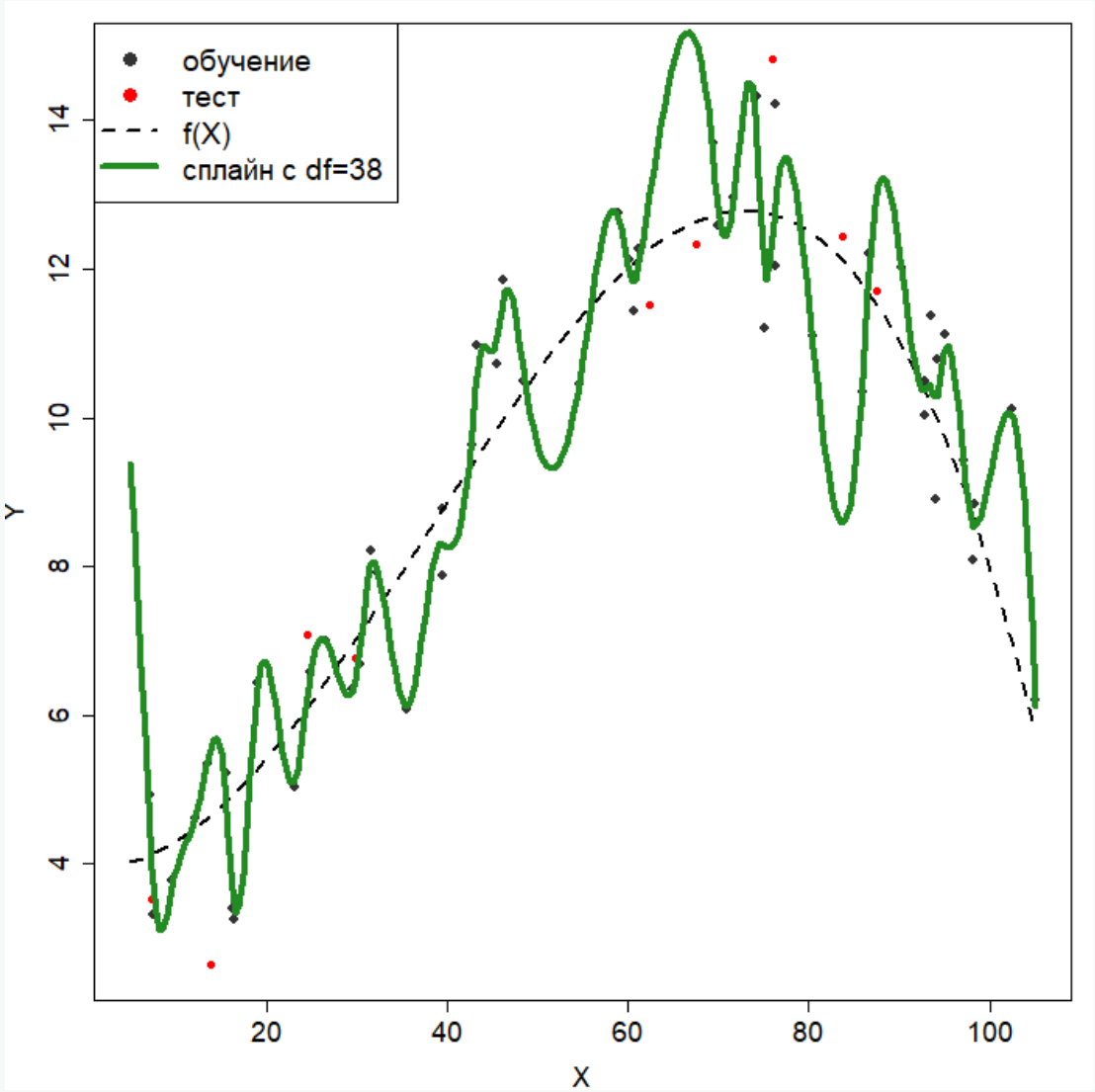
   $MSE_{TEST} = 5.56$

2. Сплайн с $df = 6$

   $MSE = 0.95$

   $MSE_{TEST} = 1.21$

3. Сплайн с $df = 38$



1. Линейный тренд

$MSE = 5.05$

$MSE_{TEST} = 5.56$

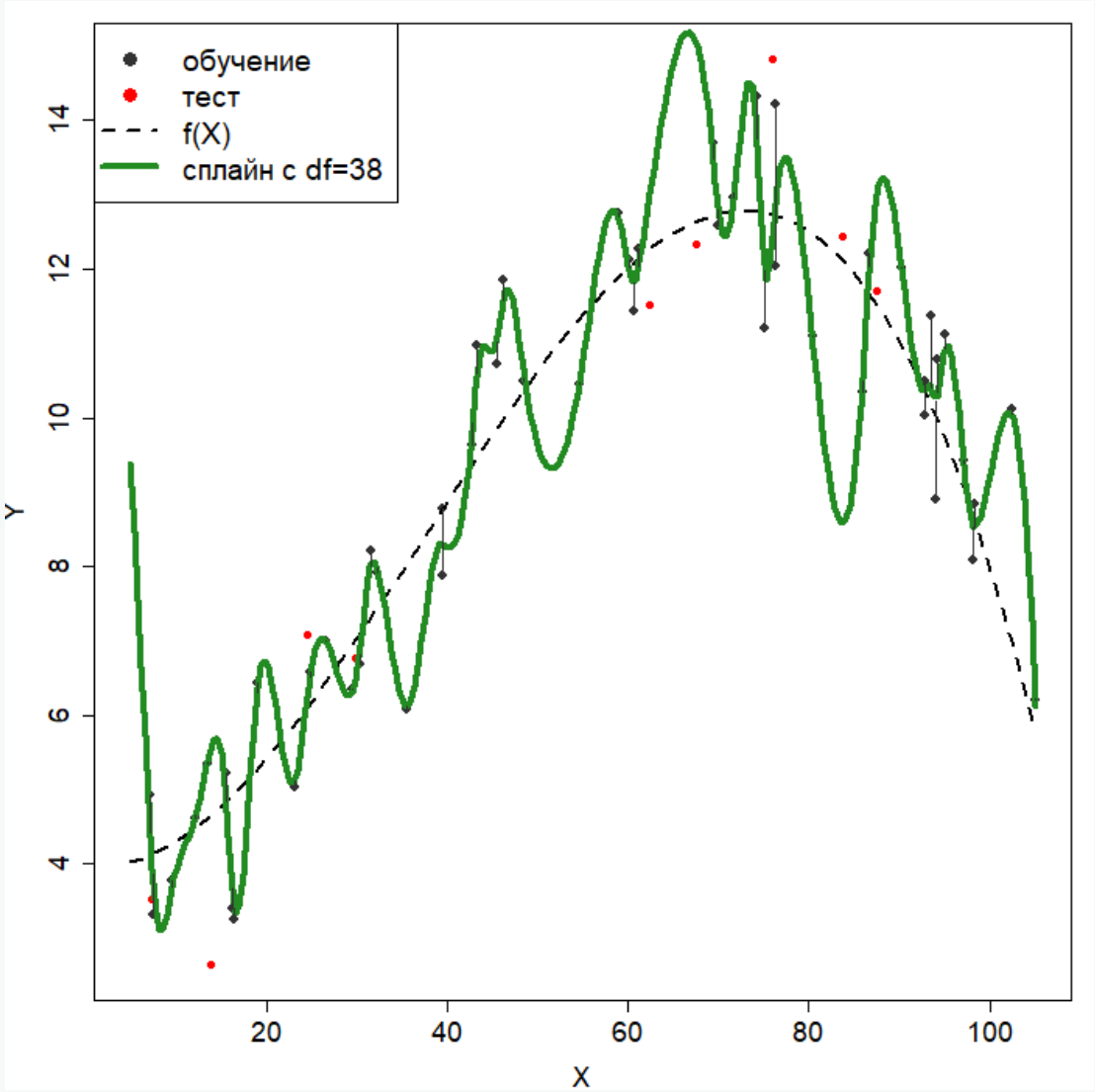
2. Сплайн с $df = 6$

$MSE = 0.95$

$MSE_{TEST} = 1.21$

3. Сплайн с $df = 38$

$MSE = 0.17$



1. Линейный тренд

   $MSE = 5.05$

   $MSE_{TEST} = 5.56$

2. Сплайн с $df = 6$

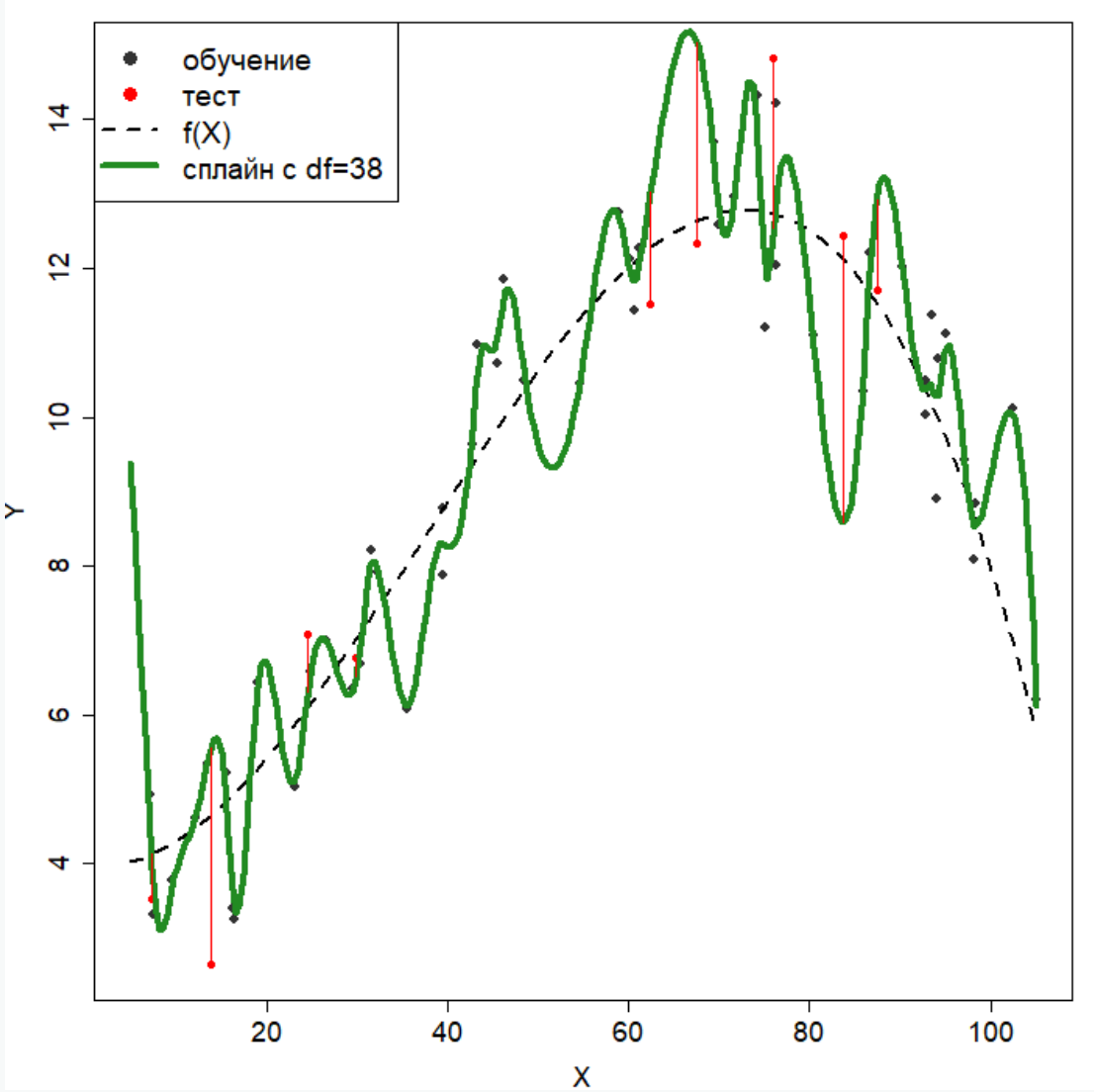
   $MSE = 0.95$

   $MSE_{TEST} = 1.21$

3. Сплайн с $df = 38$

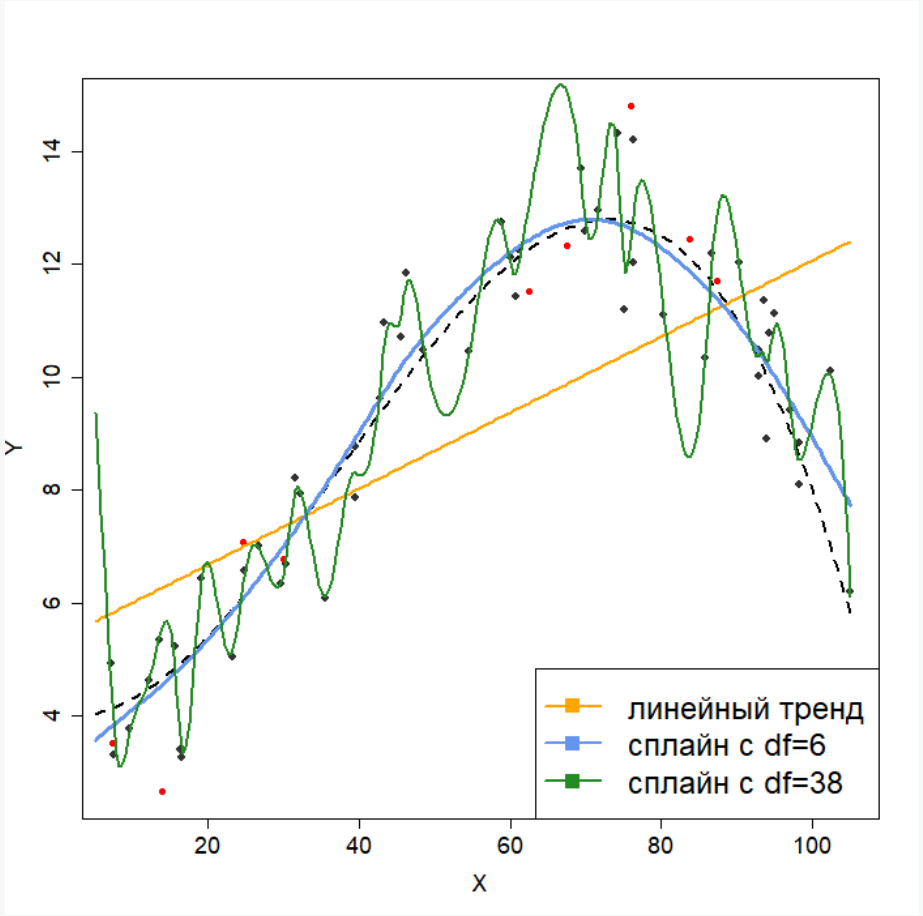
   $MSE = 0.17$

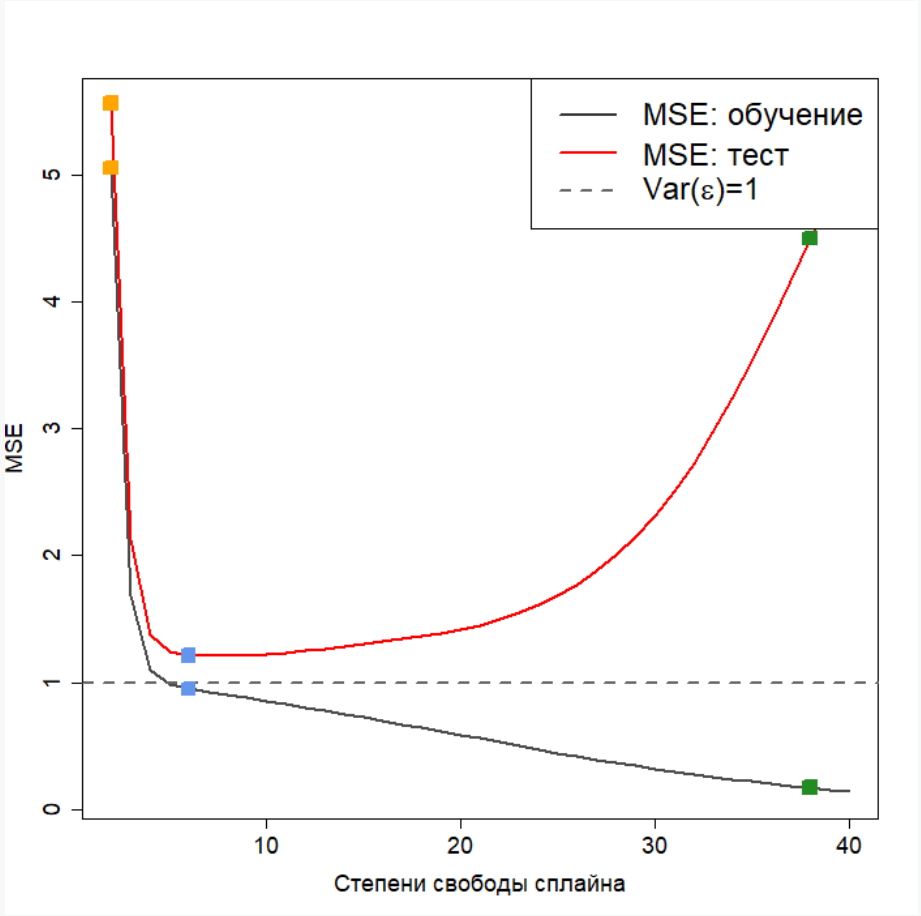
   $MSE_{TEST} = 4.5$



Сравнение точности моделей

Данные примера №1: $f(X) = 4 - 0.02X + 0.0055X^2 - 4.9 \cdot 10^{-5} \cdot X^3$





Смещение и дисперсия

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

- $E(y_0 - \hat{f}\,(x_0))^2$ – математическое ожидание ошибки на тестовой выборке;
- $Var(\hat{f}\,(x_0))$ – **дисперсия модели**;
- $[Bias(\hat{f}\,(x_0))]^2$ – **квадрат смещения модели**;
- $Var(\epsilon)$ – дисперсия случайной составляющей (неустранимая ошибка).

Смещение и дисперсия наглядно

Слева: данные, справа: модель с небольшим числом степеней свободы

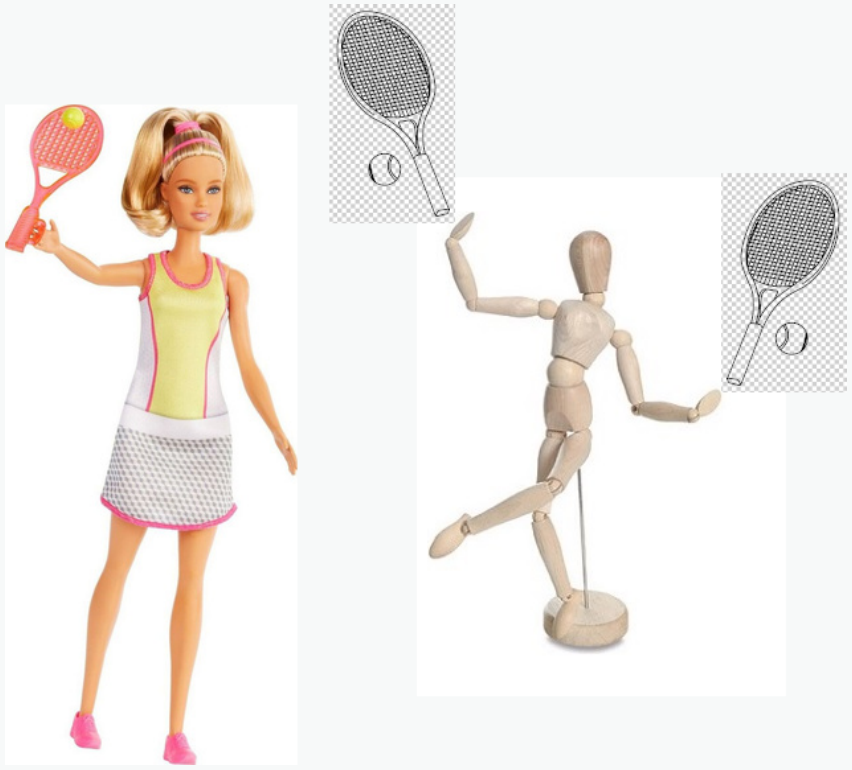
Высокое смещение модели, низкая дисперсия модели

Смещение и дисперсия наглядно

Слева: данные, справа: модель со средним числом степеней свободы (попали в форму функции)

Низкое смещение модели, средняя дисперсия модели

Смещение и дисперсия наглядно



Слева: данные, справа: модель с высоким (избыточным) числом степеней свободы

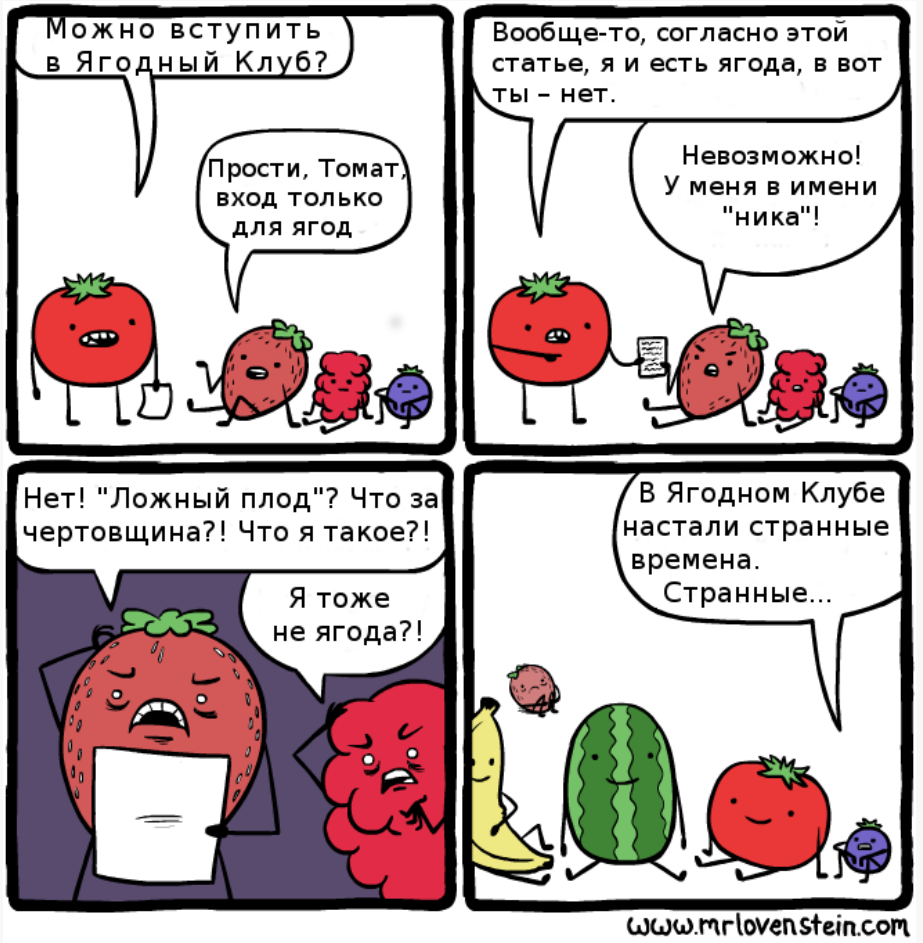
Низкое смещение модели (на обучающей!), высокая дисперсия

План лекции

- Что такое статистическое обучение
- Типы решаемых задач
- Непрерывный Y: понятия, связанные с точностью модели

- Дискретный Y: основные измерители точности

- Литература и ресурсы



www.mrlovenstein.com/comic/643

Задача классификации с учителем

Y – качественная переменная: $Y = \{1, 2, \ldots, C\}, C \geqslant 2$ – метки классов $\omega_1, \omega_2, \ldots, \omega_C$.

Ошибка модели – оценка *частоты ошибок классификации*:

$$E_{TRAIN} = \frac{1}{n_{TRAIN}} \cdot \sum_{i \in TRAIN} I(y_i \neq \hat{y}_i) \quad ; \quad E_{TEST} = \frac{1}{n_{TEST}} \cdot \sum_{i \in TEST} I(y_i \neq \hat{y}_i)$$

$$I(y_i \neq \hat{y}_i) = \begin{cases} 0, & y_i = \hat{y}_i, \\ 1, & y_i \neq \hat{y}_i. \end{cases}$$

Матрица неточностей (*confusion matrix*), или таблица сопряжённости:

$$M = \{m_{ij}\}_{i,j=1}^{C}$$

– показывает количество объектов, принадлежащих классу ω_i (i – номер строки) и отнесённых моделью к классу ω_j (j – номер столбца).

Задача классификации с учителем

Если Y – **бинарная** переменная: $Y = \{0, 1\}$, размерность матрицы неточностей 2x2:

		Предсказанное значение `$\hat{Y}$`	
		– `$(H_0)$`	+ `$(H_1)$`
Истинное значение `$Y$`	– `$(H_0)$`	**TN**	**FP**
	+ `$(H_1)$`	**FN**	**TP**

"+" означает наличие признака; "–" – отсутствие признака.
TN – истинно отрицательные случаи, **FP** – ложноположительные, **FN** – ложноотрицательные, **TP** – истинно положительные.

		Предсказанное значение $\hat{Y}$	
		– (H_0)	+ (H_1)
Истинное значение Y	– (H_0)	**TN**	**FP**
	+ (H_1)	**FN**	**TP**

Чувствительность (*sensitivity*, *TP rate*, *recall*):

$$TPR = P(\text{модель:} + \,|\, \text{факт:} +) = \frac{TP}{(TP + FN)}$$

Специфичность (*specificity*, *TN rate*):

$$SPC = P(\text{модель:} - \,|\, \text{факт:} -) = \frac{TN}{(TN + FP)}$$

		Предсказанное значение `$\hat{Y}$`	
		– `$(H_0)$`	+ `$(H_1)$`
Истинное значение `$Y$`	– `$(H_0)$`	**TN**	**FP**
	+ `$(H_1)$`	**FN**	**TP**

Ценность положит-ого прогноза (*positive predictive value*, *precision*):

$$PPV = P(\text{факт: } + \,|\, \text{модель:} +) = \frac{TP}{(TP + FP)}$$

Ценность отрицательного прогноза (*negative predictive value*):

$$NPV = P(\text{факт: } - \,|\, \text{модель:} -) = \frac{TN}{(TN + FN)}$$

Ошибки

Доля ложноотрицательных исходов (*FN rate*):

$$FNR = 1 - TPR$$

Доля ложных срабатываний (*fall-out rate*, *FP rate*):

$$FPR = 1 - SPC$$

Доля ложного обнаружения (*false discovery rate*):

$$FDR = 1 - PPV$$

чем ниже, тем лучше

Сводные характеристики качества

Точность, или верность (*Accuracy*):

$$Acc = P(y_i = j \,|\, y_i \in \omega_j) = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

F-мера – гармоническое среднее ценности положительного прогноза и чувствительности:

$$F1 = \frac{2}{(\frac{1}{PPV} + \frac{1}{TPR})}$$

Корреляция Мэтьюса (*ϕ-коэффициент*):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}; \quad MCC \in [-1,\, 1]$$

- $MCC = -1$: предсказанные классы противоположны истинным;
- $MCC = 0$: предсказанные классы не связаны с истинными (случайны);
- $MCC = 1$: предсказанные классы идеально совпадают с истинными.

План лекции

- Что такое статистическое обучение
- Типы решаемых задач
- Непрерывный Y: понятия, связанные с точностью модели
- Дискретный Y: основные измерители точности

- Литература и ресурсы

Литература

- *Рашка С.* Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения / пер. с англ. А.В. Логунова. – М.: ДМК Пресс, **2017**. – 418 с.: ил. Репозиторий с примерами к книге (англ.): https://github.com/rasbt/python-machine-learning-book-2nd-edition

- *Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р.* Введение в статистическое обучение с примерами на языке R / пер. с англ. С.Э. Мастицкого. – М.: ДМК Пресс, **2016** – 450 с. Репозиторий с примерами к книге на русском языке: https://github.com/ranalytics/islr-ru

- *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А.Слинкина. – М.: ДМК Пресс, **2015**. – 400 с.

- *Бринк Х., Ричардс Дж., Феверолф М.* Машинное обучение. – Спб.: Питер, **2018**. – 336 с.

- *Анналин Ын, Кеннет Су* Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. – Спб.: Питер, **2019**. – 208 с.

Онлайн ресурсы

- Open Data Science на habr.com
- Списки ресурсов по ML на английском: github.com/josephmisiti/awesome-machine-learning
- Скрипты к практикам: https://github.com/aksyuk/MTML/tree/main/Labs
- Слайды к лекциям: https://github.com/aksyuk/MTML/tree/main/Lections

Что установить

- Anaconda Python не ниже версии 3.8

Где завести аккаунты

- github.com
- kaggle.com