

Основные идеи машинного обучения

Ридли Михаил Кристофорович
Ридли Александра Николаевна

Положение среди других наук

МО изучает методы построения алгоритмов, способных к обобщению и обучению

Распознавание образов ≈ машинное обучение

Анализ данных ⋂ машинное обучение ≈ Ø

Машинное обучение с искусственный интеллект

Индуктивное обучение — МО
дедуктивное — экспертные системы

Знание ЛА, ТВиМС, ЧМ, МОП и ДА

Вопросы МО

- Какой алгоритм решает задачу хорошо? Как его создать?
- Сколько и какой информации нужно для обучения?
- Какие данные выбрать, как их выбор повлияет на качество обучения?
- Как свести задачу обучения к аппроксимации или оптимизации какой-либо функции?
- Практические вопросы выбора моделей данных и моделей алгоритмов.
- Вопросы практических приложений.

Типы МО

- Обучение с учителем (supervised)
явная обратная связь от «учителя»
- Частичное обучение (semi-supervised)
частичная обратная связь от «учителя»
- Активное обучение (active)
обратная связь по запросу
- Стимулируемое обучение (reinforced)
отложенная обратная связь
- Обучение без учителя (unsupervised learning)
обратная связь отсутствует

Обучение с учителем и + частичное

- Для некоторых объектов в исходных данных известны правильные ответы (их знает «учитель»)
- Учитель \approx обучающая выборка пар (объект, ответ)
- Задача обучения — нахождение закономерности, позволяющей узнать ответ для любого объекта

Активное обучение

- Ответы изначально неизвестны
- Гипотеза: алгоритм может обучаться на малых выборках, если будет самостоятельно выбирать данные
- Алгоритм составляет запросы (query), которые позволяют ему обучиться

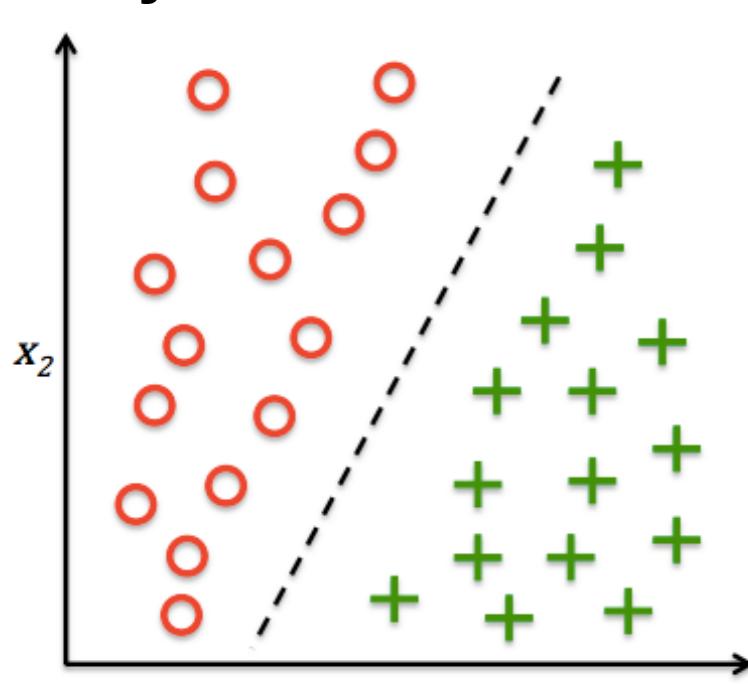
Обучение с подкреплением

- Правильных ответов нет
- Алгоритм («агент») взаимодействует со средой и получает обратную связь
- Учитель = реальная среда
- Online-алгоритм: поиск баланса между исследованием среды и её эксплуатацией

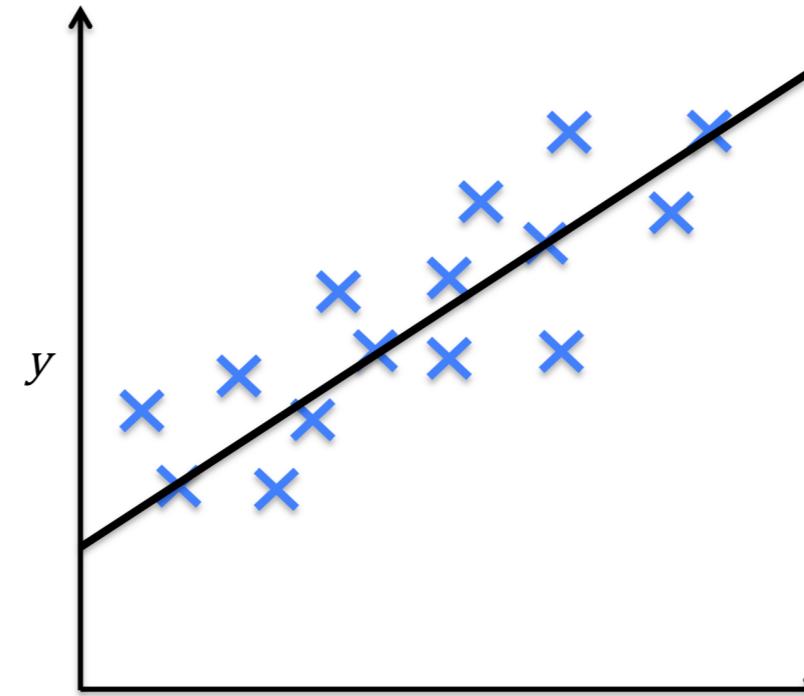
Обучение без учителя

- Работа с данными без сторонней информации: ищутся не зависимости, а связи между объектами
- Критерии качества очень разные, сложно выбрать
- Многие методы Data mining относятся к этому типу

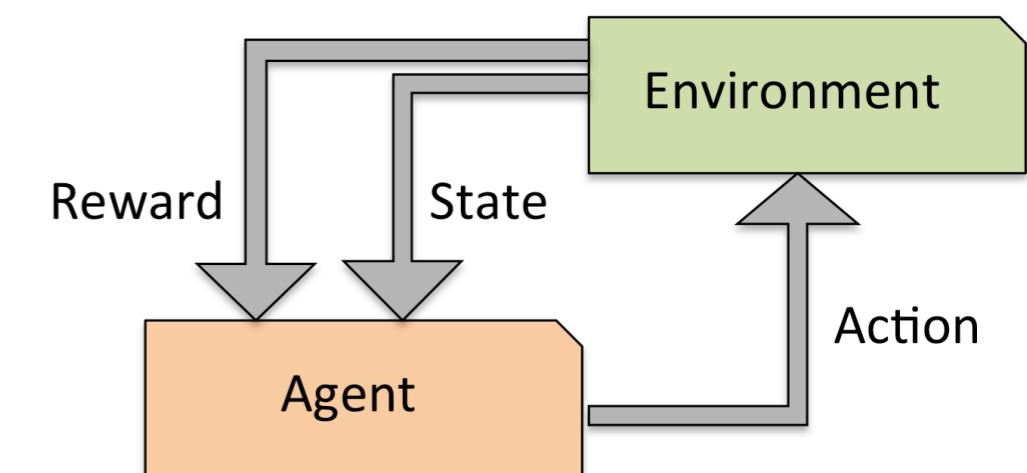
С учителем:



Классификация

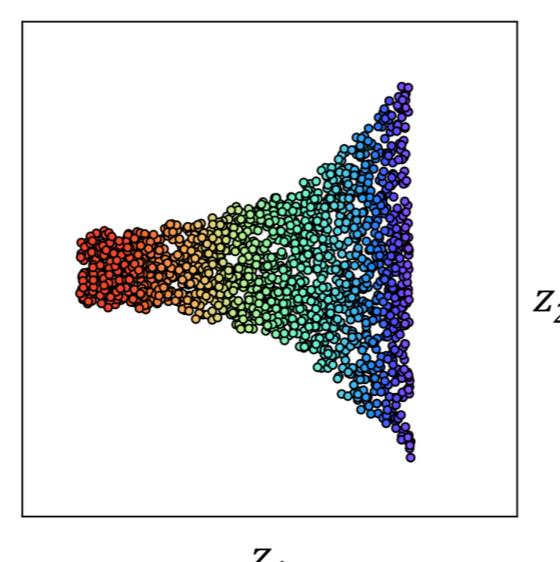
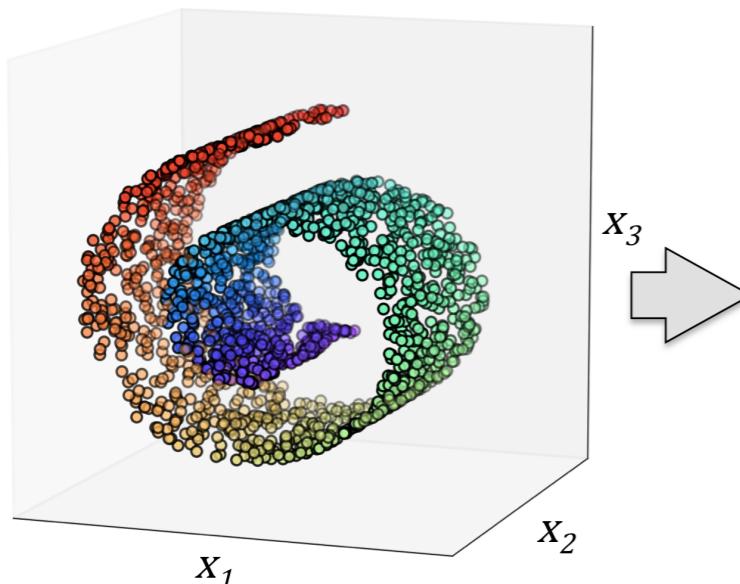


Регрессия

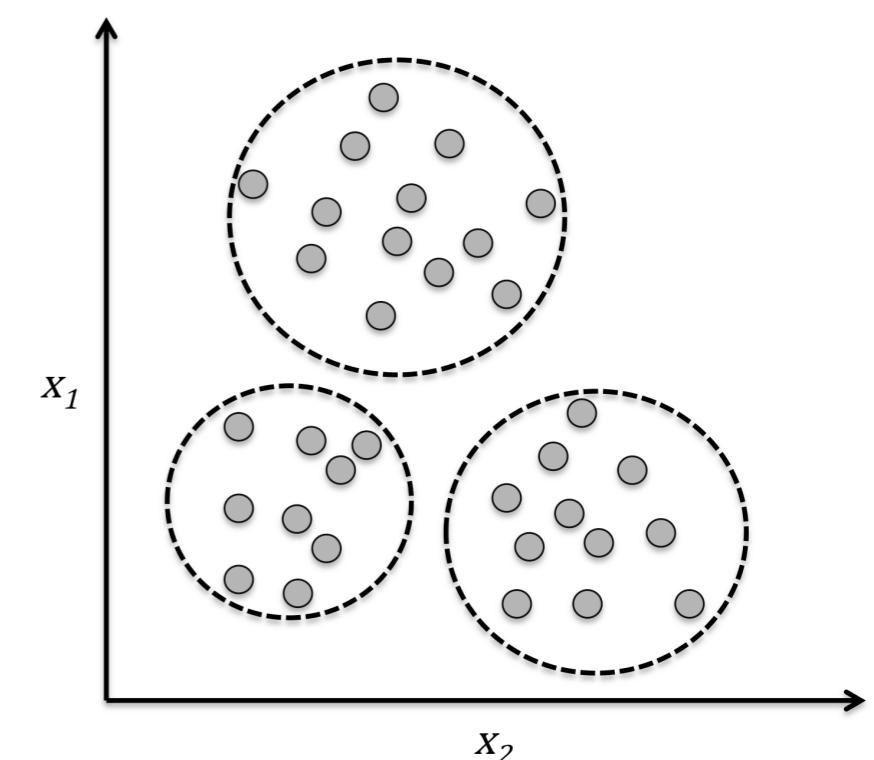


Обучение с подкреплением

Без учителя:



Уменьшение размерности



Без учителя

Задача обучения по прецедентам

X — множество объектов

Y — множество ответов

$y: X \rightarrow Y$ — неизвестная зависимость

Исходные данные:

$\{x_1, \dots, x_l\} \subset X$ — обучающая выборка

$y_i = y(x_i), i = 1, \dots, l$ — известные ответы

Требуется найти:

$a: X \rightarrow Y$ — решающую функцию (алгоритм), приближающую y на всём множестве X

3 главных вопроса

- Как задаются объекты и какими могут быть ответы?
- К каком смысле «у приближает a »?
- Как строить функцию/алгоритм a ?

Как задаются объекты

- Признаковое описание (чаще всего)
- Временные ряды или сигналы
- Изображение или видеоряд
- Описание отношений (матрица расстояний или граф)

Признаковое описание объектов

$f_j : X \rightarrow D_j, j = 1, \dots, n$ — признаки объектов (features)

Типы признаков (результатов измерения над объектом):

- $D_j = \{0, 1\}$ — бинарный признак
- $|D_j| < \infty$ — номинальный признак
- $|D_j| < \infty, D_j$ упорядочено — порядковый признак
- $D_j = \mathbb{R}$ — количественный признак

Вектор $(f_1(x), \dots, f_n(x))$ — признаковое описание объекта x

Матрица «объекты-признаки» (feature data):

$$F = \|f_j(x_i)\|_{l \times n} = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \cdots & \cdots & \cdots \\ f_1(x_l) & \cdots & f_n(x_l) \end{pmatrix}$$

Как задаются ответы

Задачи классификации (classification):

- $Y = \{-1, +1\}$ — классификация на 2 класса
- $Y = \{1, \dots, M\}$ — на M непересекающихся классов
- $Y = \{0, 1\}^M$ — на M классов, которые могут пересекаться

Задачи восстановления регрессии (regression):

- $Y = \mathbb{R}$ или $Y = \mathbb{R}^m$

Задачи ранжирования (ranking):

- Y — конечное упорядоченное множество

Классический пример: цветки ириса

4 признака, 3 класса, длина выборки — 120

Samples

(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

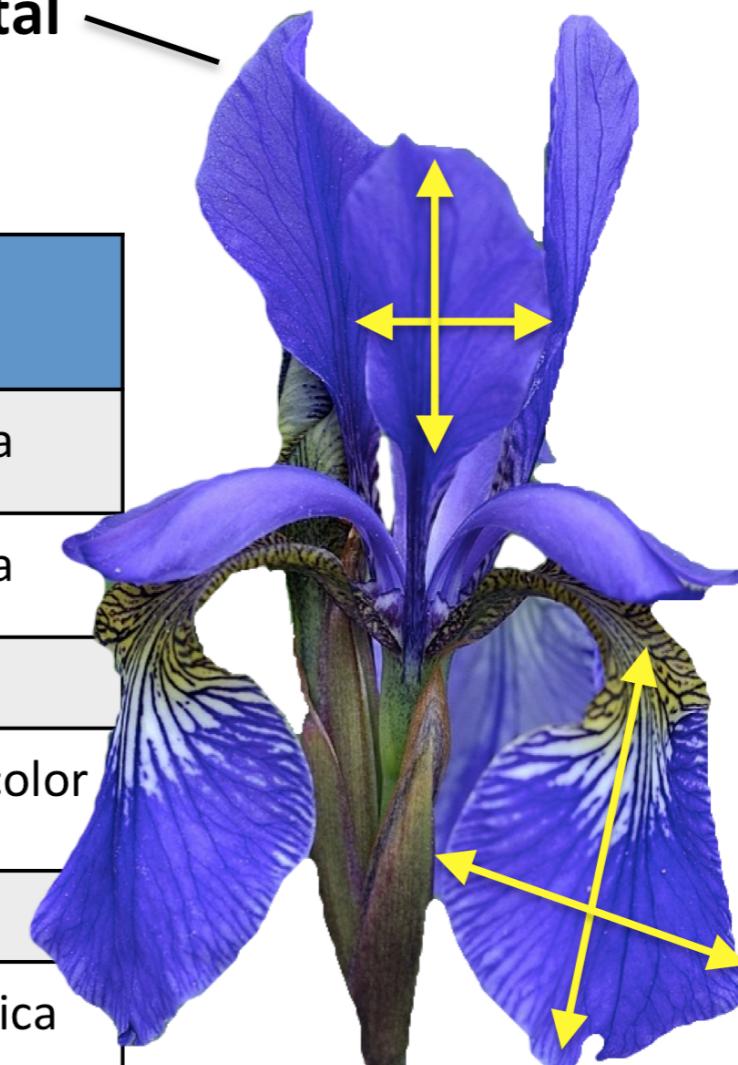
Features

(attributes, measurements, dimensions)

Petal

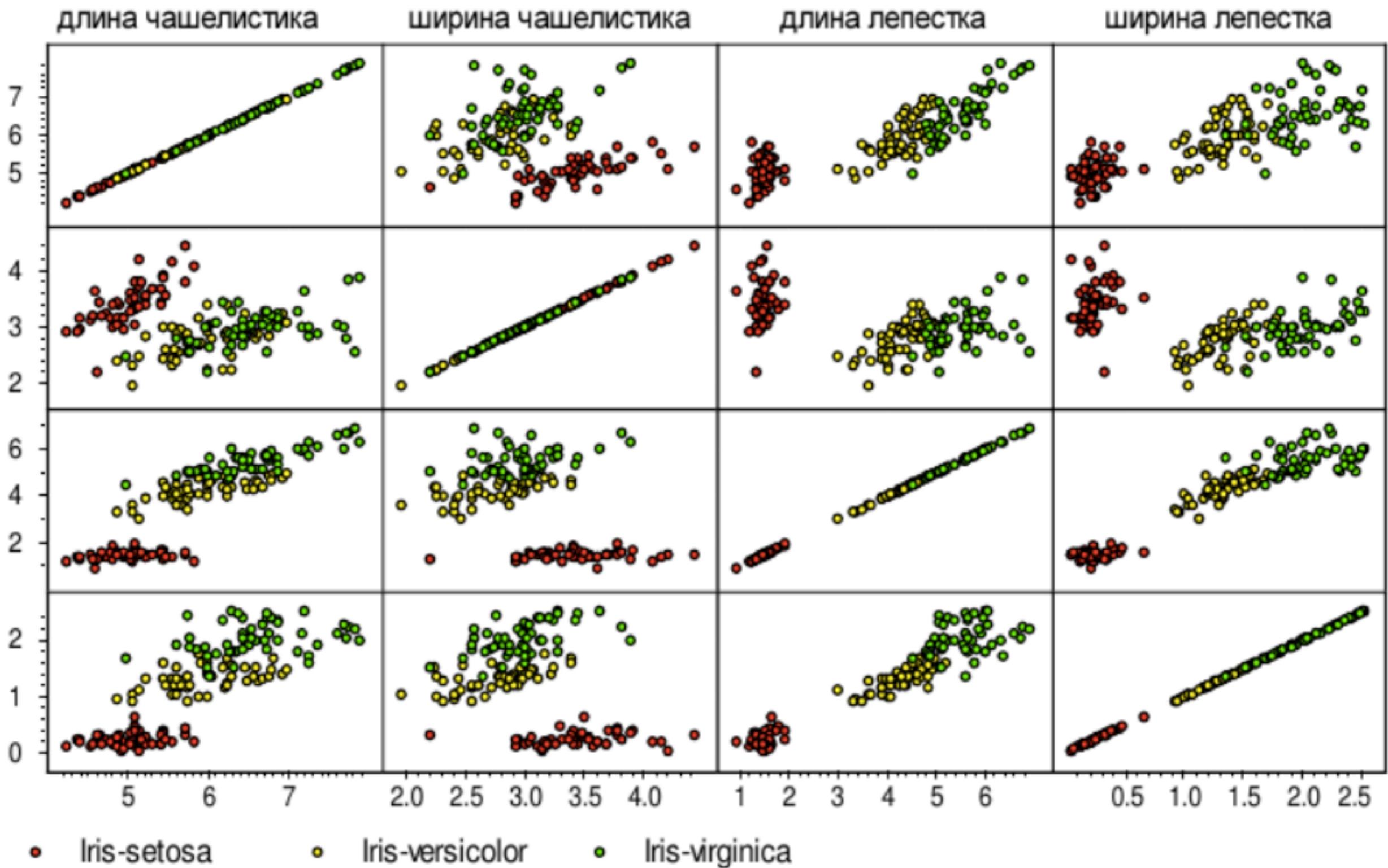
Class labels
(targets)

Sepal



Классический пример: цветки ириса

4 признака, 3 класса, длина выборки — 120



Модель алгоритмов (predictive model)

Модель — параметрическое семейство функций

$$A = \{g(x, \theta) | \theta \in \Theta\},$$

где $g : X \times \Theta \rightarrow Y$ — фиксированная функция,
 Θ — множество допустимых значений параметра θ

Пример

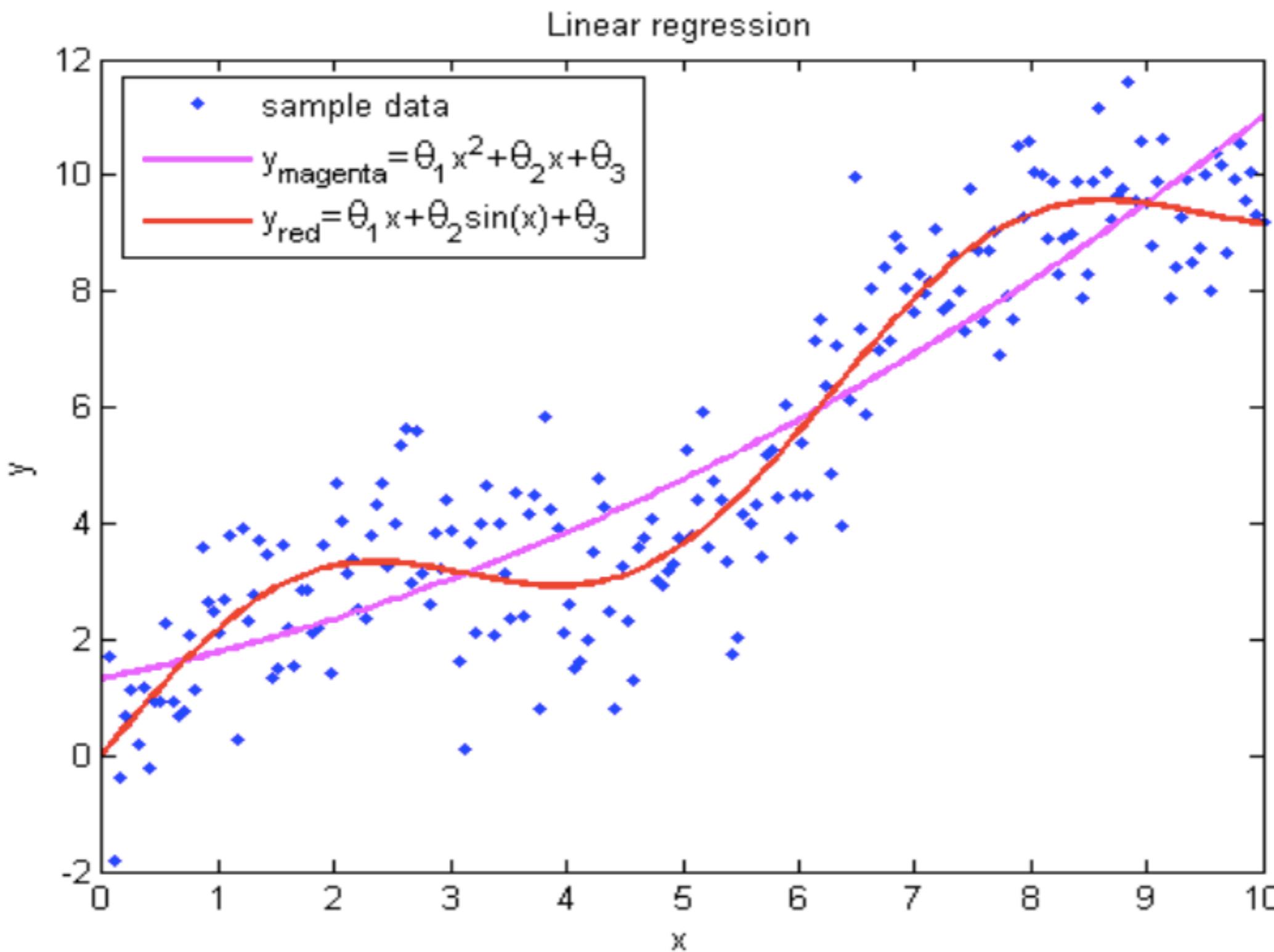
Линейная модель с вектором параметров $\theta = (\theta_1, \dots, \theta_n)$, $\Theta = \mathbb{R}^n$

$g(x, \theta) = \sum_{j=1}^n \theta_j f_j(x)$ — для регрессии и ранжирования, $Y = \mathbb{R}$

$g(x, \theta) = \text{sign} \sum_{j=1}^n \theta_j f_j(x)$ — для классификации, $Y = \{-1, +1\}$

Пример выбора признаков (задача регрессии)

$X = Y = \mathbb{R}$, $I = 200$, $n = 3$ признака:
 $\{x, x^2, 1\}$ или $\{x, \sin x, 1\}$



Метод обучения (learning algorithm)

Метод обучения — отображение вида $\mu : (X \times Y)^I \rightarrow A$

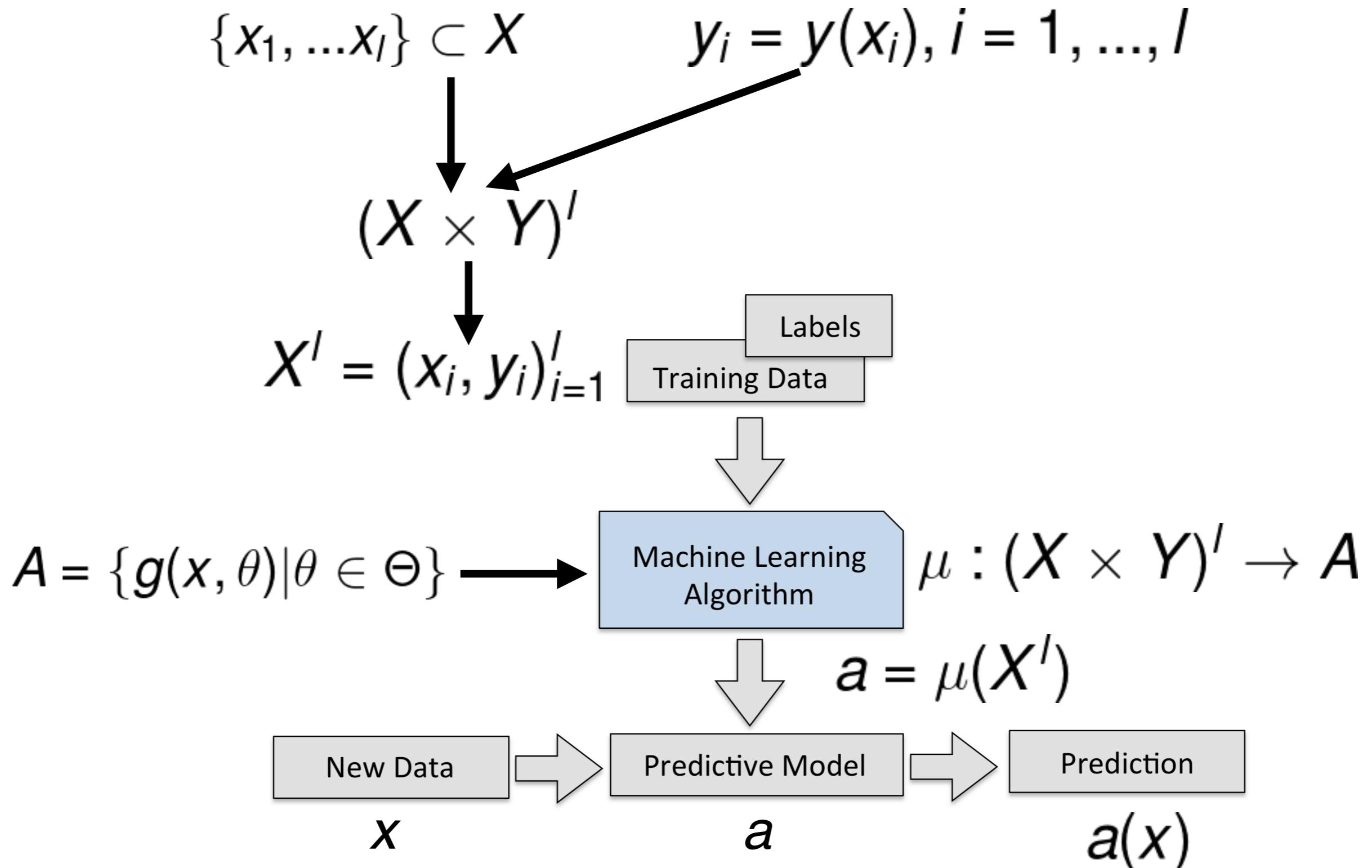
которое произвольной выборке $X^I = (x_i, y_i)_{i=1}^I$ ставит в соответствие некоторый алгоритм $a \in A$.

Почти всегда есть 2 этапа:

- обучение (training)
метод μ по выборке X^I строит алгоритм $a = \mu(X^I)$
- применение (testing)
алгоритм a для новых объектов x выдаёт ответы $a(x)$

Смысл: алгоритм μ находит гипотезу в модели, которая наилучшим образом приближает целевую функцию, используя известные значения. По сути — подбор коэффициентов.

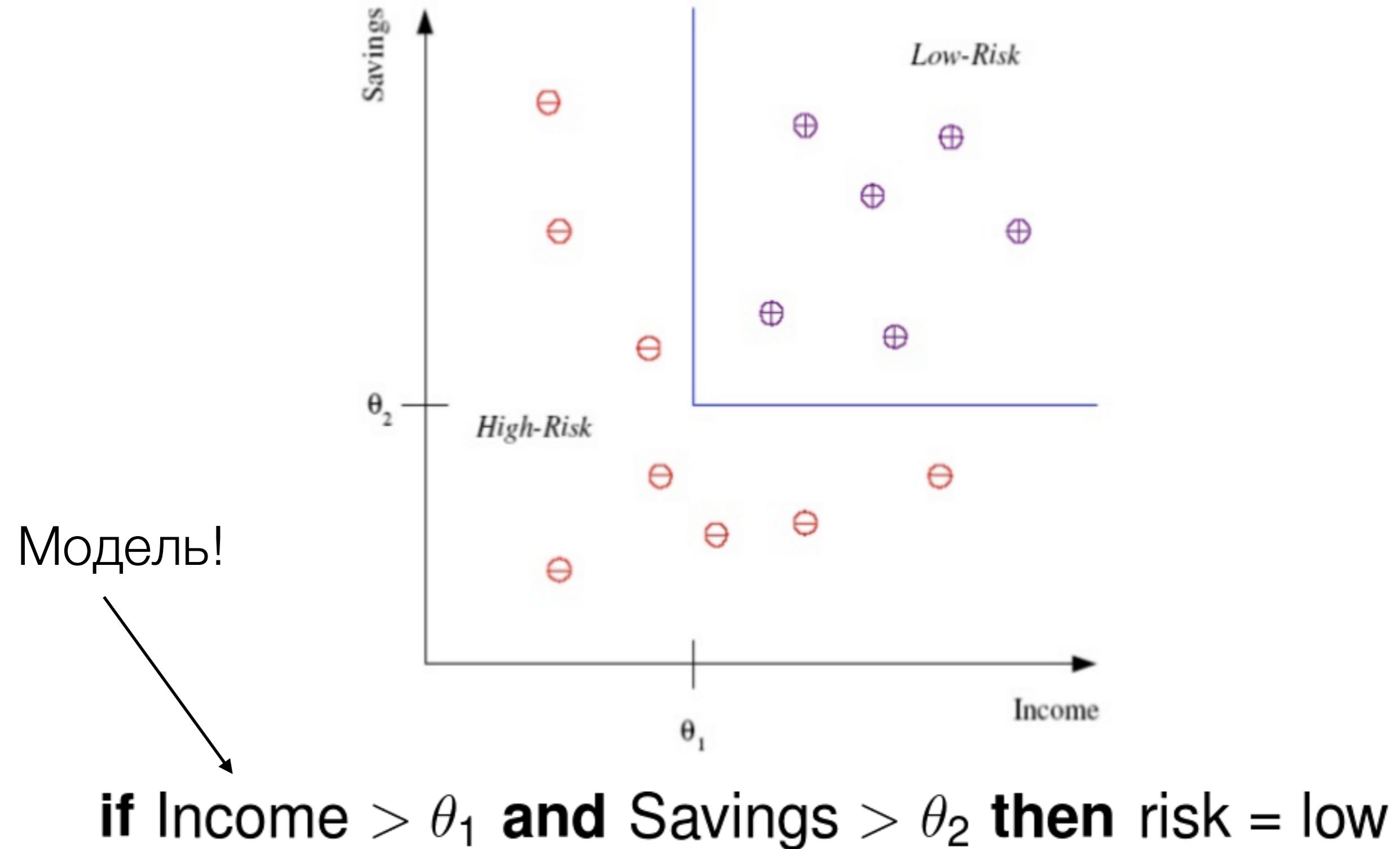
Метод обучения (learning algorithm)



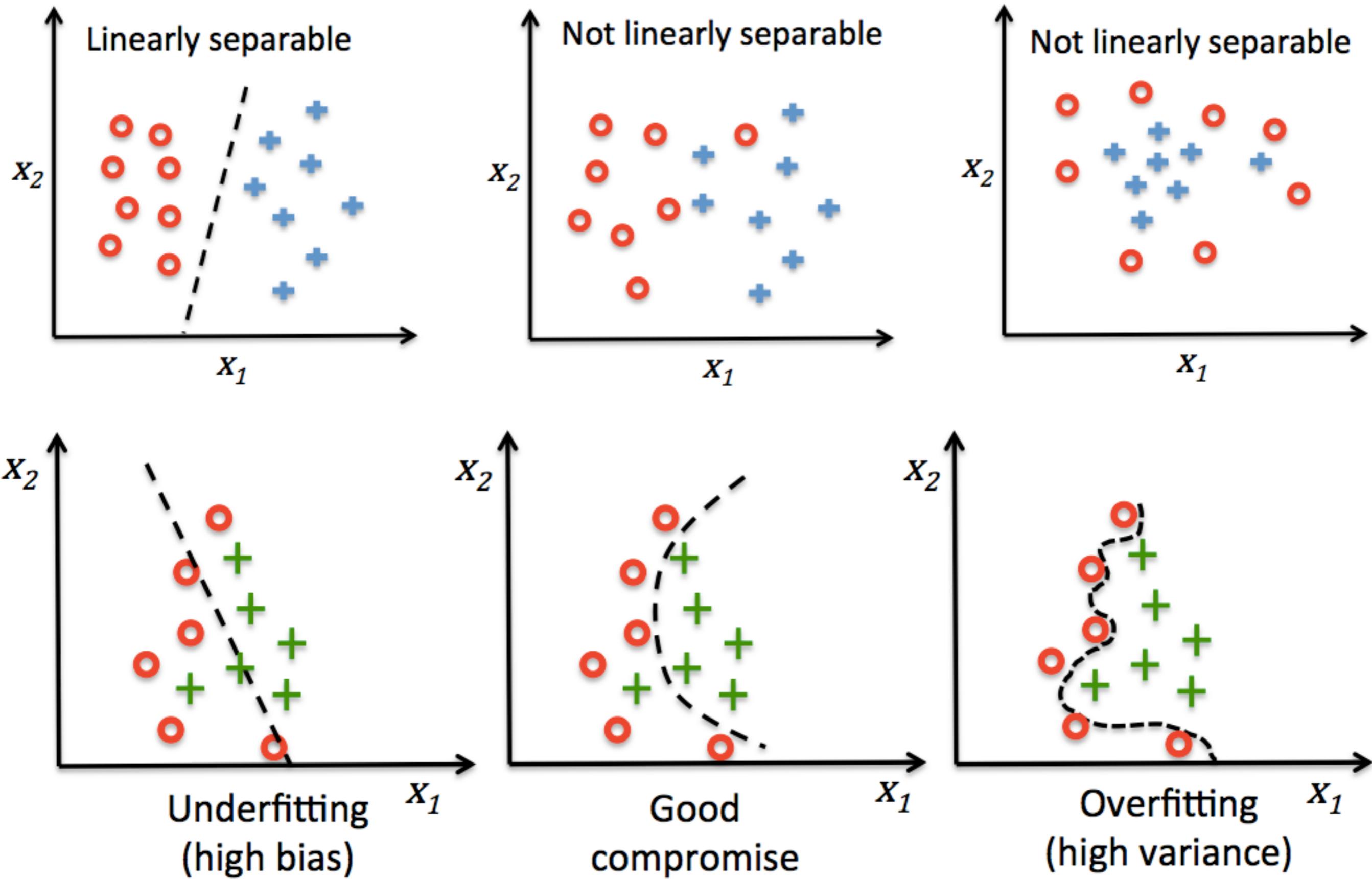
Пример: классификация двумя прямыми

Принять решение о выдаче кредита (low-risk, high-risk)

Известны только сбережения (savings) и доходы (income)



Пример: классификация двумя прямыми



Этапы обучения и применения

Этап обучения:

Метод μ по выборке $X^l = (x_i, y_i)_{i=1}^l$ строит алгоритм $a = \mu(X^l)$

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix} \xrightarrow{\mu} a$$

Этап применения:

Алгоритм a для новых объектов x'_i выдаёт ответы $a(x'_i)$

$$\begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x'_1) \\ \dots \\ a(x'_k) \end{pmatrix}$$

Функционалы качества

$\mathcal{L}(a, x)$ — функция потерь (loss function), т.е. величина ошибки алгоритма $a \in A$ на объекте $x \in X$.

Функция потерь для задач классификации:

$$\mathcal{L}(a, x) = [a(x) \neq y(x)]^* \text{ — индикатор ошибки}$$

Функции потерь для задач регрессии:

$$\mathcal{L}(a, x) = |a(x) - y(x)| \text{ — абсолютное значение ошибки}$$

$$\mathcal{L}(a, x) = (a(x) - y(x))^2 \text{ — квадратичная ошибка}$$

Эмпирический риск — функционал качества алгоритма a на X' :

$$Q(a, X') = \frac{1}{I} \sum_{i=1}^I \mathcal{L}(a, x_i).$$

* — $[x] = 1$, если x — Истина, в противном случае — 0

Обучение сводится к задаче оптимизации

Метод минимизации эмпирического риска:

$$\mu(X') = \arg \min_{a \in A} Q(a, X')$$

Пример: метод наименьших квадратов ($Y = \mathbb{R}$, \mathcal{L} квадратична):

$$\mu(X') = \arg \min_{\theta} \sum_{i=1}^I (g(x_i, \theta) - y_i)^2$$

Проблема обобщающей способности:

- надём ли мы «истину» или просто переобучимся, то есть подгоним функцию $g(x_i, \theta)$ под заданные точки?
- будет ли $a = \mu(X')$ приближать функцию u на всём X ?
- будет ли $Q(a, X^k)$ мало на новых данных — контрольной выборке $X^k = (x'_i, y'_i)_{i=1}^k$, $y'_i = y(x_i)$?

Пример переобучения

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

Признаковое описание: $x \mapsto (1, x^1, x^2, \dots, x^n)$

Модель полиномиальной регрессии:

$a(x, \theta) = \theta_0 + \theta_1 x + \dots + \theta_n x^n$ — полином степени n

Обучение методом наименьших квадратов:

$$Q(\theta, X^l) = \sum_{i=1}^l (\theta_0 + \theta_1 x_i + \dots + \theta_n x_i^n - y_i)^2 \rightarrow \min_{\theta_1, \dots, \theta_n}$$

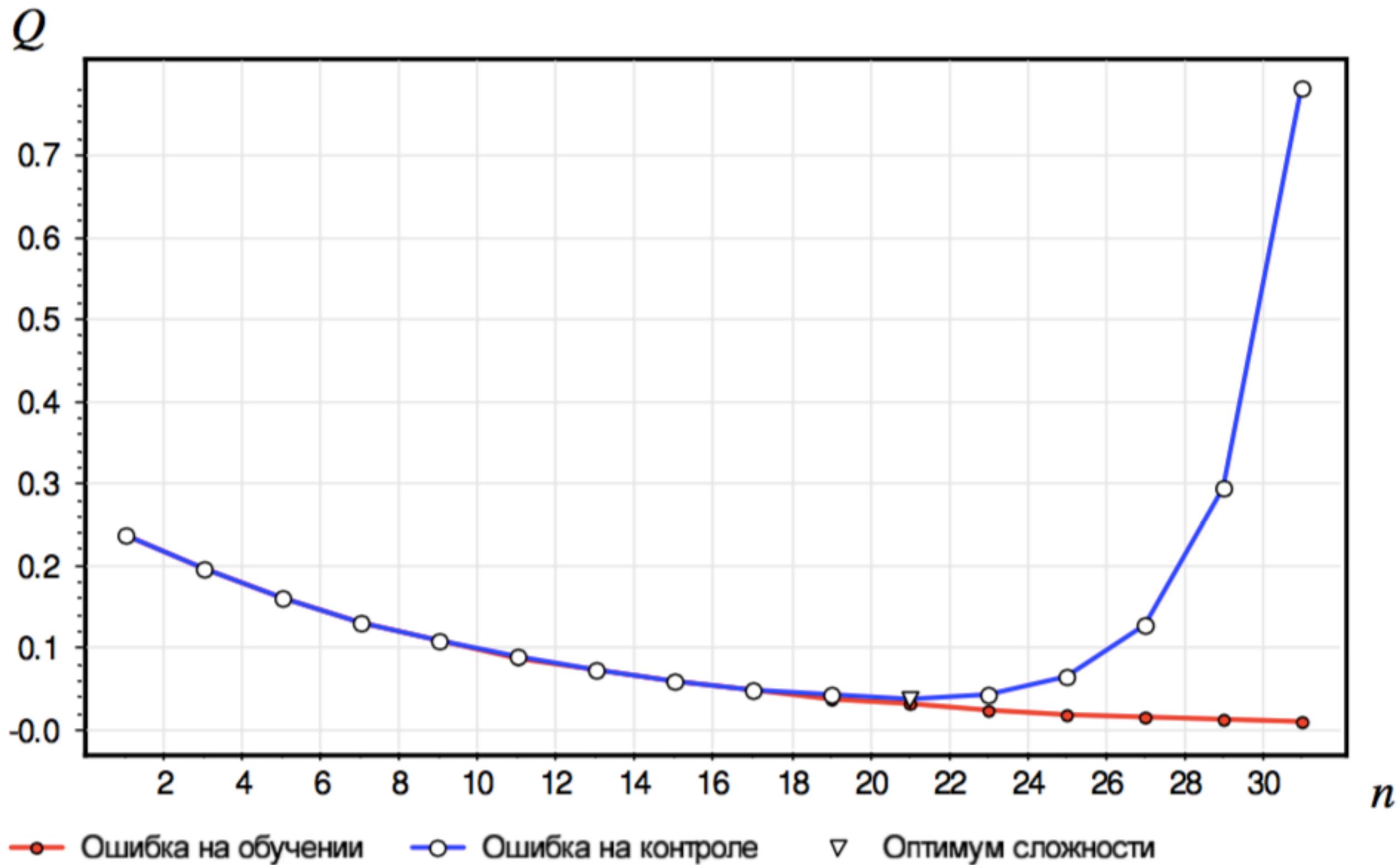
Обучающая выборка: $X^l = \left\{ x_i = 4 \frac{i-1}{l-1} - 2 \mid i = 1, \dots, l \right\}$

Контрольная выборка: $X^k = \left\{ x_i = 4 \frac{i-0.5}{l-1} - 2 \mid i = 1, \dots, l-1 \right\}$

Что происходит с $Q(\theta, X^l)$ и $Q(\theta, X^k)$ при увеличении n ?

Пример переобучения (при $l = 50$, $n = 1..31$)

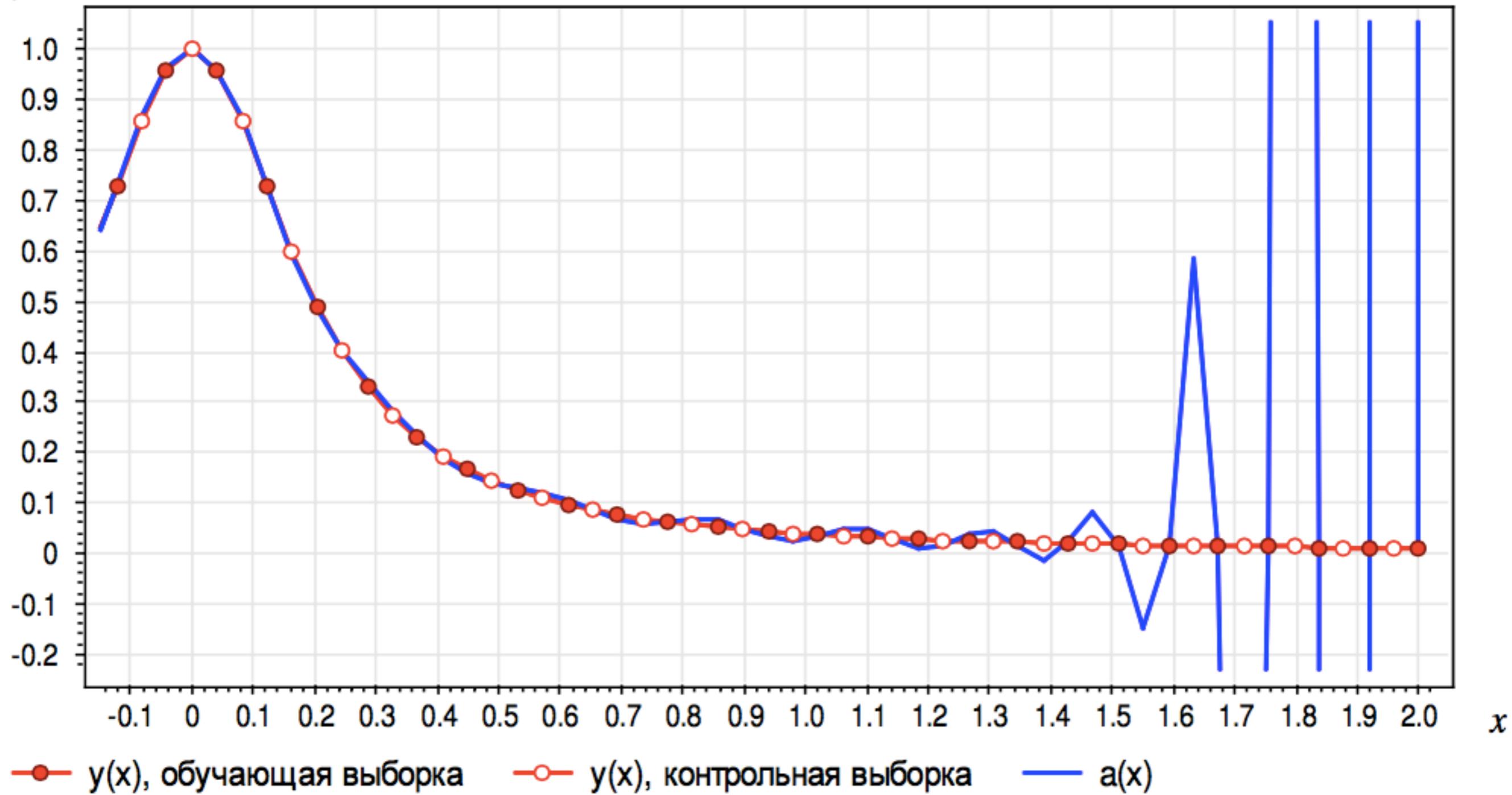
Переобучение — это когда $Q(\mu(X^l), X^k) \gg Q(\mu(X^l), X^l)$



Пример переобучения (при l = 50, n = 38)

Зависимость $y(x) = \frac{1}{1 + 25x^2}$

$y(x), a(x)$



Переобучение — одна из основных проблем МО

- Из-за чего оно возникает?
 - избыточная сложность пространства параметров, лишние степени свободы в модели «тратятся» на чрезмерно точную подгонку под обучающую выборку
 - переобучение есть всегда, когда есть оптимизация по конечной (заведомо неполной) выборке
- Как обнаружить переобучение?
 - эмпирически, разбивая выборку на train и set
- Избавиться от него нельзя, но как минимизировать?
 - минимизировать одну из теоретических оценок
 - накладывать ограничения на (регуляризация)
 - осторожно минимизировать HoldOut, LOO или CV

Эмпирические оценки обобщающей способности

- Эмпирический риск на тестовых данных (HoldOut)

$$HO(\mu, X^I, X^k) = Q(\mu(X^I), X^k) \rightarrow \min$$

- Скользящий контроль (LeaveOneOut, LOO), $L = I + 1$

$$LOO(\mu, X^L) = \frac{1}{L} \sum_{i=1}^L Q(\mu(x^L \setminus \{x_i\}), \{x_i\}) \rightarrow \min$$

- Кросс-проверка (CrossValidation), $L = I + k$, $X^L = X_n^I \sqcup X_n^k$

$$CV(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} Q(\mu(X_n^I), X_n^k) \rightarrow \min$$

- Эмпирическая оценка вероятности переобучения

$$Q_\varepsilon(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} [Q(\mu(X_n^I), X_n^k) - Q(\mu(X_n^I), X_n^I) \geq \varepsilon] \rightarrow \min$$

Медицинская диагностика

Объект — пациент в определённый момент времени

Классы: диагноз / способ лечения / исход заболевания

Примеры признаков:

- **бинарные**: пол, головная боль, слабость, тошнота, ...
- **порядковые**: тяжесть состояния, желтушность, ...
- **количественные**: возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, ...

Особенности задачи:

- много пропусков в данных
- алгоритм должен быть интерпретируемым
- нужна оценка вероятности (риска, успеха, исхода)

Кредитный скоринг

Объект — заявка на выдачу банком кредита физлицу

Классы: bad или good

Примеры признаков:

- **бинарные**: пол, наличие указанного телефона, ...
- **номинальные**: место проживания, профессия, ...
- **порядковые**: образование, должность, ...
- **количественные**: возраст, зарплата, стаж работы, доход семьи, сумма кредита, ...

Особенности задачи:

- нужно оценивать вероятность дефолта обязательств $P(\text{bad})$

Предсказание оттока клиентов

Объект — абонент в определённый момент времени

Классы: уйдёт или не уйдёт в следующем месяце

Примеры признаков:

- **бинарные**: корпоративный клиент, подключённость услуг, ...
- **номинальные**: тарифный план, регион проживания, ...
- **количественные**: длительность разговоров, доля звонков на номера других операторов, количество СМС, частота оплаты, ...

Особенности задачи:

- нужно оценивать ухода
- сверхбольшие выборки
- по «сырым» данным не видно полезных признаков

Категоризация текстовых документов

Объект — текстовый документ

Классы: рубрики иерархического тематического каталога

Примеры признаков:

- **номинальные**: автор, издание, год, ...
- **количественные**: для каждого термина — частота в тексте, в заголовках, в аннотации, ...

Особенности задачи:

- мало документов, для которых известен ответ
- документ может относиться к нескольким рубрикам
- в каждом ребре дерева — свой бинарный классификатор!

Прогноз стоимости недвижимости

Объект — квартира в Москве

Примеры признаков:

- **бинарные**: наличие балкона, лифта, мусоропровода, охраны, парковки, ...
- **номинальные**: район города, тип дома (кирпич/панели/монолит/блоки), ...
- **количественные**: число комнат, жилая площадь, расстояние до центра/МКАД, расстояние до метро, возраст дома, ...

Особенности задачи:

- выборка неоднородна, стоимость изменяется во времени
- разнотипные и разномасштабные признаки
- для линейной модели нужны преобразования признаков

Прогнозирование объёма продаж

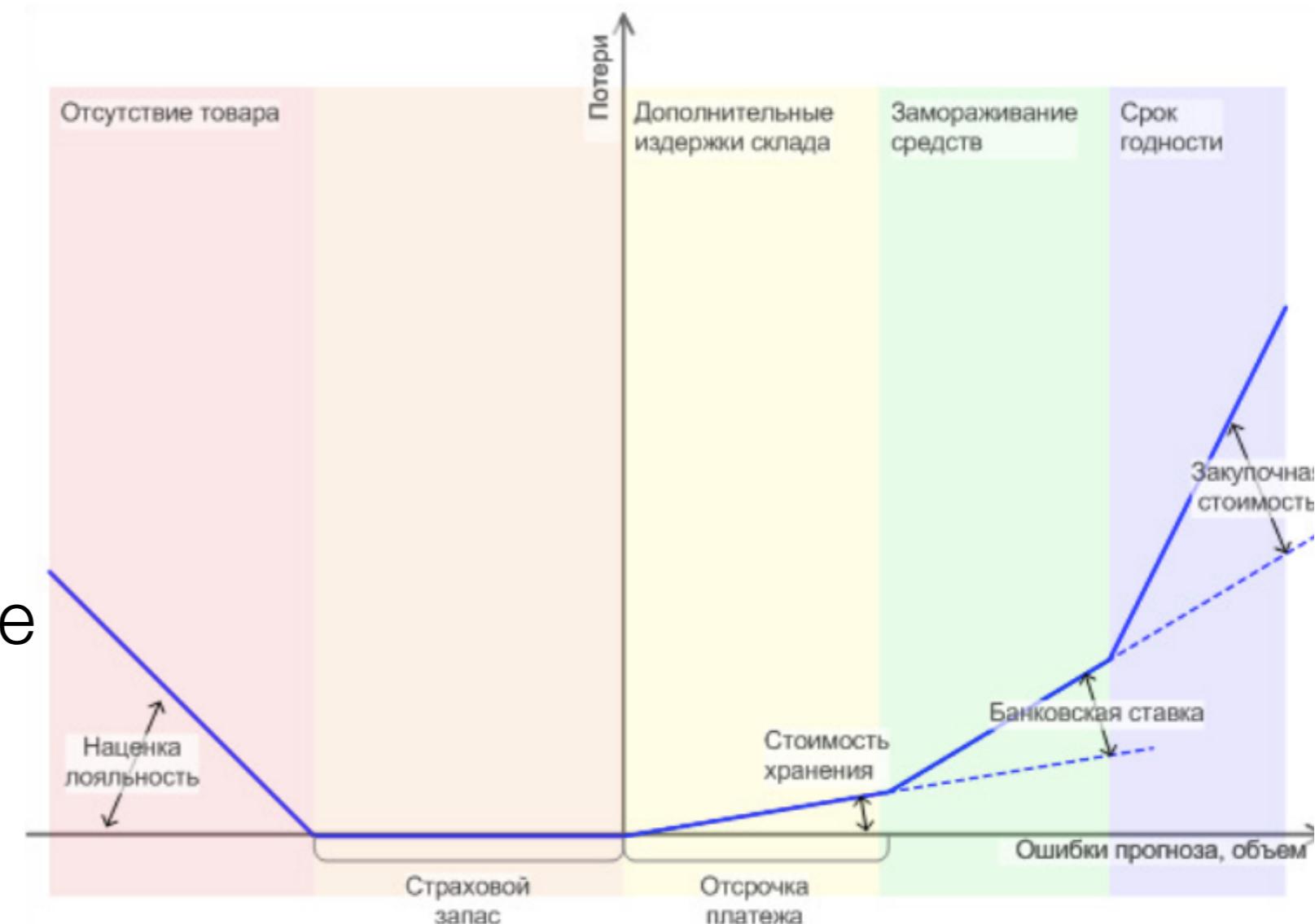
Объект — тройка (товар, магазин, день)

Примеры признаков:

- **бинарные**: выходной, праздник, промоакция, ...
- **количественные**: объём продаж в предыдущие дни и сезоны, ...

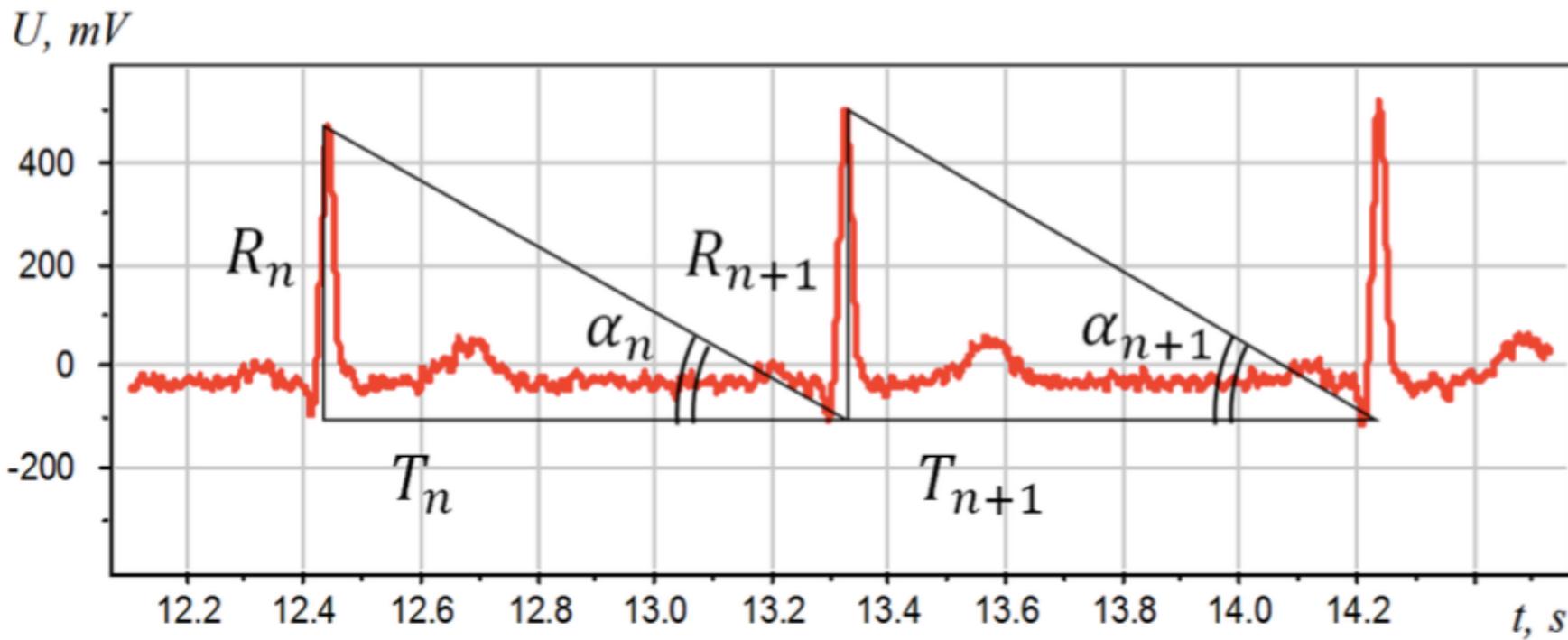
Особенности задачи:

- функция потерь не квадратична и даже не симметрична
- разреженные данные



ЭКГ-диагностика

Объект — электрокардиограмма, 1000 точек/сек



Интервалограмма T_n и амплитудограмма R_n

Их изменчивость зависит от состояния организма

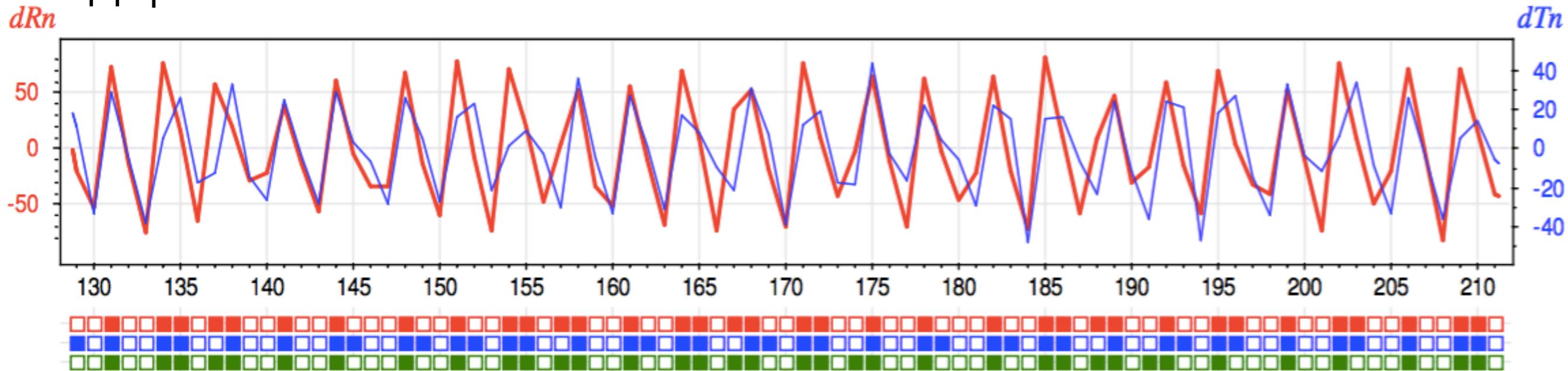
Это знак

$dR_n = R_{n+1} - R_n$	+	-	+	-	+	-
$dT_n = T_{n+1} - T_n$	+	-	-	+	+	-
$d\alpha_n = \alpha_{n+1} - \alpha_n$	+	+	+	-	-	-
кодограмма[n]	A B C D E F					

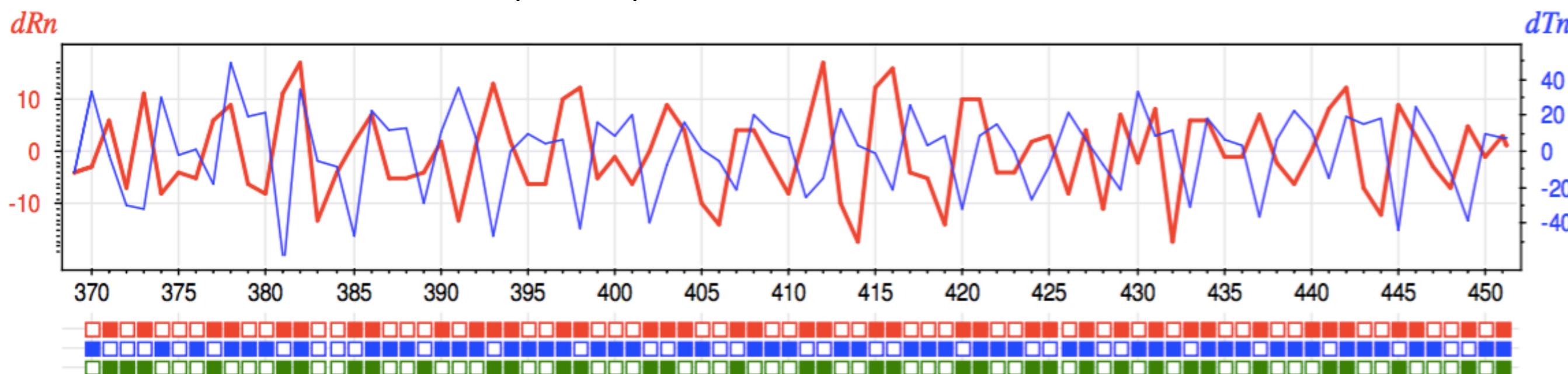
ЭКГ-диагностика

Приращения (в кардиоциклах n) dR_n , dT_n , $d\alpha_n$

Здоровый человек:



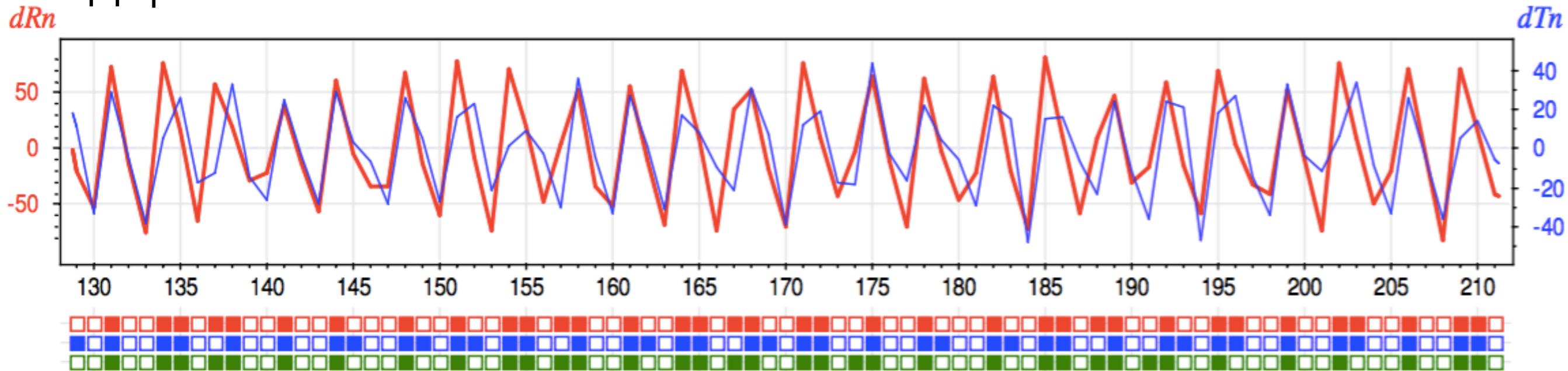
Больной человек (язва):



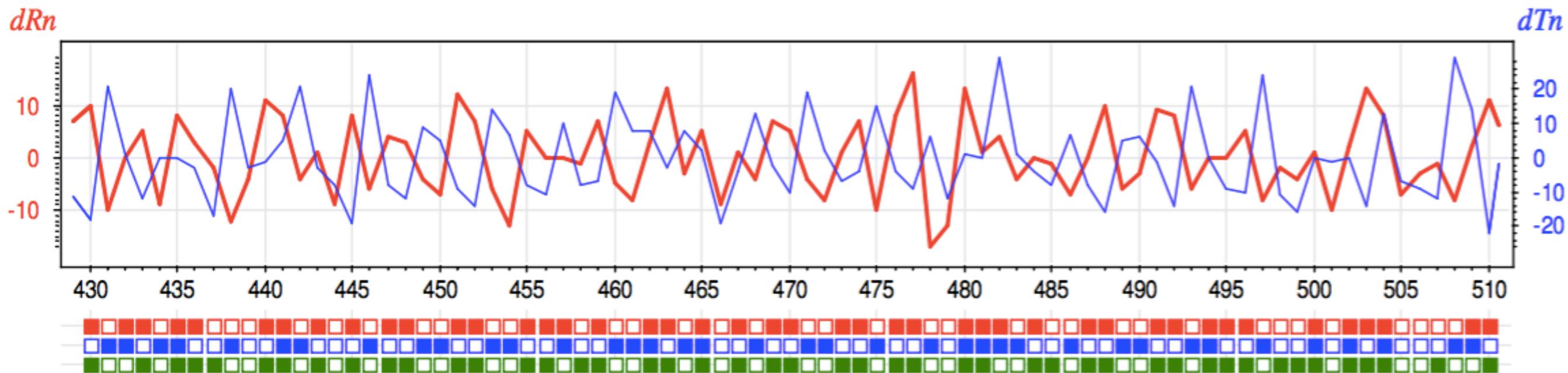
ЭКГ-диагностика

Приращения (в кардиоциклах n) dR_n , dT_n , $d\alpha_n$

Здоровый человек:



Больной человек (рак):



ЭКГ-диагностика

Кодограмма — строка в 6-знаковом алфавите (A-F)

DBF EAC FDA AF BAB DDA AD F A A FF E AC F B A E FF A AB FF A A FF A A FF A A E B F A E B F E A A F C A A F F A A D
F C A F F A A D F C A D F C C D F D A C F F A C D F A E F F A C F F E A D F C A F B C A D F F E C F F A A F F A A F F C A C F C A E F F C A D
D A A D B F A A F F A E B F A A B F A C D F F A A F B A A D F A A D F D A A F C E C F C E D F C E E F C A E F B E C B B B A A D B A A C F F A A F F A
C F F C E C F D A A B D A E F F A A F F C E D B F A A F F A E F B A C F B A E D F E A A F F C A F F D A A F F A E B D A A D B B A D F D A F F
E A B F C C A F D E E B D E C F F A C F F A A B F A A D F B A A F F A C F F A E F F A C F F C E C F B A A F F A A F F A A F F A A D F B
A A B F A C D F D A E F F A A D B A A E F F E A F B C E C F D E C C F B A A F F A A D F D A C D F A A F F A A D F C A A D F C A E F B A A F F C A D F E
A F F C E C F C E C F F A A F F A B C F D A A A F F A D B F C A E F F A A B F A C B F A A E B F A E B F C A F F B A A F F A A F F D A C F D A A B F B
C A F F A E C F F A C F F A C D F C A D F D A A B F A A E D D A B B F C A C D B A A F F A A F F C A D F A A D F D A C F F A E D F C A C F C A E B C E

Вектор признаков — частоты 216 триграмм

FFA - 42	CFF - 14	EFF - 10	FCE - 8	CEC - 6	CAE - 4	EAC - 3	EAA - 2
FAA - 33	AEF - 13	DAA - 10	AEB - 7	ADB - 5	DAC - 4	DDA - 3	CED - 2
AFF - 32	FDA - 13	ECF - 9	DFD - 7	FFE - 5	DBF - 4	CAC - 3	CAA - 2
AAF - 30	FAE - 12	FFC - 9	ACD - 6	EBF - 5	BFC - 4	EDF - 3	BCA - 2
ADF - 18	FAC - 12	FEA - 9	CDF - 6	CFD - 5	CFB - 4	EFB - 3	BBA - 2
FCA - 18	FBA - 11	DFC - 8	DFA - 6	AFB - 4	AED - 3	DBA - 3	DFF - 2
ACF - 17	BFA - 11	ABF - 8	CAF - 6	AAE - 4	FFF - 3	FCC - 2	BDA - 2
AAD - 15	BAA - 11	AAB - 8	CAD - 6	CFC - 4	FBC - 3	AFC - 2	DAE - 2

У каждой болезни свой набор информативных признаков (частот триграмм), которые вместе встречаются в кодограмме больного человека

Эксперименты

Конкретные прикладные задачи:

- решить задачу как можно лучше
- важно понимание задачи и данных
- важно придумывать информативные признаки
- конкурсы <http://kaggle.com>

Наборы прикладных задач:

- цель — протестировать методы как можно лучше
- нет необходимости понимать суть задачи
- признаки уже кем-то придуманы
- репозиторий 308 задач: <http://archive.ics.uci.edu/ml>

Эксперименты

Тестирование новых методов (преимущество — мы знаем истинную функцию y)

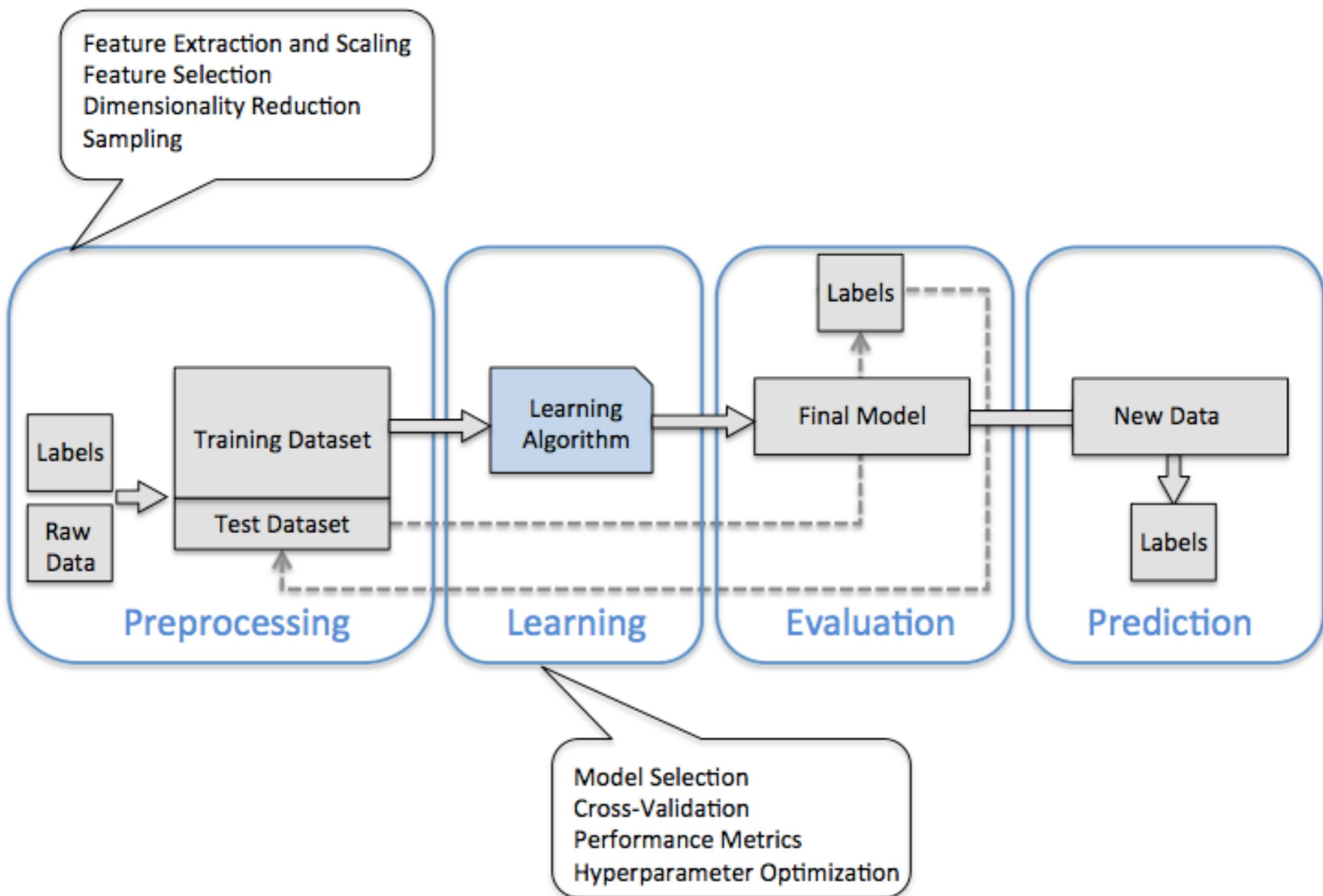
Модельные данные:

- цель — отладить метод, найти границы применимости
- объекты из придуманного распределения
- ответы для придуманной функции

Полумодельные данные:

- цель — тестирование помехоустойчивости модели
- объекты из реальных задач (+ шум)
- ответы для полученного решения (+ шум)

Общий процесс МО



Резюме

Этапы решения задач МО:

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

Терминология:

объект, ответ, признак, алгоритм, модель алгоритмов, метод обучения, эмпирический риск, переобучение

Будем использовать:

Python 2.7, Pandas, NumPy, scikit-learn