

ДОНСКОЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ

КАФЕДРА АВТОМАТИЗАЦИЯ ПРОИЗВОДСТВЕННЫХ ПРОЦЕССОВ

ЛЕКЦИЯ № 02
Представление о
машинном обучении.
Термины и определения.

СОСТАВИТЕЛЬ: КАНД. ТЕХН. НАУК БЫКАДОР В.С.

Определение машинного обучения

Общее определение

Машинным обучением называется систематическое обучение алгоритмов и систем, в результате которого их знания или качество работы возрастают по мере накопления опыта.

Формальное определение

Говорят, что компьютерная программа обучается на опыте E относительно некоторого класса задач T и меры качества P , если её качество на задачах, принадлежащих T , измеренное в соответствии с P , улучшается с увеличением опыта E .

Том Митчелл

Укрупнённая классификация машинного обучения

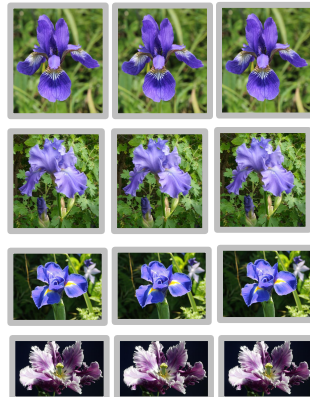


Компоненты машинного обучения

2 Объекты предметной области



1 Обучающий набор



Признаки

Модель машинного обучения

$$\text{Dis}_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^d |x_j - y_j|^p \right)^{\frac{1}{p}}$$
$$\mu = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x}$$

Результаты

Обучение
на данных

1 С начало обучение модели.

2 Затем использование модели.

Признаки

Алгоритм
обучения

Признаки

Признаки позволяют формально описать исследуемый объект.

Другими словами признаки позволяют перейти от непосредственно объекта к абстрактным представлениям о нём. Последнее обстоятельство позволяет математически формализовать описание объекта и затем это математическое описание использовать в моделях машинного обучения.

По сути, модели, определяются в терминах признаков.

Какие могут быть признаки у данного объекта? →



Вариант № 1:

Описать совокупностью пикселей изображения цветка Ириса.

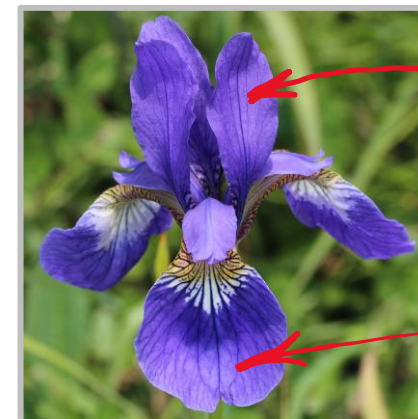
При этом каждый пиксель будет описывать свою яркость одним числом $\in [0 \dots 255]$.

№ пикселя	Яркость
0	0
1	0
2	34
3	231
...	...
5000	5
...	...

Но биологи уже исследовали цветы ириса и выделили 4-ре характерных признака, по анализу которых можно определить вид конкретного ириса. →

Вариант № 2:

- 1) Длина чашелистика.
- 2) Ширина чашелистика.
- 3) Длина лепестка.
- 4) Ширина лепестка.



Лепесток
(petal)

Чашелистик
(sepal)

Модели и задачи

Модель является центральной концепцией машинного обучения, поскольку модель получается в результате обучения на данных с целью решения поставленной задачи.

Модели определяются в терминах признаков.

Разнообразие моделей чрезвычайно большое. Причиной такого большого многообразия моделей является разнообразие задач, которые приходится решать при помощи машинного обучения.

Вот некоторые из основных задач:

- 1) классификация (например выполнить классификацию цветов ириса);
- 2) кластеризация;
- 3) регрессия;
- 4) выявление ассоциаций;
- 5) поиск правил и другие.

Вот некоторые из основных групп моделей:

- 1) линейные модели;
- 2) вероятностные модели;
- 3) метрические модели;
- 4) модели на основе правил;
- 5) древовидные модели и другие.

Примеры задач

Классификация



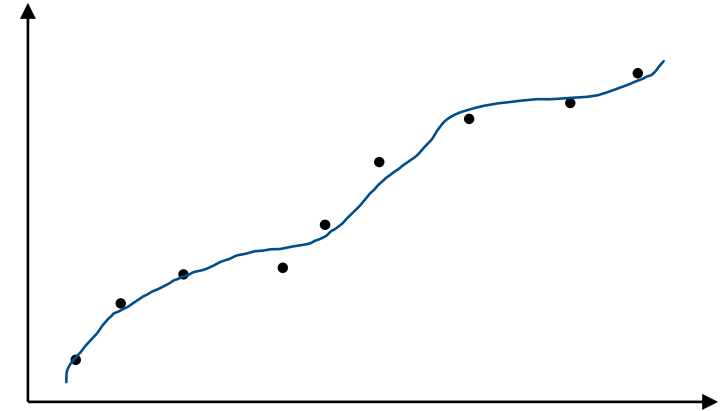
Класс **setosa**

Класс **versicolor**

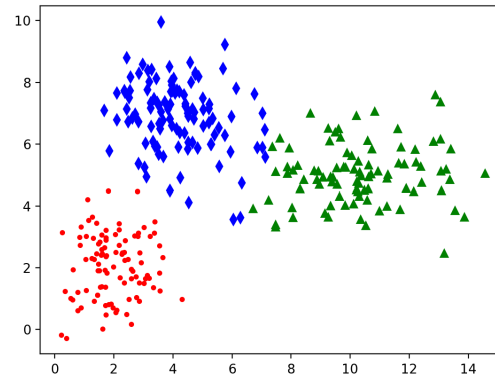
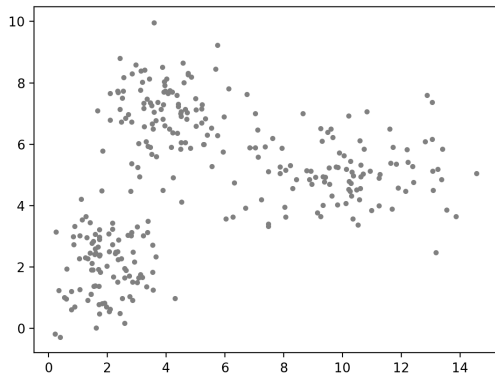
Класс **virginica**



Регрессия



Кластеризация



- выявление ассоциаций;
- поиск правил и другие виды задач.

Примеры моделей

Линейная модель

$$f(\mathbf{x}) = a + b \cdot \mathbf{x}$$

Вероятностная модель

$$P(Y|X) = \frac{P(X, Y)}{\sum_y P(Y = y, X)}$$

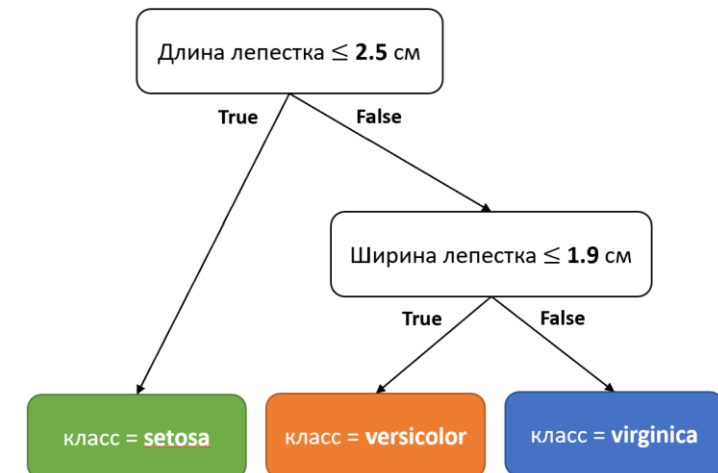
Метрическая модель

$$\text{Dis}_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^d |x_j - y_j|^p \right)^{\frac{1}{p}}$$

Модель на основе правил

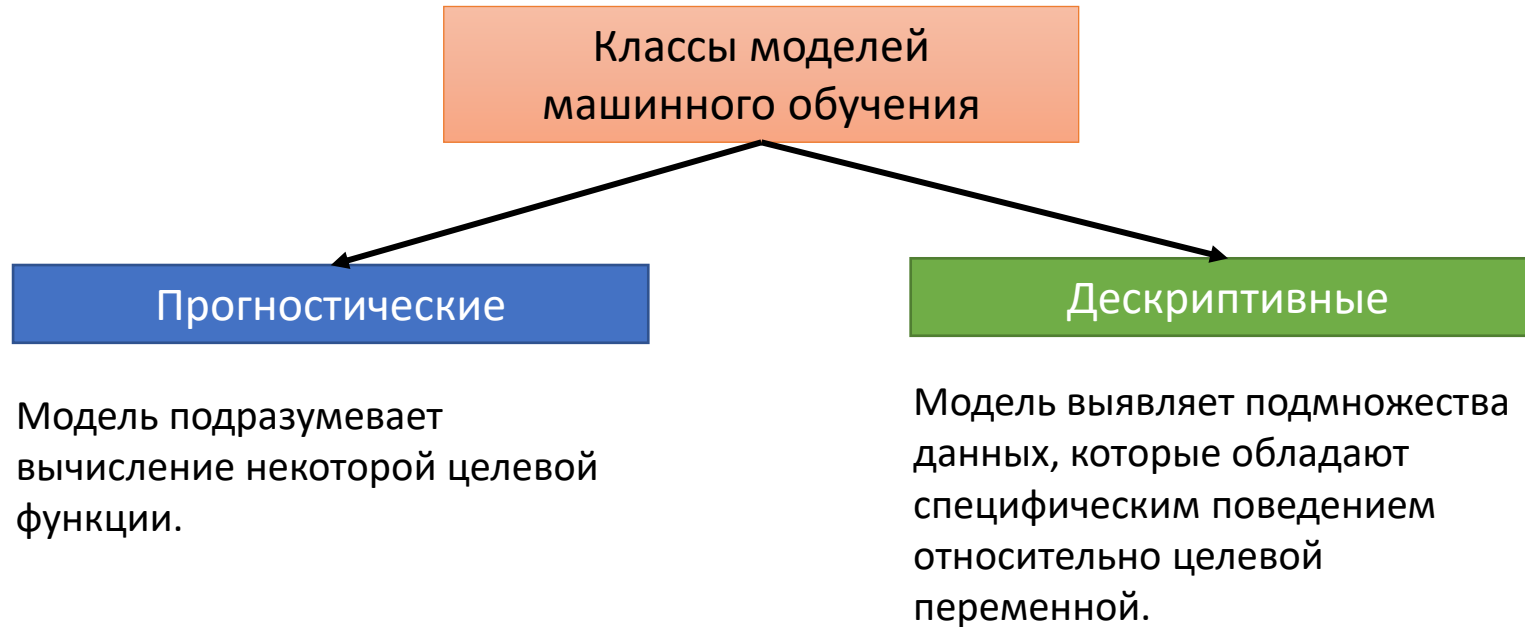
if Жабры = да \wedge Зубы = нет **then** Класс 1
if Жабры = нет \wedge Зубы = да **then** Класс 2
if Жабры = да \wedge Зубы = да **then** Класс 3

Древовидная модель



Модели машинного обучения

Модели машинного обучения можно разделить на два больших класса.



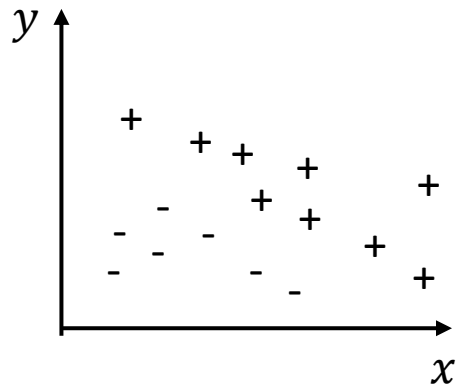
С учётом способов обучения машинных моделей, можно получить следующие разновидности машинного обучения.

	Прогностическая модель	Дескриптивная модель
Обучение с учителем	Классификация, регрессия	Выявления подгрупп
Обучение без учителя	Прогностическая кластеризация	Дескриптивная кластеризация, выявление ассоциативных правил

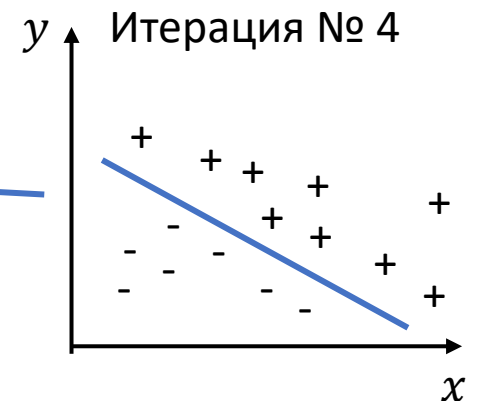
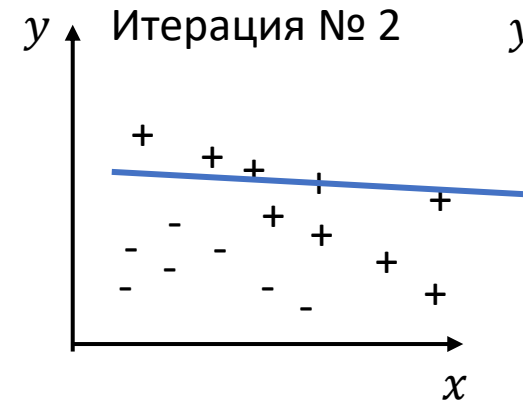
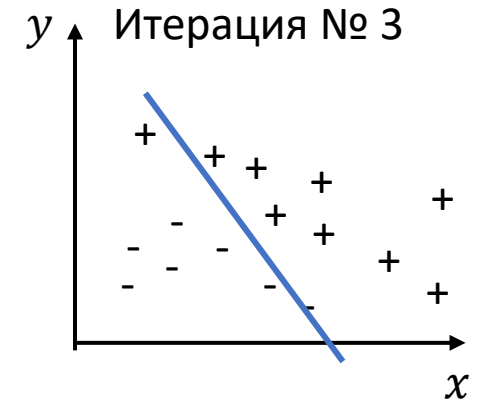
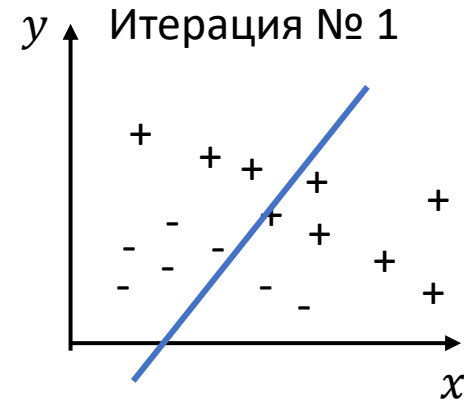
Представление об алгоритме обучения

Рассмотрим на примере линейного классификатора.

Пусть имеется два набора данных «+» и «-» и необходимо найти разделяющую их границу, которую можно задать следующей линейной моделью: $y(x) = a \cdot x + b$.



Тогда задача алгоритма обучения, искать такие значения коэффициентов a и b , чтобы наилучшим образом разделить два типа объектов.



Таким образом алгоритм обучения подбирает коэффициенты модели, то есть «обучает» модель.

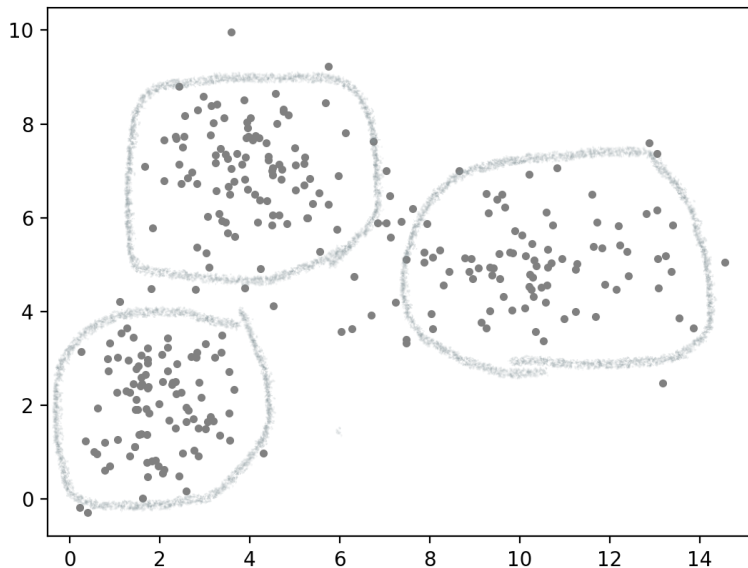
Представление об алгоритме обучения

Рассмотрим на примере кластеризации методом K средних.

Пусть имеется произвольный набор данных о котором известно, что у него имеет три кластера.

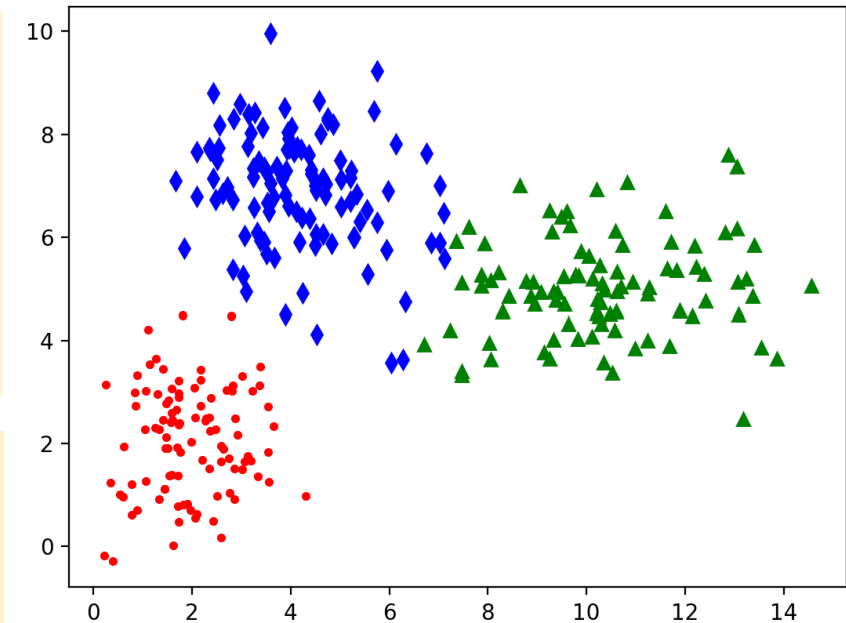
Необходимо разделить произвольный набор данных на эти три кластера, используя метрическую модель:

$$Dis_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^d |x_j - y_j|^2}, \mathbf{y} = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x}$$



Задача алгоритма обучения => отыскать все точки для каждого кластера. Анализируя их расстояния до предполагаемых центров каждого кластера. Центры алгоритм тоже вычисляет.

Таким образом, алгоритм распределил точки по кластерам, на основе выбранной модели машинного обучения.



Типы машинного обучения

```
graph TD; A[Типы машинного обучения] --> B[Машинное обучение с учителем]; A --> C[Машинное обучение без учителя]; A --> D[Машинное обучение с подкреплением];
```

Машинное обучение с учителем

Машинное обучение без учителя

Машинное обучение с подкреплением

Машинное обучение с учителем

Машинное обучение с учителем является одним из самых наиболее часто используемых и успешных видов машинного обучения.

Основной машинного обучения с учителем является наличие так называемых размеченных данных, то есть когда для выборки данных уже заранее известны правильные ответы. Модель обучается на размеченном наборе данных, такой набор называется **обучающим набором данных**. После того как модель будет обучена на размеченном наборе данных, она может применяться для анализа новых данных, которые ранее не встречались. Естественно данные должны быть из той же области, что и размеченные данные.

Формально процесс машинного обучения с учителем, можно представить так: *задача T заключается в том, чтобы обучиться отображению f множества входов $x \in \mathcal{X}$ (признаков) во множество выходов $y \in \mathcal{Y}$.*

Обучающий набор данных формирует как раз тот опыт E , который необходим системе машинного обучения.

Для машинного обучения с учителем требуется размеченный набор данных, который как правило подготавливается человеком.



$$\begin{matrix} \text{red arrow} \\ \rightarrow \end{matrix} x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \xrightarrow{\quad f \quad} y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

где: x_i - длина, ширина чашелистика и лепестка.

y_i - классы setosa, versicolor, virginica.

Машинное обучение с учителем

Небольшие обучающие наборы данных, с небольшим количеством признаков, представляют в виде так называемой матрицы плана, размером $N \times D$.

Где

N – количество строк в матрице, которое представляет количество размеченных примеров;

D – количество столбцов в матрице, которое представляет количество признаков.

$N \gg D$ – то это «большие данные».

$D \gg N$ – то это «широкие данные».

Пример матрицы плана для ирисов

Номер образца	Длина чашелистика, см	Ширина чашелистика, см	Длина лепестка, см	Ширина лепестка, см	Метка класса
0	5.2	3.4	1.6	0.1	setosa
1	4.9	3.0	1.5	0.3	setosa
...					
50	7.1	3.0	4.5	1.1	versicolor
51	6.8	3.2	4.2	1.4	versicolor
...					
148	5.2	3.1	5.5	2.1	virginica
149	5.1	3.2	5.2	1.9	virginica

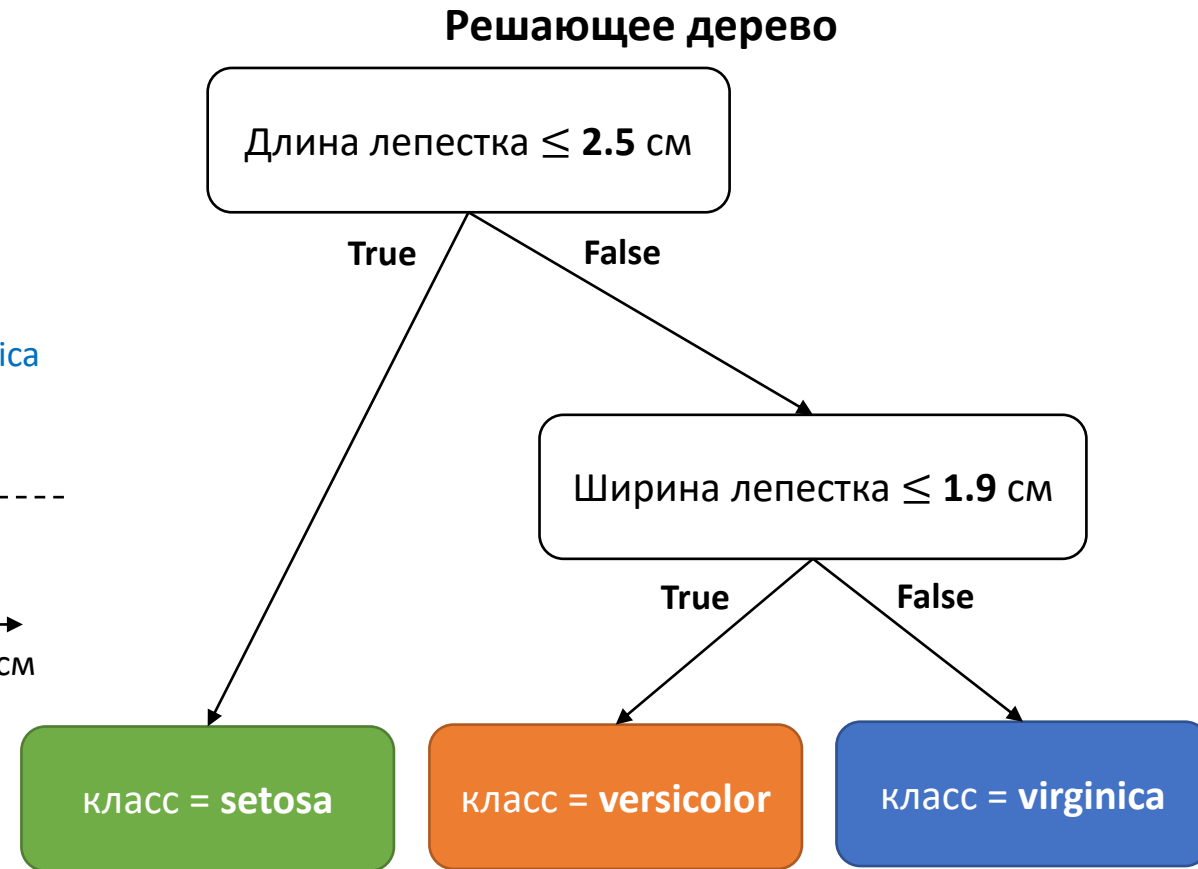
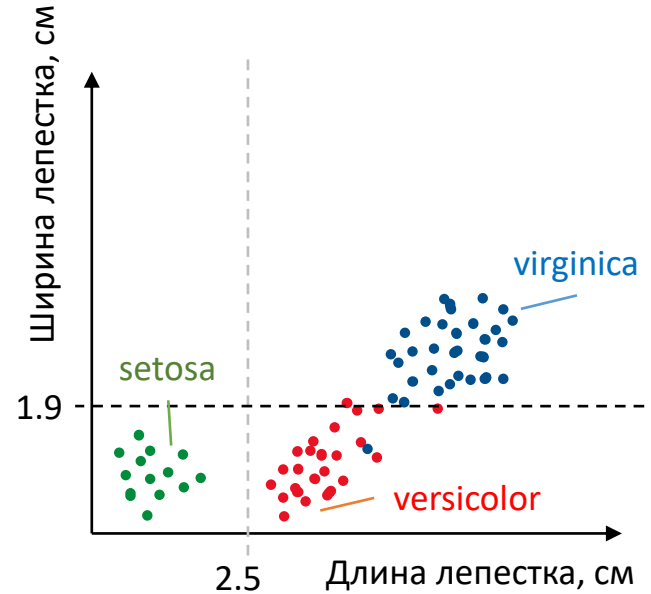
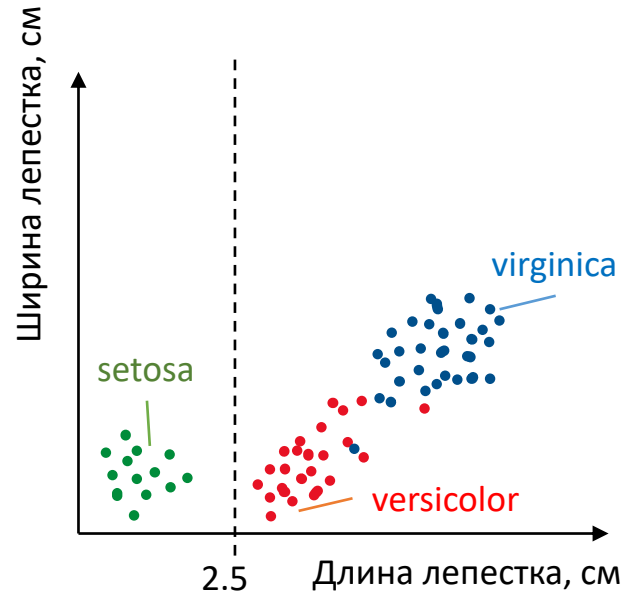
$D = 4$

Машинное обучение с учителем

Схематично, то есть в общем виде без точного алгоритма, рассмотрим обучение модели машинного обучения с учителем.

В качестве модели можем выбрать, например, **решающее дерево**.

Можно выполнить визуализацию данных по парам признаков и исследовать как данные распределены в координатах соответствующих признаков. Допустим, что достаточно рассмотреть пару признаков «длина лепестка – ширина лепестка».



Обобщающая способность, переобучение и недообучение модели

В **машинном обучении с учителем** необходимо помеченные данные разделять на два набора:

1. обучающий набор данных ($\approx 75\%$ от размеченного набора);
2. тестовый набор данных ($\approx 25\%$ от размеченного набора).

Это связано с тем, что модель обучается на обучающем наборе, а проверка правильности её обучения выполняется на тестовом наборе данных. **Нельзя проверять правильность обученности машинной модели по обучающему набору.**

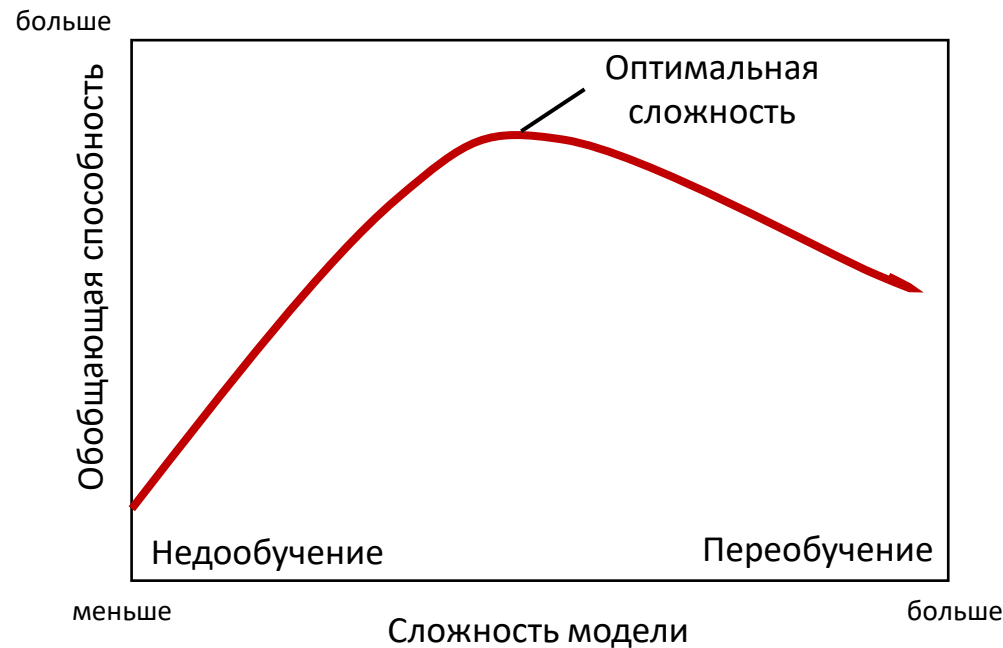
Обобщающая (generalize) способность. Если машинная модель может выдавать точные прогнозы на ранее не встречающихся данных (тестовый набор), то говорят, что машинная модель обладает **способностью обобщать**.

Переобучение (overfitting) имеет место, когда выбранная машинная модель слишком точно подстраивается под конкретные особенности обучающего набора данных. В результате переобученная машинная модель превосходно работает на обучающем наборе данных, но не способна обобщить результаты на новых данных. Переобучение происходит когда машинная модель слишком сложна для имеющегося объёма информации в обучающем наборе данных.

Недообучение (underfitting) имеет место для слишком простых моделей машинного обучения, которые не позволяют описать всё многообразие и изменчивость данных даже на обучающем наборе.

Обобщающая способность, переобучение и недообучение модели

Таким образом существует некоторая оптимальная точка в сложности модели, которая позволяет получить наилучшую обобщающую способность для модели машинного обучения.



Чем больше разнообразных точек в обучающем наборе данных, тем более сложную модель машинного обучения можно использовать не беспокоясь о её переобучении.

НО! Простое дублирование или сбор очень похожих данных **НЕ УВЕЛИЧИВАЕТ** разнообразие данных!

Обобщающая способность, переобучение и недообучение модели

Простой пример. Магазин по продаже лодок хочет получить машинную модель, которая будет давать ответ «кто их потенциальный покупатель».

№	Возраст	Есть авто	Дом/кв.	Кол-во детей	Семейное положение	Купил лодку
1	65	Да	Дом	2	Женат	Нет
2	50	Да	Дом	3	Женат	Да
3	22	Нет	Кв	0	Холост	Нет
4	25	Нет	Кв	1	Разведён	Нет
5	44	Нет	Кв	1	Разведён	Нет
6	39	Да	Кв	2	Женат	Нет
7	53	Да	Дом	2	Разведён	Да
8	64	Нет	Дом	3	Женат	Да
9	58	Да	Дом	3	Женат	Да
10	33	Нет	Кв	2	Разведён	Нет

Обобщающая способность, переобучение и недообучение модели

Простой пример. Магазин по продаже лодок хочет получить машинную модель, которая будет давать ответ «кто их потенциальный покупатель».

№	Возраст	Есть авто	Дом/кв.	Кол-во детей	Семейное положение	Купил лодку
2	50	Да	Дом	3	Женат	Да
8	64	Нет	Дом	3	Женат	Да
9	58	Да	Дом	3	Женат	Да

Модель на основе правил:

if Возраст \geq 50 \wedge Дом = TRUE \wedge Детей \geq 3 \wedge Семейное положение = Женат then “наш клиент”.

Такая модель машинного обучения будет **переобученной**, так как слишком мало данных в обучающем наборе, всего N = 10 записей.

if Возраст \geq 50 then “наш клиент”.

Такая модель машинного обучения будет **недообученной**, так как такая модель скорее всего не позволит охватить всё многообразие и изменчивость данных.

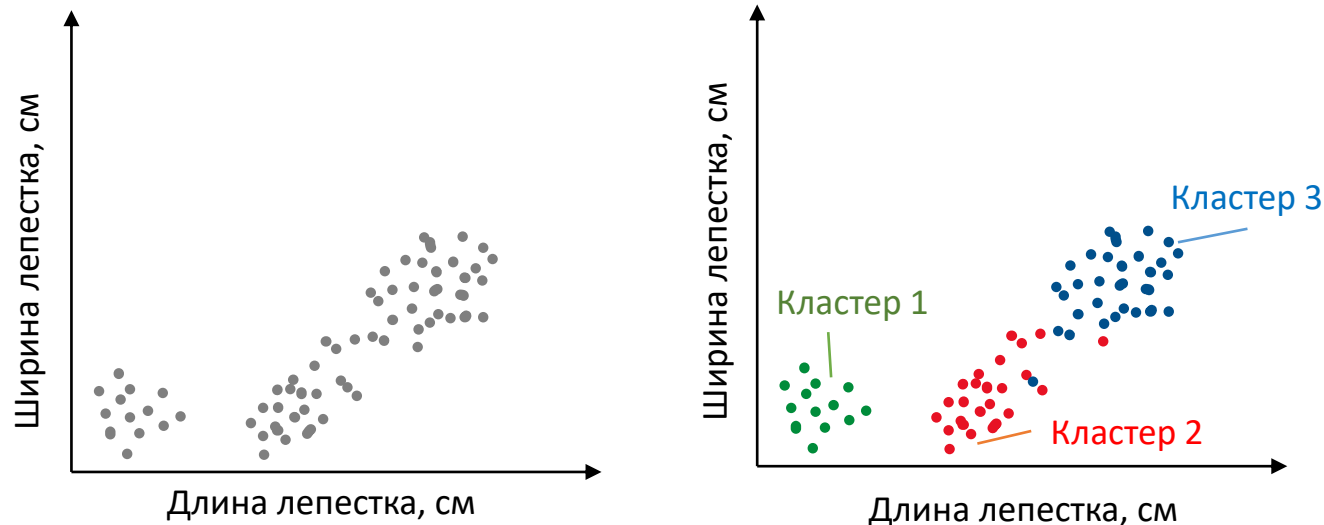
Машинное обучение без учителя

Машинное обучение с учителем по сути «тренирует» машинную модель ассоциировать набор признаков с множеством выходных меток. В этом смысле машинное обучение с учителем представляет собой что-то наподобие аппроксимации.

Машинное обучение без учителя не требует обучающего набора данных, а пытается «извлечь смысл» из исходных данных. То есть всё, что у нас есть, так это только входные данные и нет никаких размеченных им выходов.

Классическим примером машинного обучения без учителя является уже не раз упоминавшийся выше кластеризация.

Если ввести меру «компактности», например, это может быть метрическое правило, то результатом машинного обучения без учителя должно стать разбиение исходных данных на три кластера.



Математические обозначения

Предмет рассмотрения в машинном обучении называется **объектом** (цветы ириса, сообщения электронной почты).

Множество всех возможных рассматриваемых объектов называется **пространством объектов** \mathcal{X} .

Простейшее пространство объектов возникает тогда, когда объекты описываются фиксированным множеством признаков. **Область значений** признака обозначается как \mathcal{F}_i , тогда получим, что каждый объект \mathcal{X} можно записать как $\mathcal{X} = \mathcal{F}_1 \times \mathcal{F}_2 \times \dots \times \mathcal{F}_d$, то есть каждый объект является вектором длины d , состоящим из значений признаков.

Признак = атрибут = прогностический параметр = объясняющая переменная = независимая переменная = ковариат = предиктор.

Пространство меток обозначается как \mathcal{L} (используется в машинном обучении с учителем).

Пространство выходов обозначается как \mathcal{Y} .

Для решения задачи требуется **модель машинного обучения**, которая является отображением пространства объектов в пространство выходов $f: \mathcal{X} \rightarrow \mathcal{Y}$, здесь f модель машинного обучения.

Обучающий набор помеченных объектов $Tr = \{x, l(x)\}$ здесь $l: \mathcal{X} \rightarrow \mathcal{L}$ - помечающая функция.

Тестовый набор объектов $Te = \{x, l(x)\}$.

Когда истинные функции неизвестны, то ищут их аппроксимации. В этом случае добавляют знак «крышки» над символом функции и говорят, что это аппроксимация или оценка истинной функции, например $\hat{l}: \mathcal{X} \rightarrow \mathcal{L}$.

Наблюдение может осложняться наличием «шума», который может стать **меточным шумом** и тогда вместо функции $l = l(x)$ будет наблюдаться искажённая метка, которая обозначается как l' . Так же может быть **объектный шум**, тогда вместо объекта x будет наблюдаться искаженный объект x' .

Использованные информационные источники

1. 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.: ил.
2. Мэрфи К. П. Вероятностное машинное обучение: введение / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2022. – 990 с.: ил.
3. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными.: Пер. с англ. - СПб.: ООО "Альфа-книга", 2017. - 480 с.: ил.