Lecture_00

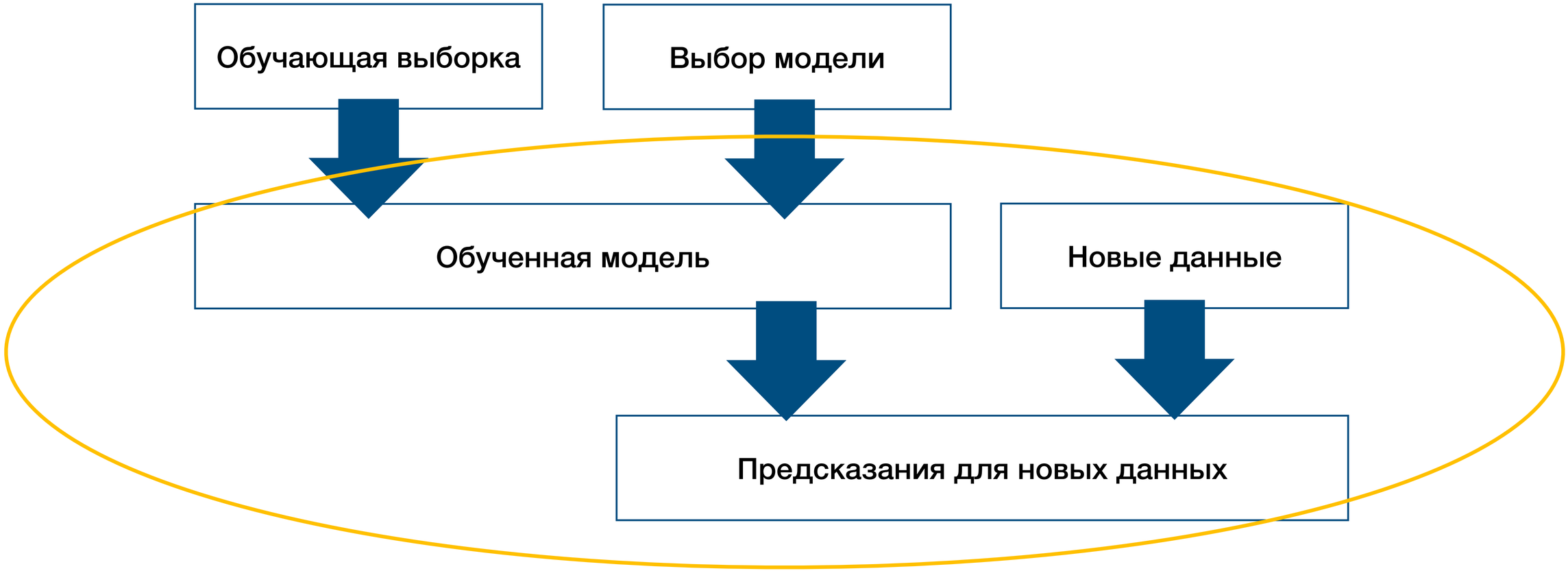
Intro

Историческая справка

ГОД	СОБЫТИЕ
1763	Опубликовано эссе Томаса Байеса, представляющее работу, лежащую в основе Теоремы Байеса.
1805	Лежандр описывает метод наименьших квадратов.
1812	Лаплас публикует работу, в которой определена Теорема Байеса.
1913	Андрей Марков описывает метод, позже называемый «Цепи Маркова».
1950	Алан Тьюринг предлагает концепцию Машинного обучения, предвещающую генетические алгоритмы.
1957	Розенблат изобретает perceptron.
1967	Изобретен метод ближайших соседей
1970	Seppo Linnainmaa публикует общий метод автоматического дифференцирования (AD)
1986	Seppo Linnainmaa применяе обратный режим автоматического дифференцирования

Постановка задачи машинного обучения

Задача: восстановить сложную зависимость по конечному числу примеров



Обучающая выборка

Матрица «объекты–признаки»

Датасет с задержками рейсов.

	Month	DayofMonth	DayOfWeek	DepTime	UniqueCarrier	Origin	Dest	Distance	dep_delayed_15min
0	c-8	c-21	c-7	1934	AA	ATL	DFW	732	N
1	c-4	c-20	c-3	1548	US	PIT	MCO	834	N
2	c-9	c-2	c-5	1422	XE	RDU	CLE	416	N
3	c-11	c-25	c-6	1015	OO	DEN	MEM	872	N
4	c-10	c-7	c-6	1828	WN	MDW	OMA	423	Y

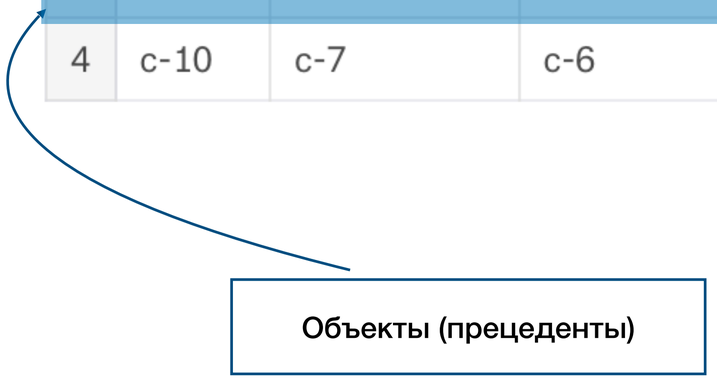
Источник: https://www.transtats.bts.gov

Обучающая выборка

Матрица «объекты–признаки»

Датасет с задержками рейсов.

	Month	DayofMonth	DayOfWeek	DepTime	UniqueCarrier	Origin	Dest	Distance	dep_delayed_15min
0	c-8	c-21	c-7	1934	AA	ATL	DFW	732	N
1	c-4	c-20	c-3	1548	US	PIT	MCO	834	N
2	c-9	c-2	c-5	1422	XE	RDU	CLE	416	N
3	c-11	c-25	c-6	1015	OO	DEN	MEM	872	N
4	c-10	c-7	c-6	1828	WN	MDW	OMA	423	Y



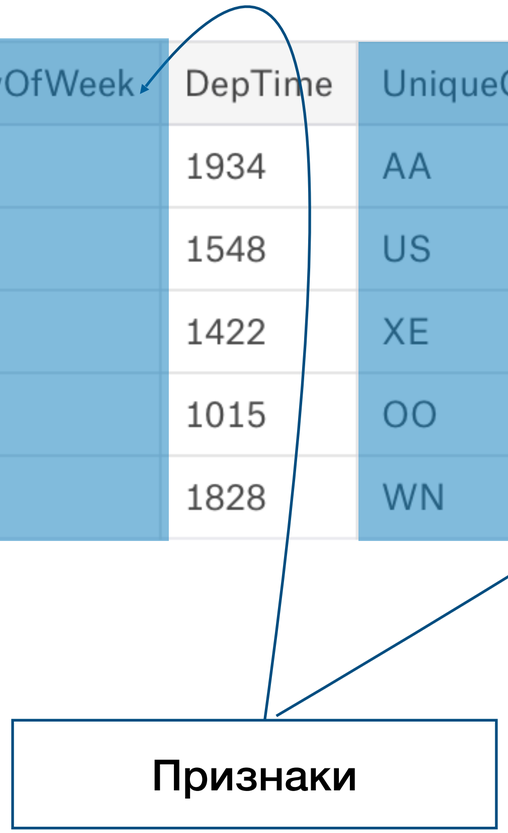
Источник: https://www.transtats.bts.gov

Обучающая выборка

Матрица «объекты–признаки»

Датасет с задержками рейсов.

	Month	DayofMonth	DayOfWeek	DepTime	UniqueCarrier	Origin	Dest	Distance	dep_delayed_15min
0	c-8	c-21	c-7	1934	AA	ATL	DFW	732	N
1	c-4	c-20	c-3	1548	US	PIT	MCO	834	N
2	c-9	c-2	c-5	1422	XE	RDU	CLE	416	N
3	c-11	c-25	c-6	1015	OO	DEN	MEM	872	N
4	c-10	c-7	c-6	1828	WN	MDW	OMA	423	Y



Источник: https://www.transtats.bts.gov

Обучающая выборка

Матрица «объекты–признаки»

Датасет с задержками рейсов.

	Month	DayofMonth	DayOfWeek	DepTime	UniqueCarrier	Origin	Dest	Distance	dep_delayed_15min
0	c-8	c-21	c-7	1934	AA	ATL	DFW	732	N
1	c-4	c-20	c-3	1548	US	PIT	MCO	834	N
2	c-9	c-2	c-5	1422	XE	RDU	CLE	416	N
3	c-11	c-25	c-6	1015	OO	DEN	MEM	872	N
4	c-10	c-7	c-6	1828	WN	MDW	OMA	423	Y

Целевая переменная

Источник: https://www.transtats.bts.gov

Признаки

Признаковое описание объекта - Вектор:

$x_i = \{d_1, d_2, d_3, \ldots d_n\}$

Множество значений признака

$d_j \in D_j$

Бинарные признаки

$D_j = \{0, 1\}$

В нашем примере:
Целевая переменная

Категориальные признаки

D_j - упорядоченное множество

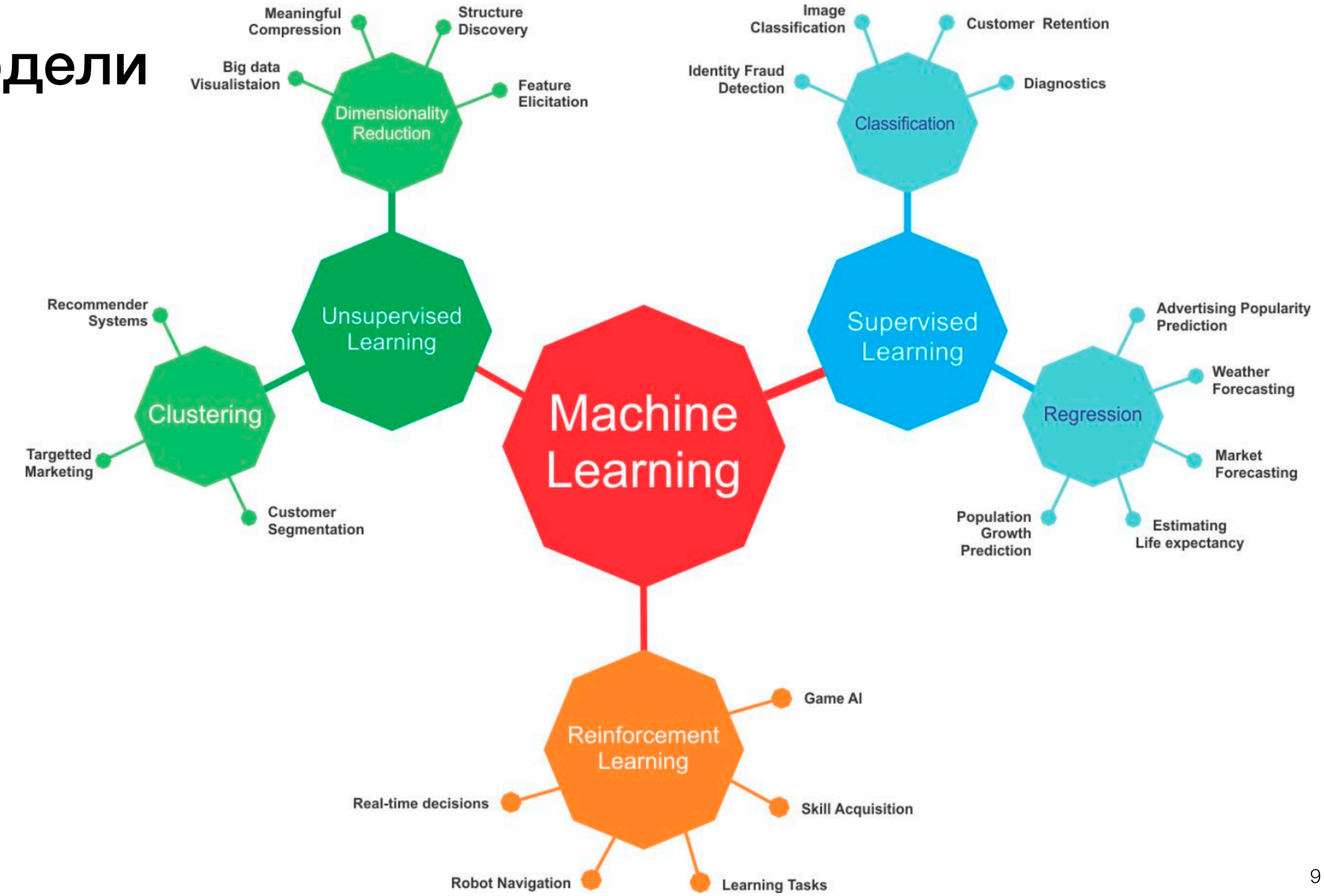
В нашем примере:
Локация отправления
Локация прибытия

Вещественные признаки

$D_j = \mathbb{R}^m$

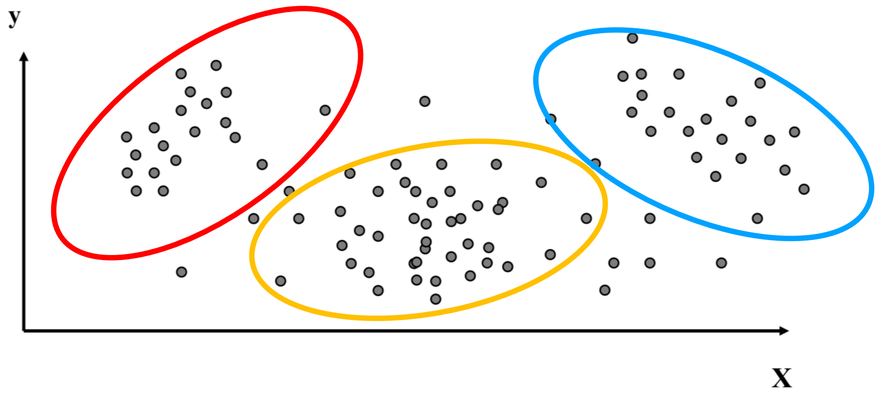
В нашем примере:
Расстояние

Выбор модели

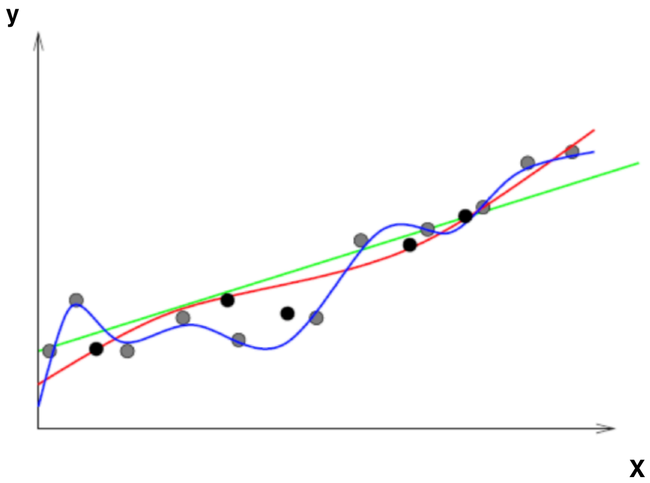


Формальная постановка задачи

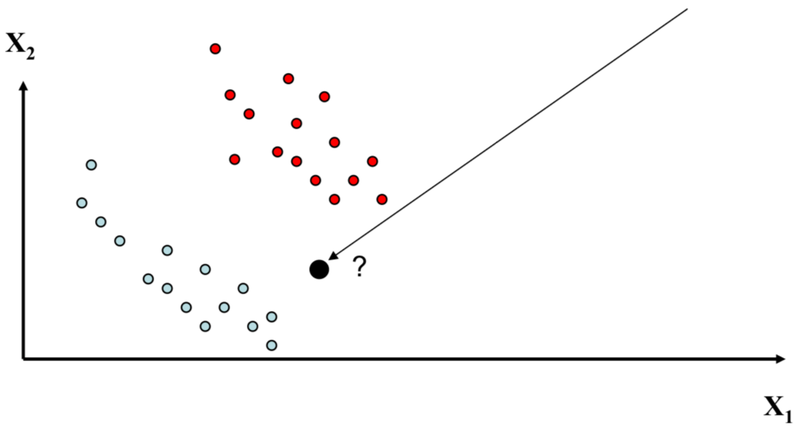
Кластеризация



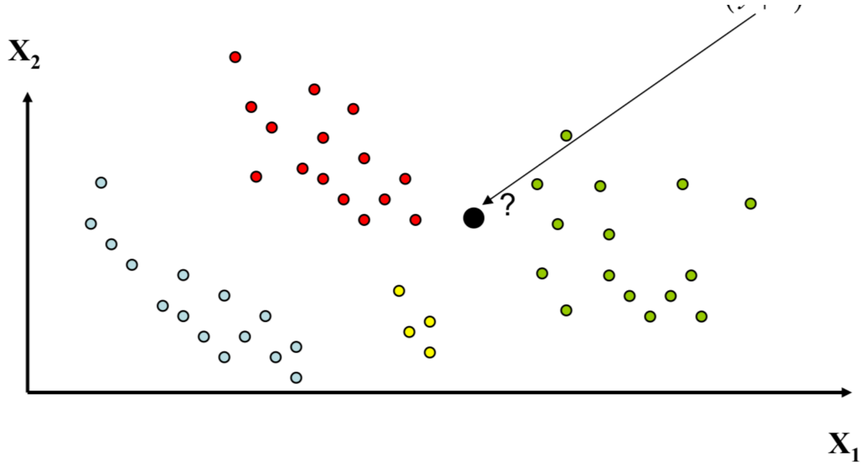
Регрессия



Классификация



Классификация - многокласс



Где взять данные?

- Google Dataset Search. Dataset Search позволяет по ключевому слову искать датасеты по всей Сети.
- Kaggle. Площадка для соревнований по машинному обучению с множеством интересных датасетов. В списке датасетов можно найти разные нишевые экземпляры — от оценок рамена до баскетбольных данных NCAA и базы лицензий на домашних животных в Сиэтле.
- UCI Machine Learning Repository. Один из старейших источников датасетов в Сети и первое место, куда стоит заглянуть в поиске интересных датасетов. Хотя они добавляются пользователями и потому имеют различную степень «чистоты», большинство из них очищены. Данные можно скачивать сразу, без регистрации.
- VisualData. Датасеты для компьютерного зрения, разбитые по категориям. Доступен поиск.
- Find Datasets | CMU Libraries. Коллекция датасетов, предоставленная университетом Карнеги Меллон.

Больше датасетов: https://tproger.ru/translations/the-best-datasets-for-machine-learning-and-data-science/