Интеллектуальный анализ данных и основы машинного обучения

Data Science Minor, осень 2017

Немного организационного

проверьте доступ к

- RStudio
- Stepik.org
- почте ВШЭ (рассылка все так же через нее)

Что такое машинное обучение

Что такое машинное обучение

обширный раздел искусственного интеллекта, изучающий методы построения алгоритмов, способных **обучаться**

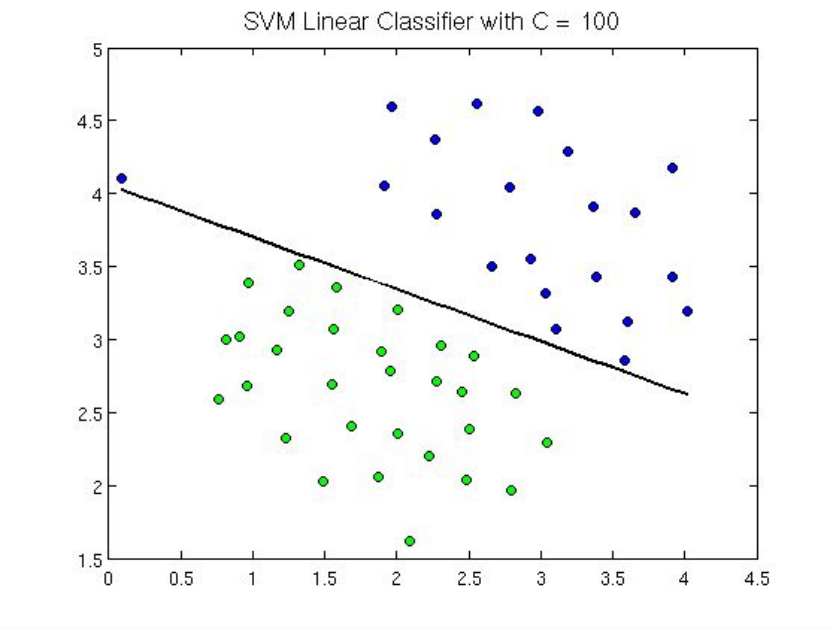
http://www.machinelearning.ru/

Что такое машинное обучение

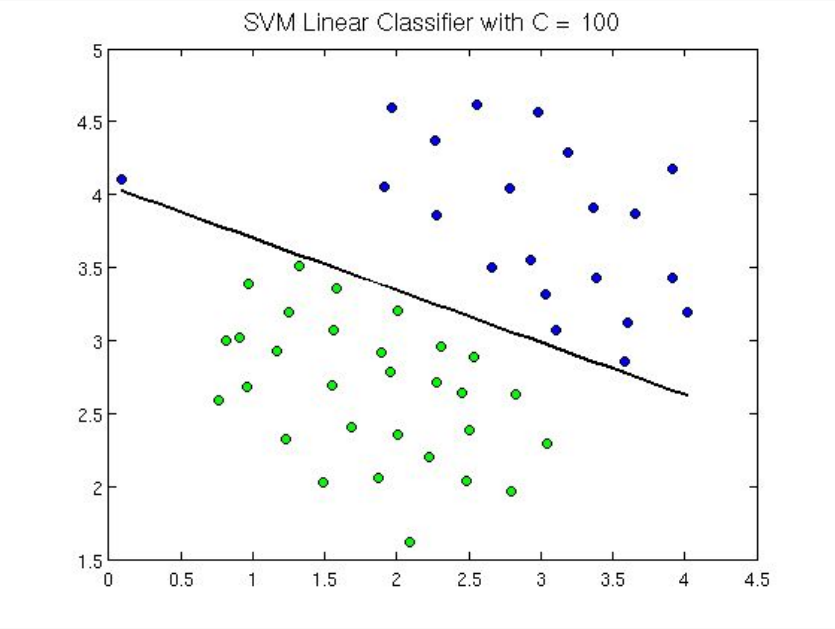
обширный раздел искусственного интеллекта, изучающий методы построения алгоритмов, способных **обучаться**

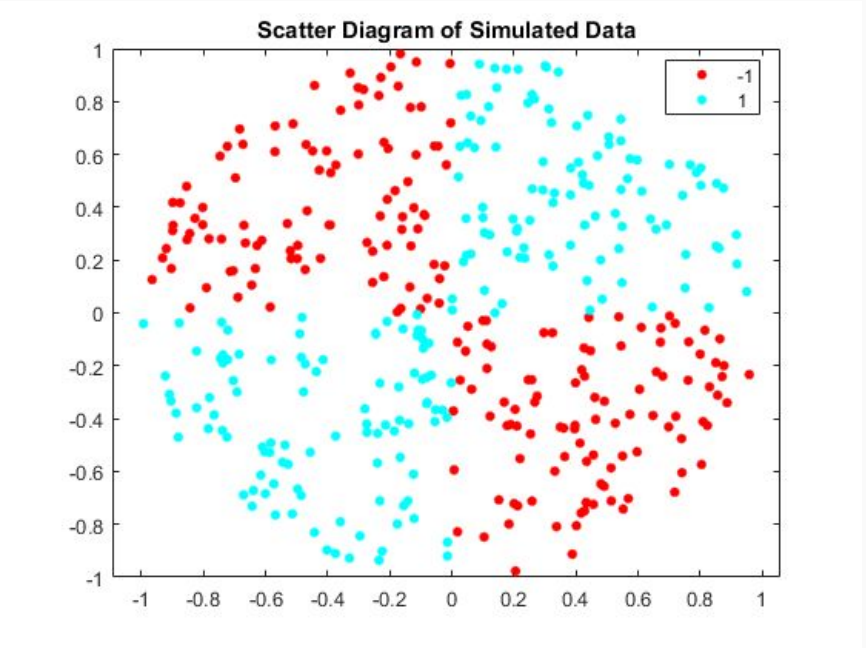
- Зачем нужны алгоритмы, способные обучаться?
- Почему сразу не запрограммировать прямое решение задачи?

Зачем обучаться

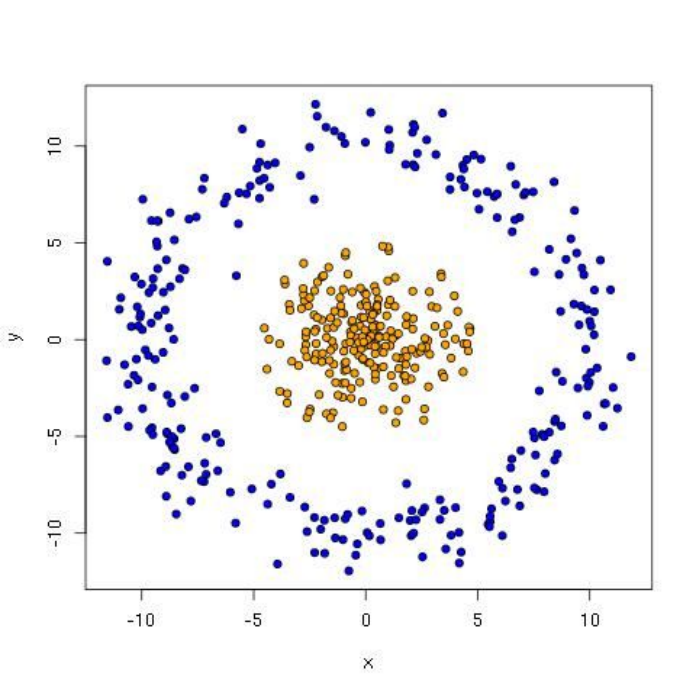


Зачем обучаться

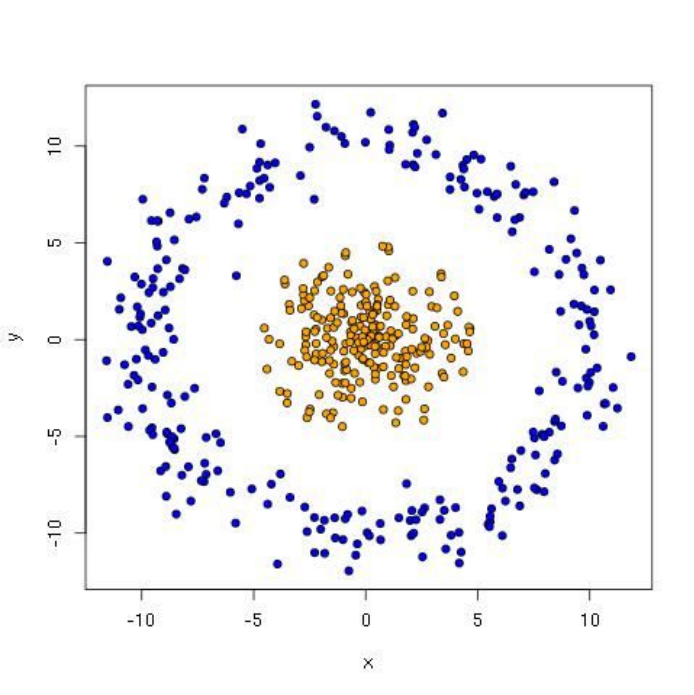


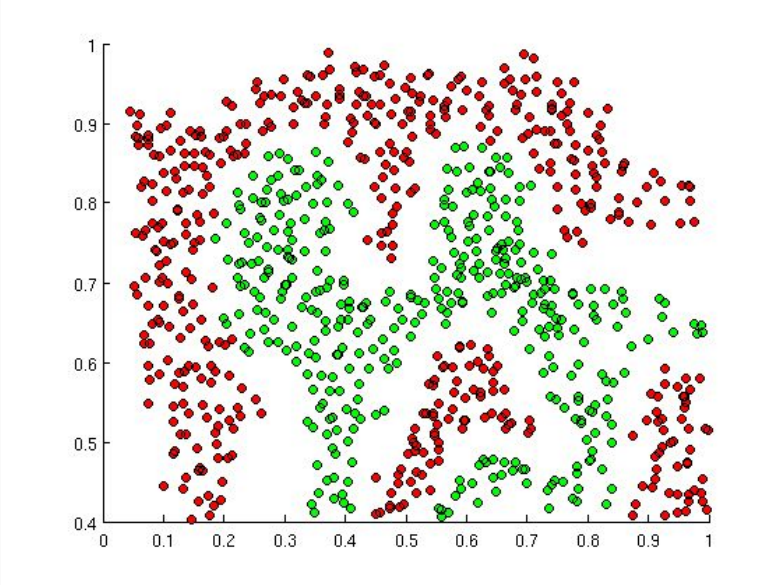


Зачем обучаться



Зачем обучаться





Зачем обучаться

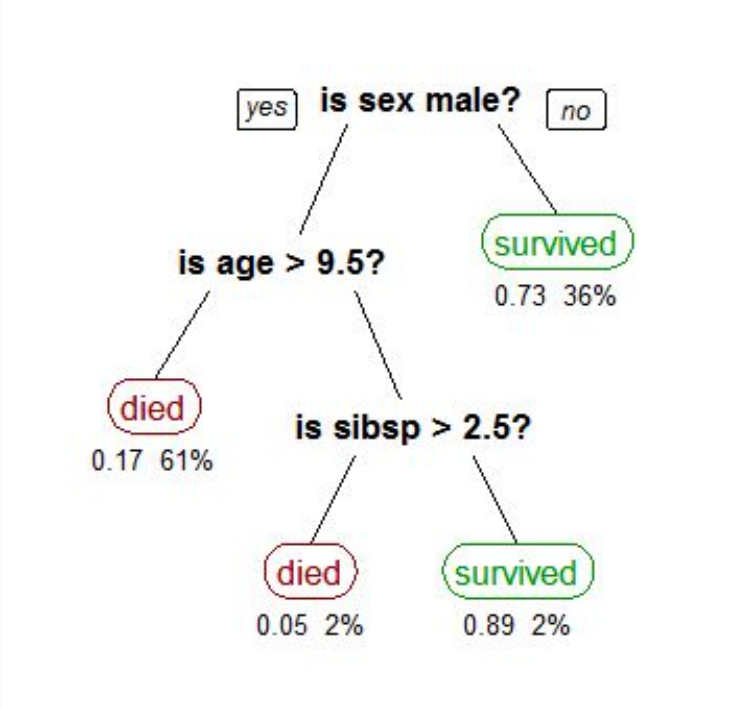
- мы не знаем правило
- большая размерность (много факторов)

Зачем обучаться

- мы не знаем правило
- большая размерность (много факторов)

- зато есть данные -- пусть компьютер (алгоритм) сам конструирует правило!

Пример правила: Decision tree



Классификационное дерево предсказания выживаемости на Титанике

*sibsp = число близких родственников на Титанике

Что мы знаем и умеем

- Основы агрегации данных
- Основные способы визуализации
- Регрессия и классификация (чуть-чуть)
- Рекомендательные системы
- Анализ текста
- Анализ сетей

Что мы знаем и умеем

- **Основы агрегации данных**
- **Основные способы визуализации**
- **Регрессия и классификация (чуть-чуть)**
- Рекомендательные системы
- Анализ текста
- Анализ сетей

Регрессия vs классификация

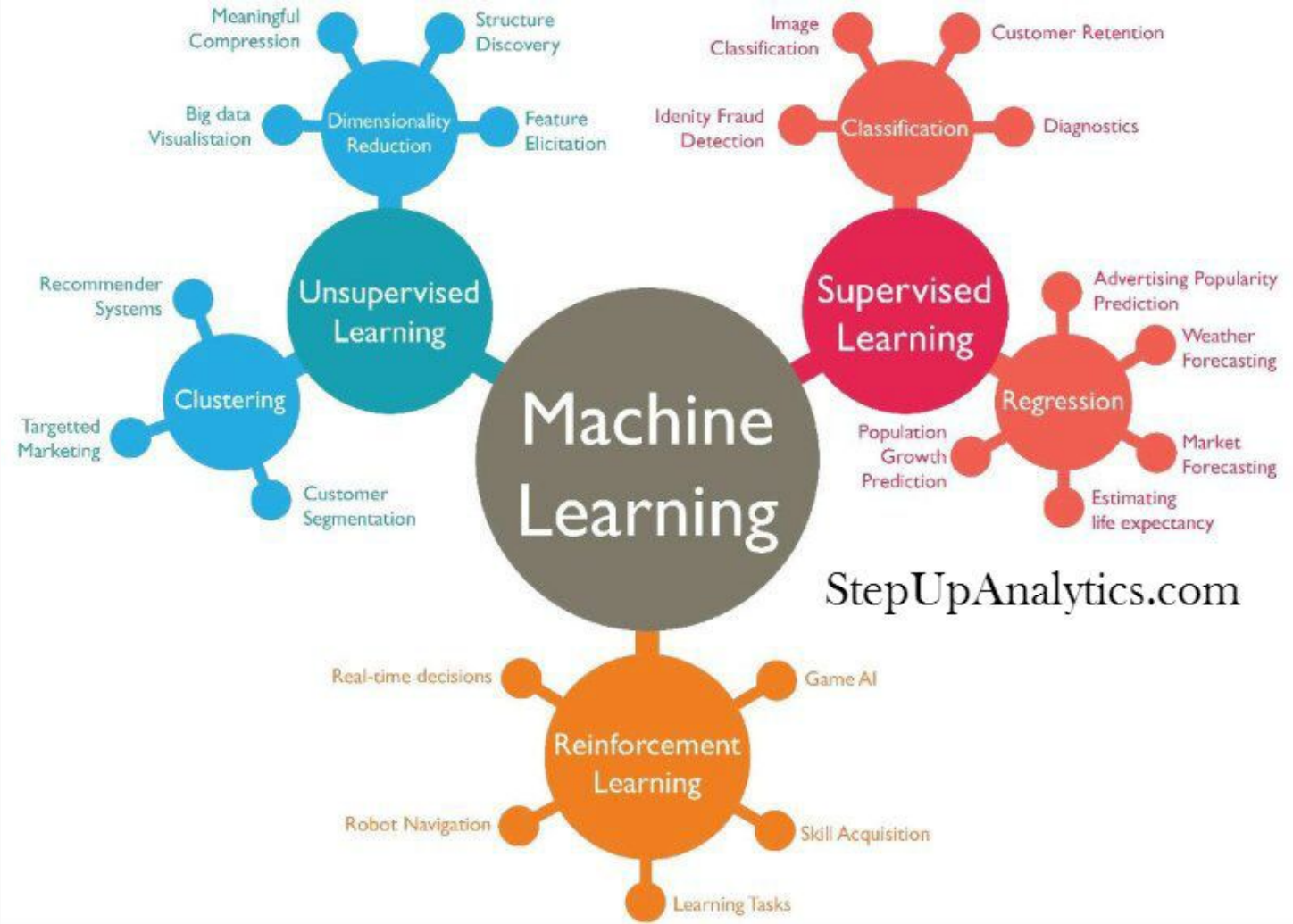
- Определить, есть ли у пациента рак легких, по результатам анализа
- Предсказать оценку студента за курс на основании его предыдущих успехов
- Оценить риск заразиться малярией на основании списка факторов
- Определить, является ли письмо спамом

Куда мы двинемся

- Посмотрим на новые модели
- Разберемся еще раз с тем, что такое хорошо и что такое плохо в предсказаниях
- Узнаем, как измерить, на сколько именно все плохо
- Попробуем понять, как можно выбирать модели

Задачи в машинном обучении

- Обучение с учителем (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами
- Обучение без учителя (unsupervised learning) – обучение, в котором нет правильных ответов, только данные
- Обучение с подкреплением (reinforcement learning) – обучение, в котором агент учится из собственных проб и ошибок



Еще немного о типах задач

Обучение с учителем: задача

Обучение с учителем – обучение, в котором есть некоторое число примеров с правильными ответами:

- обучающая выборка (training set) – набор примеров, каждый из которых состоит из признаков;
- у примеров есть правильные ответы – целевая переменная, которую мы хотим предсказывать; она может быть категориальная, непрерывная или ординальная;
- Задача: предложить способ по признакам определять значение целевой переменной, как на имеющихся данных, так и на данных, которые не вошли в нашу обучающую выборку.

Обучение с учителем: общая схема

- определяем критерии ошибки;
- строим модель, подбираем параметры, минимизирующие ошибку на обучающей выборке;
- проверяем, что на тестовой выборке ошибка приемлемая.

Примеры задач

- **Классификация**: есть некоторый дискретный набор категорий (классов), и надо новые примеры определить в какой-нибудь класс:
  - классификация текстов по темам, спам-фильтр;
  - распознавание лиц/объектов/текста;
- **Регрессия**: есть некоторая неизвестная функция, и надо предсказать её значения на новых примерах:
  - инженерные приложения (предсказать температуру, положение робота);
  - финансы – предсказать цену акций или квартиры

В чем проблема с оценкой качества

- Нужно определить ошибку на одном примере
- Нужно определить ошибку на всех примерах для общей оценки качества модели
- Нужно оценить ошибку на примерах, которых у нас нет
- Нужно понимать, полученная ошибка – это много или мало

Ошибка на одном примере: функция потерь

- Для регрессии:
  - абсолютное отклонение
  - квадратичное отклонение
- Для классификации:
  - совпало / не совпало
  - перекрестная энтропия

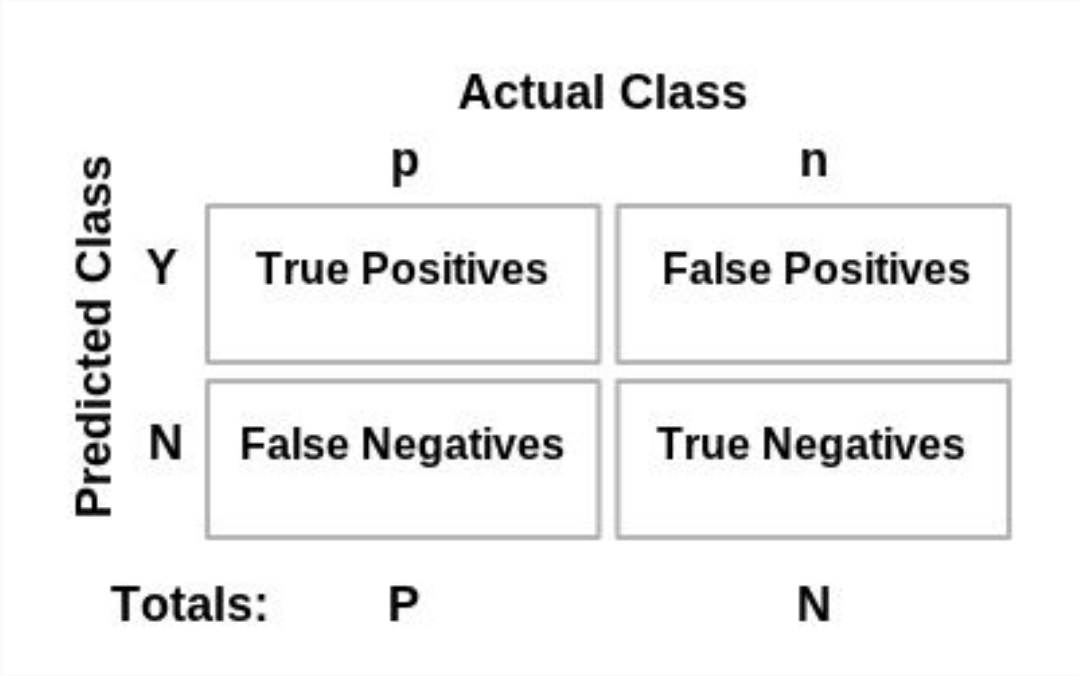
Общая ошибка: эмпирический риск

- Мы знаем ошибку на каждом примере из нашего набора
- Предположим, что каждый пример – равновероятен
- Эмпирический риск – математическое ожидание ошибки (среднее значение)
- Одна из самых распространенных метрик ошибки Root Mean Squared Error (RMSE)

Другие метрики общей ошибки

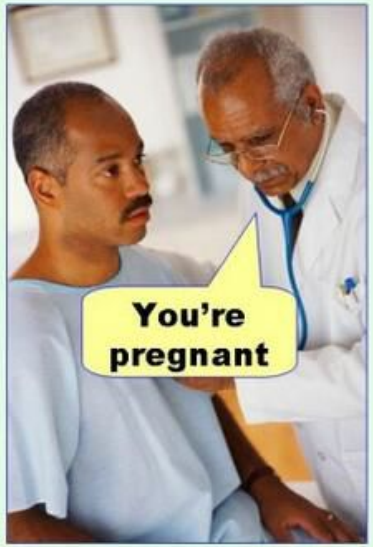
- Для задач классификации, часто рассматривают метрики, основанные на количестве правильных/неправильных ответов:
  - Точность (Accuracy)
  - Precision и recall
  - F-мера

Другие метрики общей ошибки

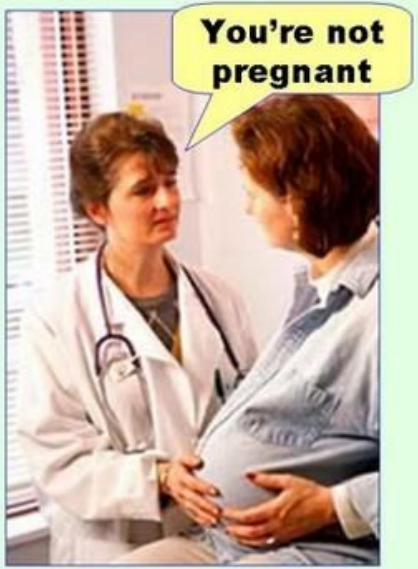


Ошибки первого и второго рода

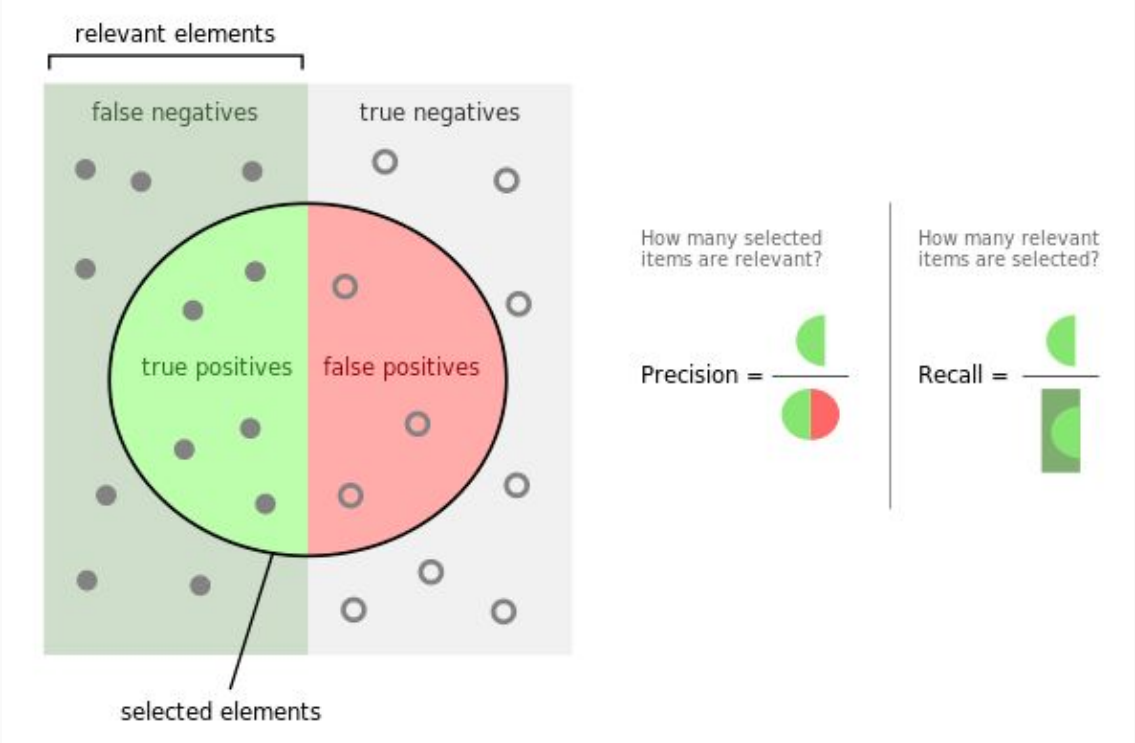
Type I error
(false positive)



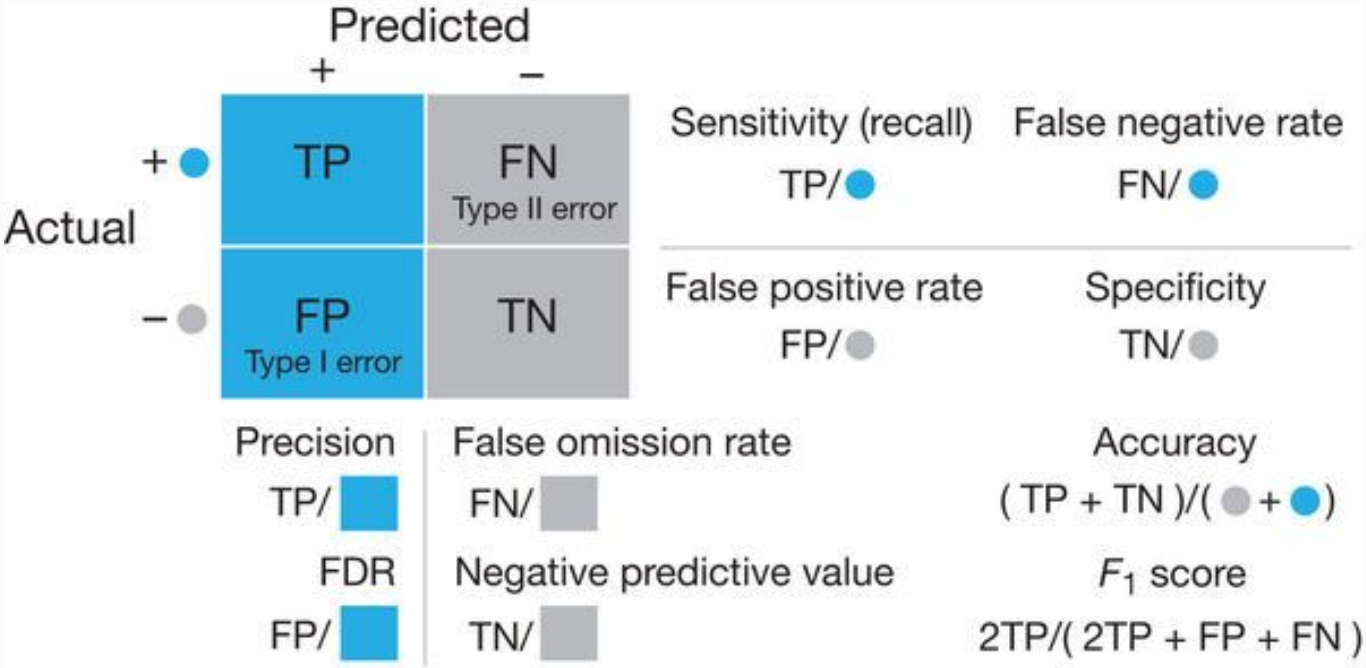
Type II error
(false negative)

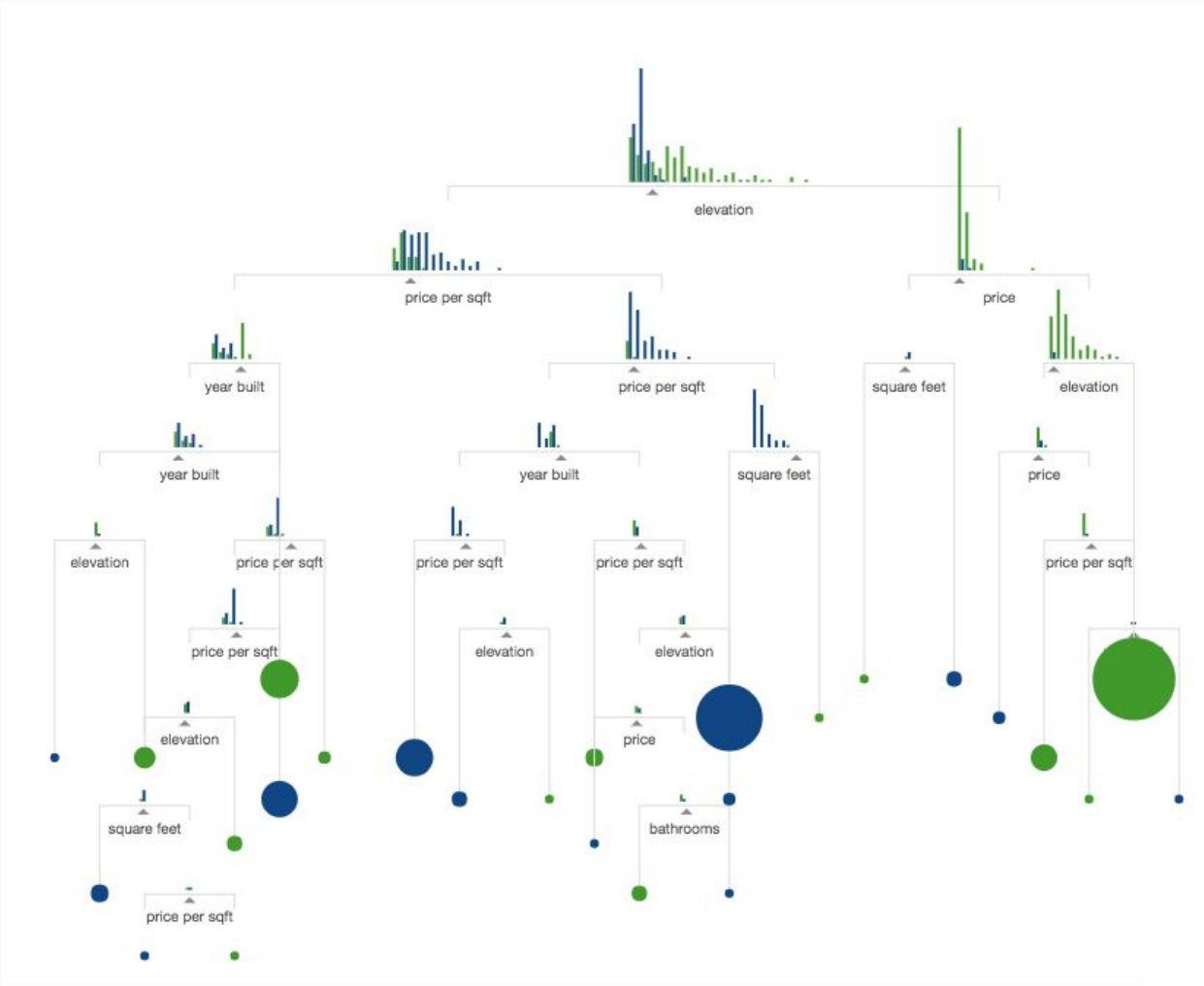


Precision и recall



Метрики: еще раз





Пример: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

В чем проблема с оценкой качества

- Нужно определить ошибку на одном примере
- Нужно определить ошибку на всех примерах для общей оценки качества модели
- **Нужно оценить ошибку на примерах, которых у нас нет**
- Нужно понимать, полученная ошибка – это много или мало

Как оценить ошибку на данных, которых нет

Никак, если не сделать дополнительных предположений

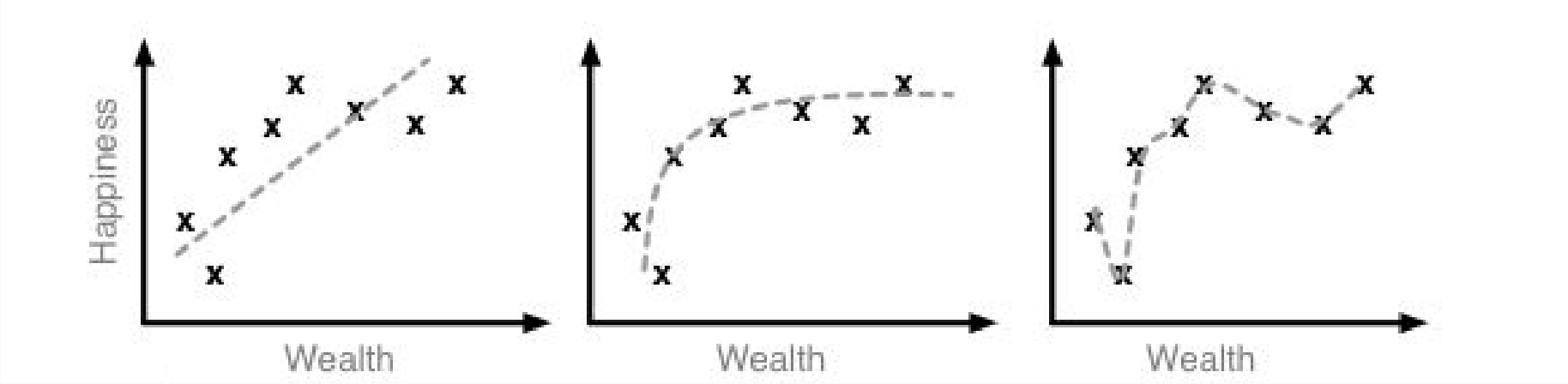
Как оценить ошибку на данных, которых нет

- Считаем, что данные, которых нет, будут похожи на наши
- Выберем случайно тестовое подмножество примеров, которое не будем использовать в обучении -- проверим именно на нем
- Так как множество выбрано случайно, то оно «похоже» на общий вид данных
- Но нам может не повезти

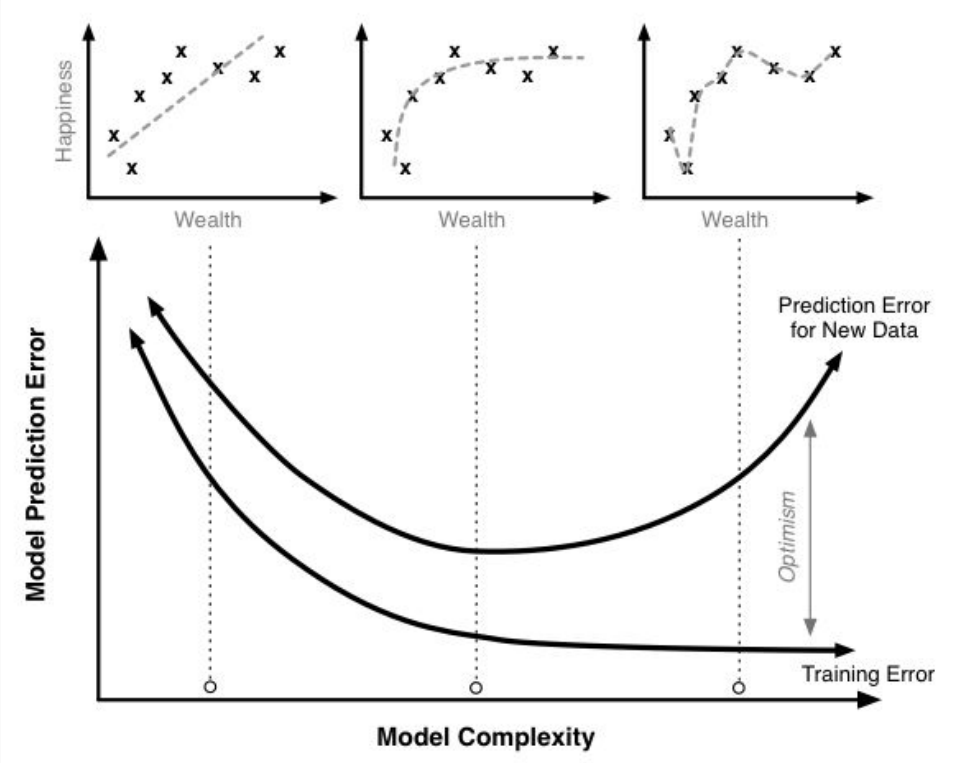
Кросс-валидация

- Исходное множество разными способами разбивается на пары обучающая выборка – контрольная выборка
- Обучаем модель на каждой обучающей выборке и проверяем на соответствующей контрольной
- Среднее значение ошибки дает более точную (несмещенную) оценку

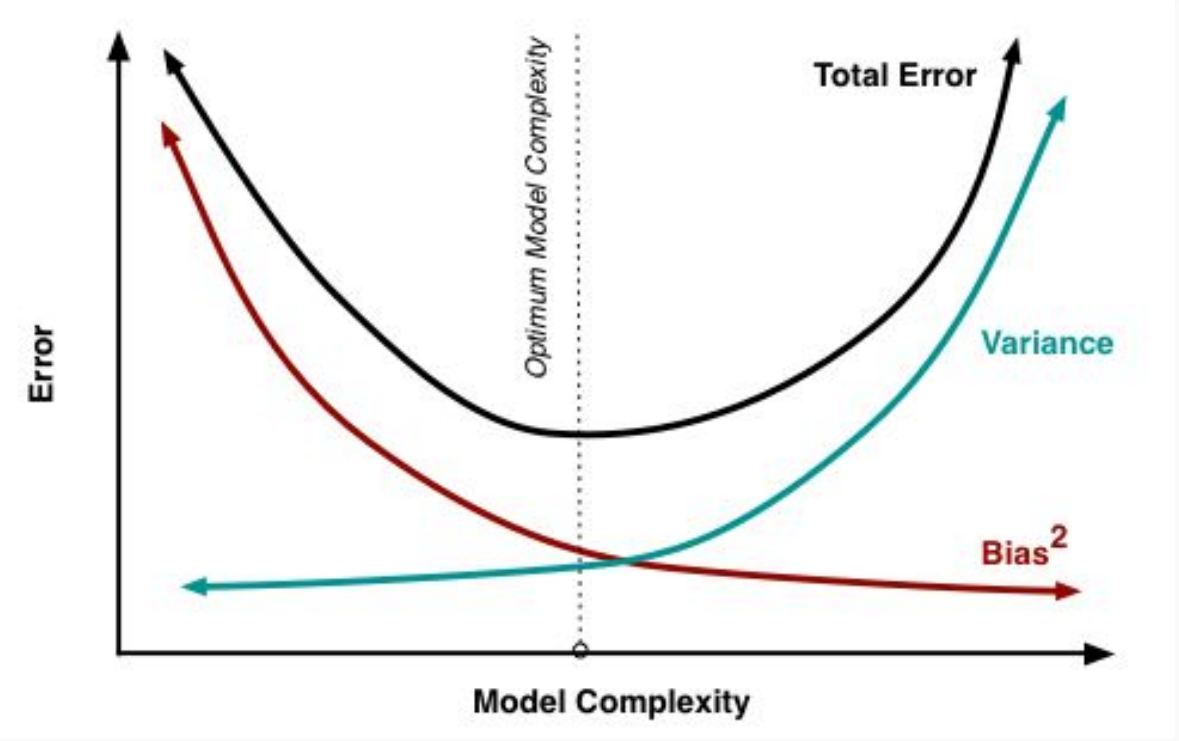
Переобучение (overfitting)



Переобучение (overfitting)



Bias–Variance tradeoff



В чем проблема с оценкой качества

- Нужно определить ошибку на одном примере
- Нужно определить ошибку на всех примерах для общей оценки качества модели
- Нужно оценить ошибку на примерах, которых у нас нет
- **Нужно понимать, полученная ошибка – это много или мало**

Задачи в машинном обучении

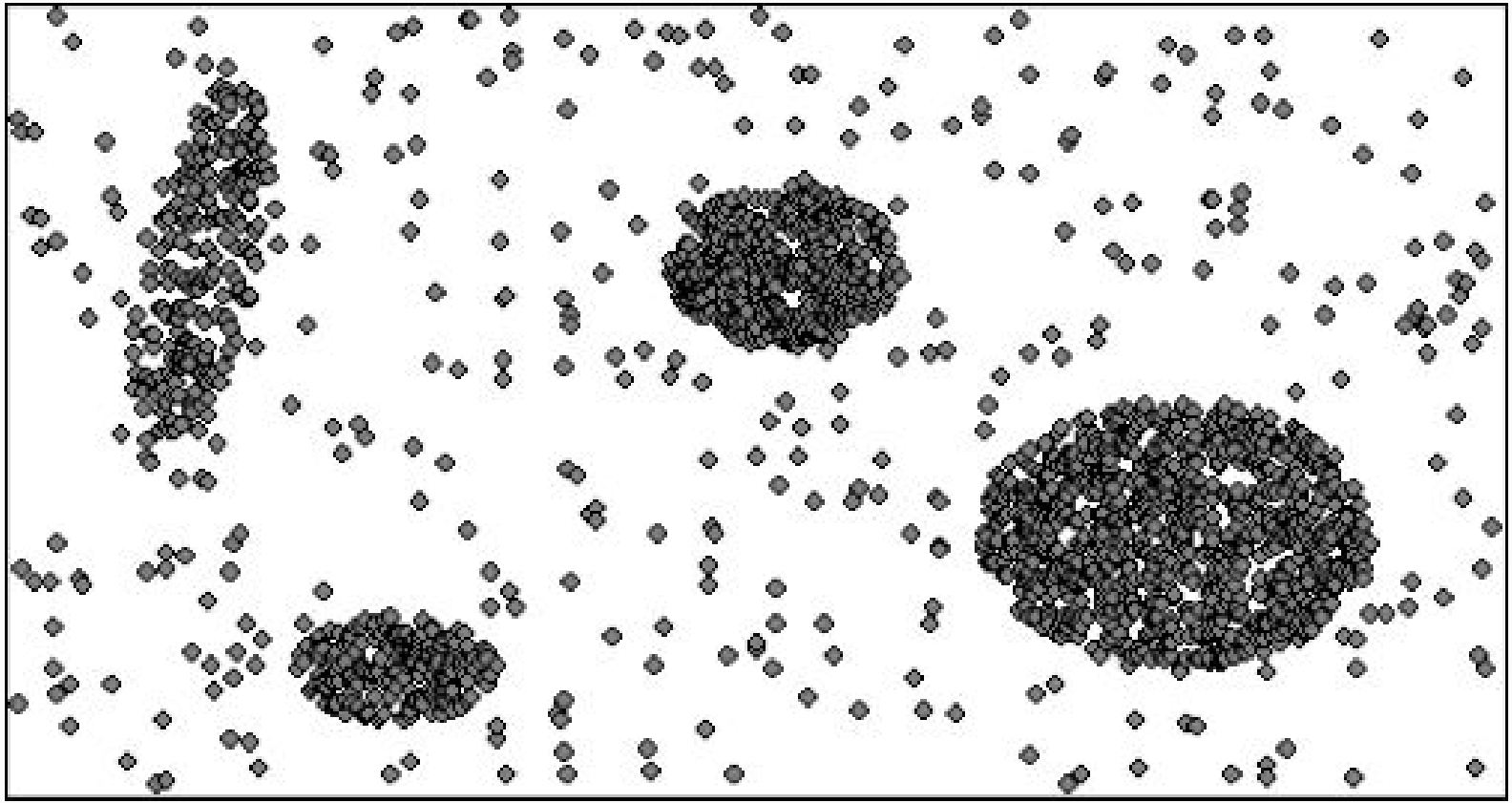
- Обучение с учителем (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами
- **Обучение без учителя (unsupervised learning) – обучение, в котором нет правильных ответов, только данные**
- Обучение с подкреплением (reinforcement learning) – обучение, в котором агент учится из собственных проб и ошибок

Обучение без учителя

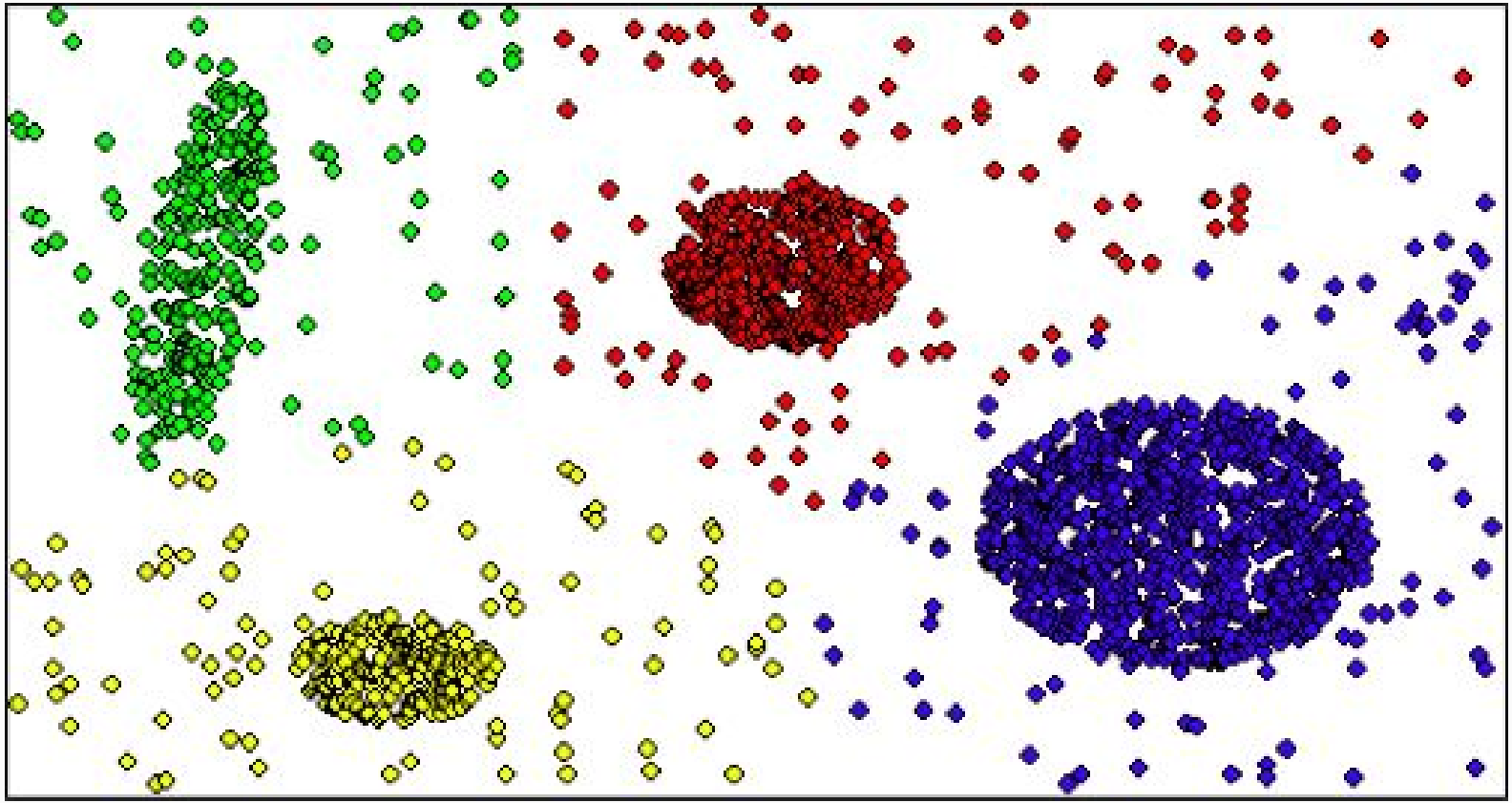
- Обучение без учителем – обучение, в котором у нас есть какие-то объекты с признаками, но нет “правильного” ответа:
  - определение “степени похожести” объектов;
  - кластеризация;
  - поиск ассоциативных правил;
  - восстановление пропусков;
  - выявление аномалий;
  - сокращение размерности;
  - визуализация.

Примеры задач

- **Кластеризация:** надо разбить данные на заранее неизвестные классы по некоторой мере похожести:
  - выделить группы или сообщества;
  - кластеризовать пользователей и персонализовать под них приложение
- **Визуализация:** хорошая визуализация позволяет увидеть некоторые закономерности
  - многомерные данные сложно отобразить
  - ищем шкалы (оси) в которых “различия” максимальны



Немного о кластеризации



Немного о кластеризации

Задачи в машинном обучении

- Обучение с учителем (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами
- Обучение без учителя (unsupervised learning) – обучение, в котором нет правильных ответов, только данные
- **Обучение с подкреплением (reinforcement learning) – обучение, в котором агент учится из собственных проб и ошибок**

Обучение с подкреплением

- Постановка задачи:
  - Есть среда, в которой действует агент. У среды есть состояние.
  - Агент может совершать действия.
  - Действия приносят некий результат (reward) (не обязательно мгновенно)
  - Надо научиться такой модели поведения, которая максимизирует результат
- Типичные задачи:
  - Управление роботом
  - Оптимизация в играх

Подводя итоги: пример

https://how-old.net/

- к какому типу задач относится пример?
- какие признаки могут использоваться? Идеи?

Подводя итоги: что будем делать

- Рассмотрим разные алгоритмы и принципы их работы
- Разберем все понятия и концепции подробнее (overfitting, bias-variance, кросс-валидация, тестовая и обучающая выборки, выбор признаков, сокращение размерности и т.д.)
- Рассмотрим типичные ошибки и способы их избежать
- Научимся интерпретировать результаты и выбирать лучшую модель
- Узнаем, как это все применить в R

Немного организационного: напоминание

проверьте доступ к

- RStudio
- Stepik.org
- почте ВШЭ (рассылка все так же через нее)