МАШИННОЕ ОБУЧЕНИЕ

ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

МИХАИЛ ЛИПКОВИЧ

ПЛАН

- Задача понижения размерности (вспомним)
- Задача кластеризации

ОБУЧЕНИЕ С УЧИТЕЛЕМ

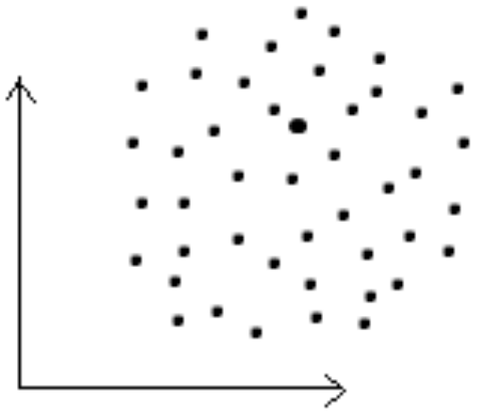
- Раньше рассматривали задачи, где у нас том или ином смысле были правильные ответы и задача состояла в том, чтобы приблизиться к ним
- В реальности получение таких ответов может быть связано с высокой стоимостью (ручная разметка, обзвон пользователей, проведение исследования)
- Но зато могут получаться модели хорошего качества

ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

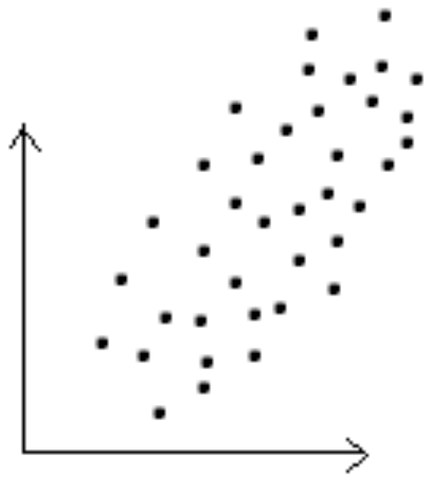
- Но можно что-то делать и когда ответов нет
- Типичные методы пытаются из самих данных выделить некоторую структуру:
  - Кластеризация
  - Понижение размерности
  - Поиск аномалий

ЧЕМ ОТЛИЧАЮТСЯ ТРИ КАРТИНКИ?

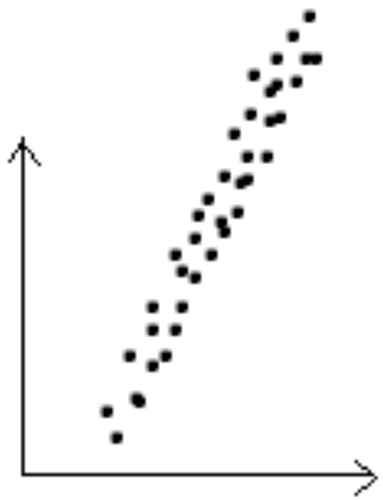
ЧЕМ ОТЛИЧАЮТСЯ ТРИ КАРТИНКИ?



Данные независимы по компонентам

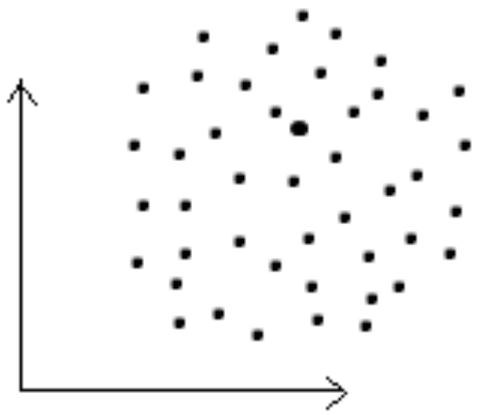


Есть линейная зависимость



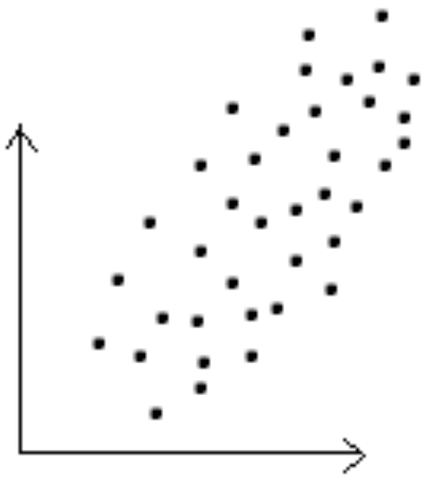
Очень сильная линейная зависимость

И ЧТО С ТОГО?

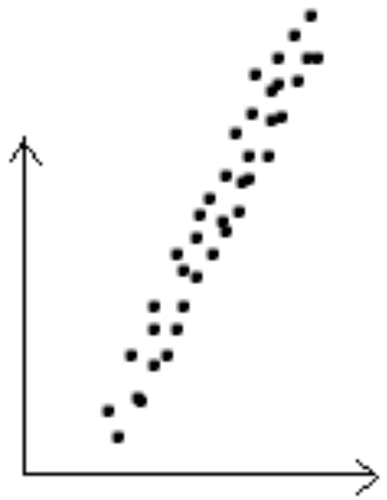


Данные независимы
по компонентам

Здесь ничего не сделать



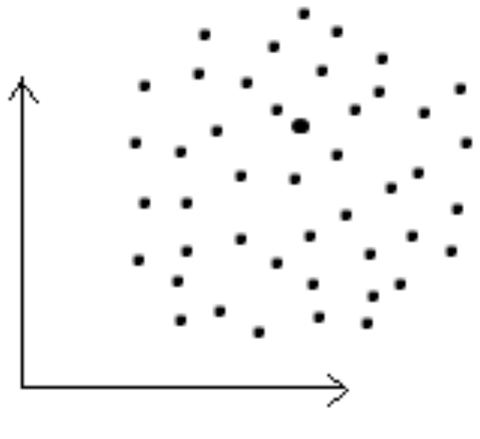
Есть линейная
зависимость



Очень сильная
линейная зависимость

А здесь зачем нам
две координаты?

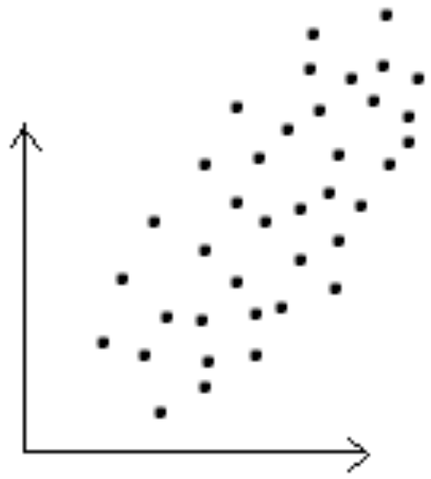
СМОТРИМ НА КОВАРИАЦИОННЫЕ МАТРИЦЫ



Данные независимы по компонентам

$$\begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}$$

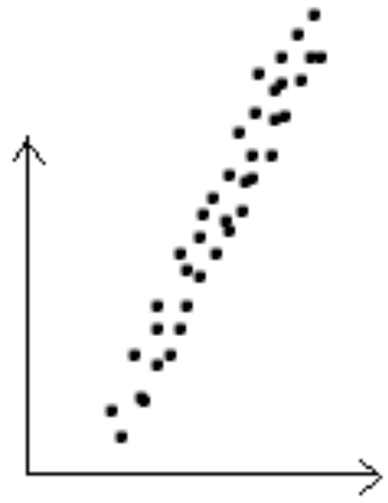
Почти нули вне диагонали



Есть линейная зависимость

....

Нечто среднее



Очень сильная линейная зависимость

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

Большие элементы вне диагонали

НАБЛЮДЕНИЯ

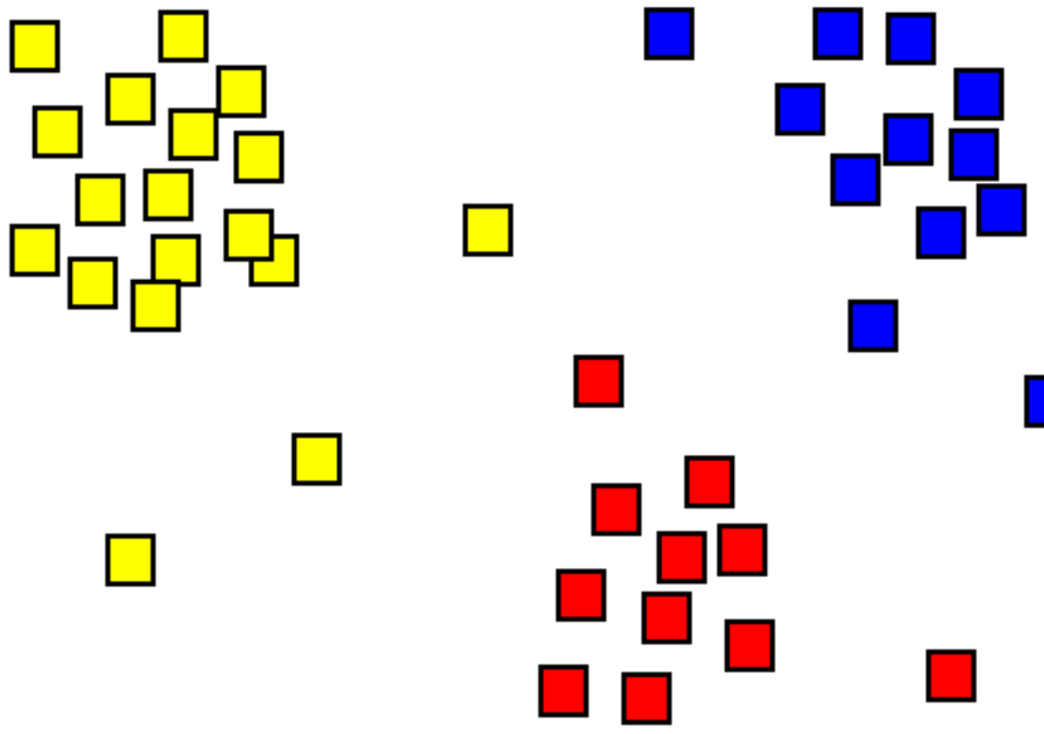
- В данных могут быть избыточные размерности (координаты)
- Доп. размерности могут быть обусловлены шумом / сильно скорелированными признаками / просто мусором. Это все только мешает при обучении или анализе данных
- Мы хотели бы привести данные к виду, когда ковариационная матрица диагональная. А далее, мы бы убрали компоненты с маленькой дисперсией

АЛГОРИТМ ТАКОЙ:

- Вычитаем среднее из каждого вектора фичей X
- Применяя SVD к X находим собственные векторы XX^T
- Умножая эти собственные векторы на наши исходные фичи, получаем фичи в новом пространстве
- Элементы матрицы Sigma показывают значимость координат
- Отсортировав собственные векторы по значимости Sigma, можем отрезать сколько угодно последних компонент

ЧТО ТАКОЕ КЛАСТЕРИЗАЦИЯ?

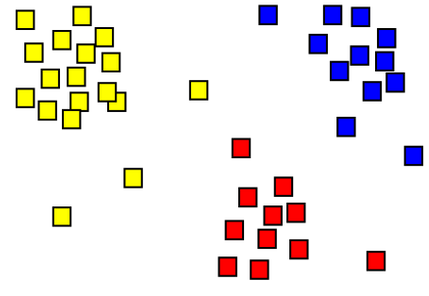
- По данным получить что-то такое:



https://en.wikipedia.org/wiki/Cluster_analysis

ЧТО ТАКОЕ КЛАСТЕРИЗАЦИЯ?

- По данным получить что-то такое:



- Что именно из себя должно представлять это разбиение зависит от наших критериев близости/плотности/разрозненности

ВИДЫ КЛАСТЕРИЗАЦИЙ

- Иерархическая кластеризация
- На основе центроидов
- На основе распределений
- На основе плотности
- Графовая кластеризация
- ....

САМЫЙ ПОПУЛЯРНЫЙ АЛГОРИТМ:

1. Задаем k - число кластеров
2. Случайно задаем k центров кластеров (центроидов)
3. По каждой точке определяем к какому кластеру она относится по близости к центроиду
4. Определяем внутри каждого кластера новый центроид как центр масс
5. Повторяем шаги 3-4 пока не сойдемся

ЗАМЕЧАНИЯ

- Откуда взять k?

ЗАМЕЧАНИЯ

- Откуда взять k?
  - Попробовать для разных k и посмотреть как себя ведут метрики.
  - В качестве метрики можно взять silhouette_score: среднее расстояние внутри кластера должно быть маленьким, среднее расстояние между кластерами должно быть большим

НЕДОСТАТКИ

- Необходимость выбора k (но иногда наоборот хорошо)
- Зависимость от начального выбора точек (можно делать несколько итераций)
- Выделяет только шарообразные кластеры