

Обучение без учителя

Владимир Вежневек, Антон Конушин
Александр Вежневек

Компьютерное зрение
МГУ ВМК, Осень 2006

Пример



- После проведения социологического исследования, как выявить группы людей сходных мнений?
- Есть большая база данных изображений, требуется разделить их на группы
- Сегментация

Обучение без учителя



- Пусть, имеется набор наблюдений:

$$X^l = \{x_1, \dots, x_l\}, X \in R^d$$

- Требуется некоторым образом сделать суждения о наблюдаемых данных
- Трудно придумать более общую постановку, неправда ли?

Обучение без учителя



- Кластеризация
 - Разбиение наблюдений на некоторые группы, с максимально близкими наблюдениями внутри групп и максимально далекими между
- Понижение размерности
 - Понижение размерности наблюдений с сохранением описательной силы
- Анализ плотности распределения
 - Получить аппроксимацию плотности распределения вероятности наблюдений или поиск их особых точек

Обучение без учителя



Кластеризация

- К-средних
- Смесь нормальных распределений
- ...



Понижение размерности

- Метод главных компонент
- SOM
- ...

■ Анализ плотности распределения

- Аппроксимация плотности распределения через обучение с учителем
- Сдвиг среднего для поиска экстремумов плотности распределения
- ...

Отличие от классификации



- Множество ответов неизвестно
- Нет четкой меры качества решений
- Задачи поставлены крайне нечетко

Кластеризация

Постановка задачи (1)



- Пусть, имеется набор наблюдений:

$$X^l = \{x_1, \dots, x_l\}, X \in R^d$$

- Требуется разбить X^l на некоторые непересекающиеся подмножества (группы, кластеры), таким образом, что объекты внутри одной группы соотносились сильнее чем объекты из разных групп

Кластеризация

Постановка задачи (2)



- Пусть, так же, имеется некоторая мера $D: X \times X \rightarrow R$, характеризующая «схожесть» между объектами
- Тогда, требуется найти некоторое разбиение:

$$\{C^i\}_1^k : C^i \subset X^l, \bigcap_{i=1}^k C^i = X^l; C^i \cap C^j = \emptyset \text{ для } i \neq j$$

- Такое, что минимизируется $D(x_i, x_j); x_i, x_j \in C^p$
- И максимизируется $D(x_i, x_j); x_j \in C^p, x_i \in C^r; r \neq p$

Кластеризация

Модель кластеров



- Под моделью кластеров будем понимать некоторое параметрическое семейство отображений из исходного пространства в множество индексов кластеров

$$C = \{c(x, \gamma) \mid c : X \times \Gamma \rightarrow \{1, \dots, k\}\}$$

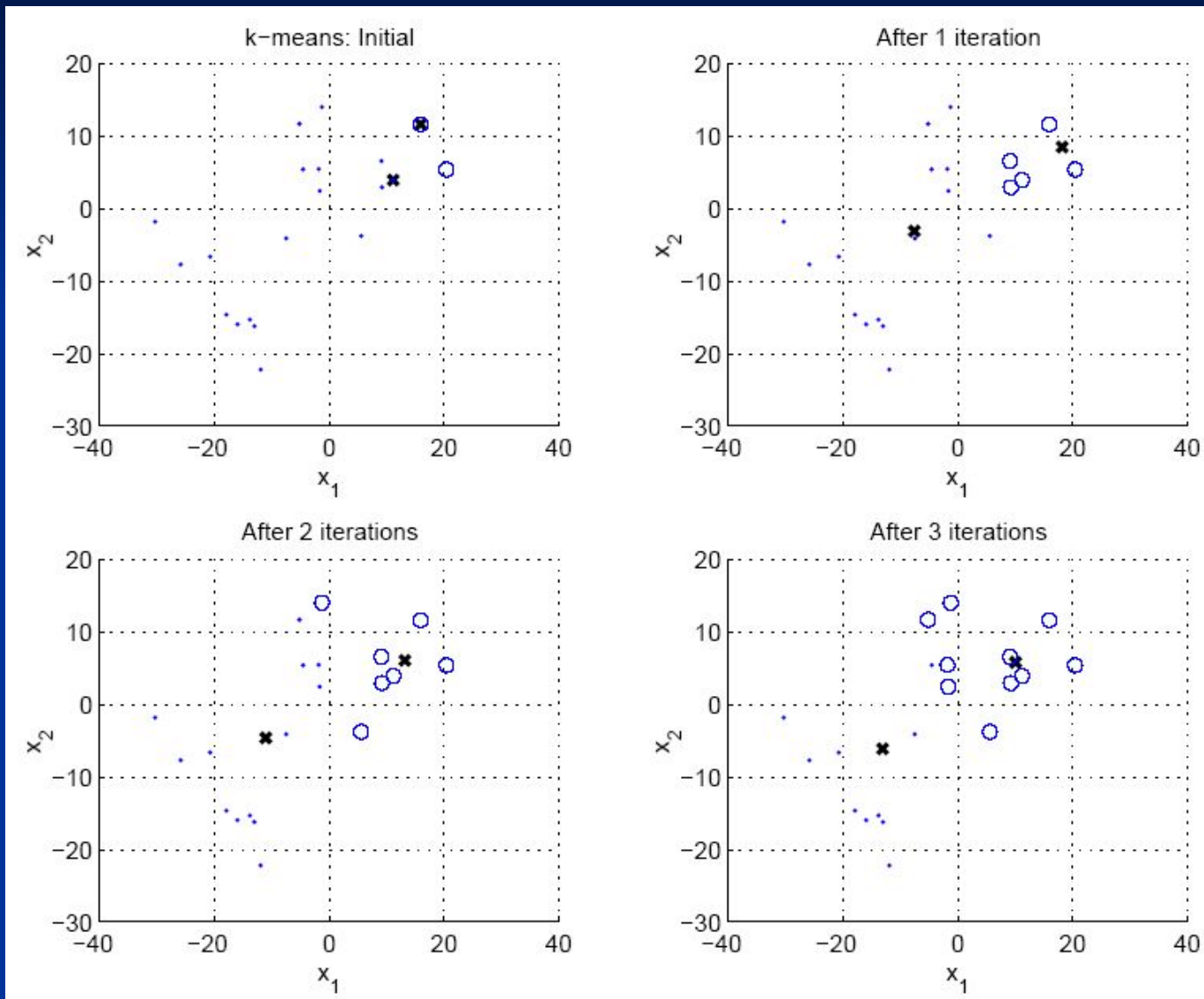
- Множество параметров, пространством поиска
- Нахождения параметров - кластеризацией

Алгоритм K-средних



- Кто хочет рассказать как он работает?
 1. Случайным образом выбрать k средних m_j , $j=1, \dots, k$;
 2. Для каждого x_i , $i=1, \dots, p$ подсчитать расстояние до каждого из m_j , $j=1, \dots, k$,
 3. Отнести (приписать) x_i к кластеру j' , расстояние до $m_{j'}$ минимально;
 4. Пересчитать средние m_j , $j=1, \dots, k$ по всем кластерам;
 5. Повторять шаги 2, 3 пока кластеры не перестанут изменяться;

Иллюстрация



Алгоритм K-средних

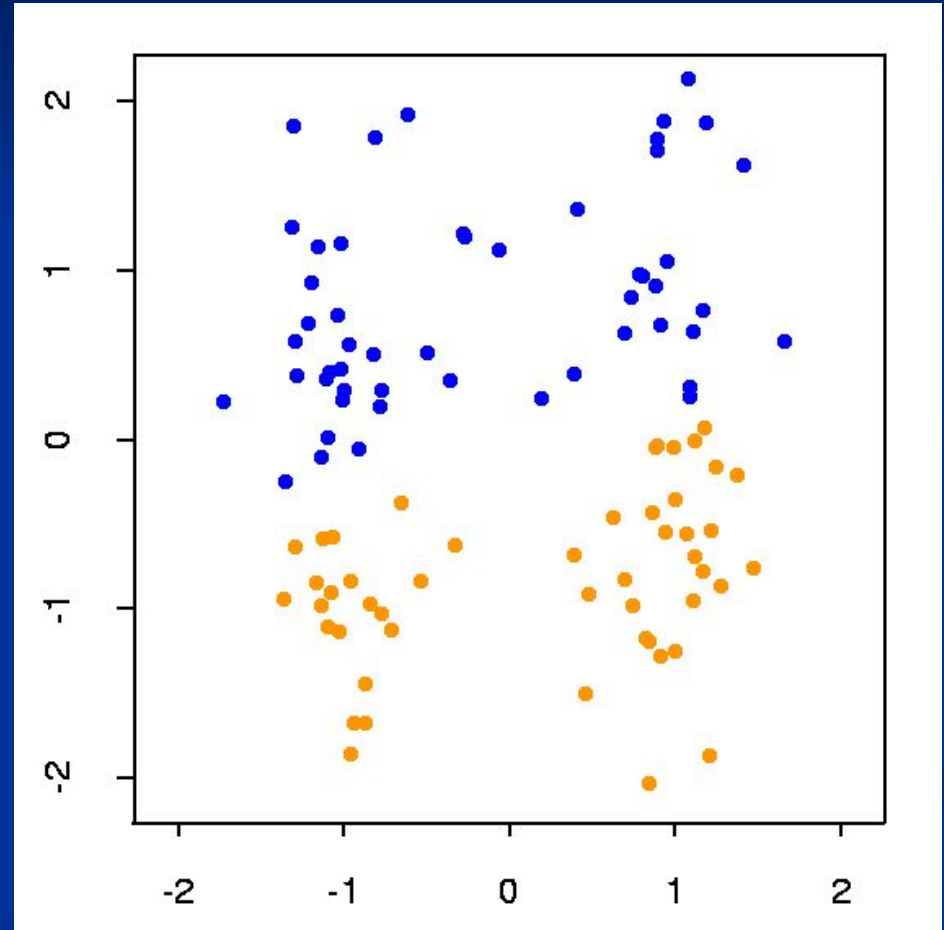


- Мера схожести $D: X \times X \rightarrow R$
 - Евклидово расстояние в пространстве X
- Модель кластеров
 - Пространство поиска - центры масс

Алгоритм К-средних



- Однопараметрический
 - Требуется знание только о количестве кластеров
- Рандомизирован
 - Зависит от начального приближения
- Не учитывает строения самих кластеров



EM алгоритм

Общая идеология



- Пусть есть вектор неизвестных величин

$$\eta = (\eta_1, \dots, \eta_p)$$

- и параметрическая модель с так же неизвестным параметром(ами) λ

- Пусть возможно рассчитать правдоподобие $P(\eta | \lambda)$

- Наша задача подобрать такие λ и η , чтобы правдоподобие было максимальным

EM алгоритм

Общая идеология



- Возьмем некоторые начальные приближения λ^0
- Итеративно $t = 1 \dots$ делаем два шага:
 - E-хрест: согласно текущему значению λ^{t-1} высчитываем наиболее вероятные значения

$$\eta^t = \arg \max_{\eta} (P(\eta | \lambda^{t-1}))$$

- Maximize: согласно текущим значениям η^t высчитываем новое значение λ^t максимизирующее функцию правдоподобия

$$\lambda^t = \arg \max_{\lambda} (P(\eta^t | \lambda))$$

- Остановимся когда правдоподобие стабилизируется

Кластеризация смесью нормальных распределений



- Будем считать, что наблюдения сгенерированы смесью нормальных распределений, то есть:

$$P(x | C^i) = \frac{P(x | C^i) P(C^i)}{\sum_{i=1}^k P(x | C^i) P(C^i)} \exp\left(-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right)$$

$(2\pi)^{-n/2} |\Sigma_i|^{-1/2} N(\mu_i, \Sigma_i)$

- Пусть k известно заранее, будем осуществлять кластеризацию EM алгоритмом

$\eta = (\eta_1, \dots, \eta_l)$ - Индексы кластеров наблюдений

$$\lambda = \{\mu_i, \Sigma_i\}_{i=1}^k$$

Кластеризация смесью нормальных распределений



- Возьмем некоторые (случайные) начальные приближения $\lambda = \{\mu_i, \Sigma_i\}_1^k$
- Итеративно для $t = 1 \dots$:
 - E: согласно текущему значению λ высчитываем наиболее вероятные значения индексов $\eta = (\eta_1, \dots, \eta_l)$ кластеров для наблюдений из X^l

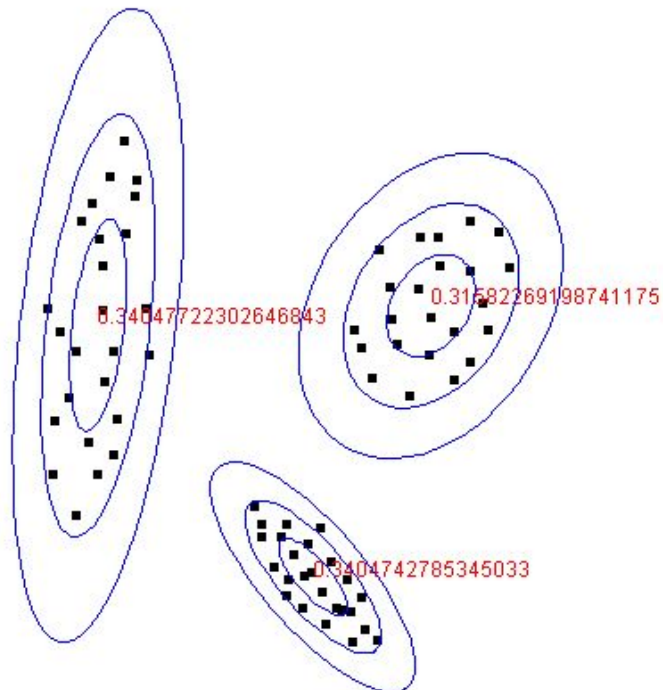
$$\eta_j^t = \arg \max_i (P(x | C^i))$$

- M: согласно текущим значениям индексов пересчитаем параметры распределений (методом максимального правдоподобия)

Иллюстрация



Mean Likelihood = -10.212002301916018



Кластеризация смесью нормальных распределений



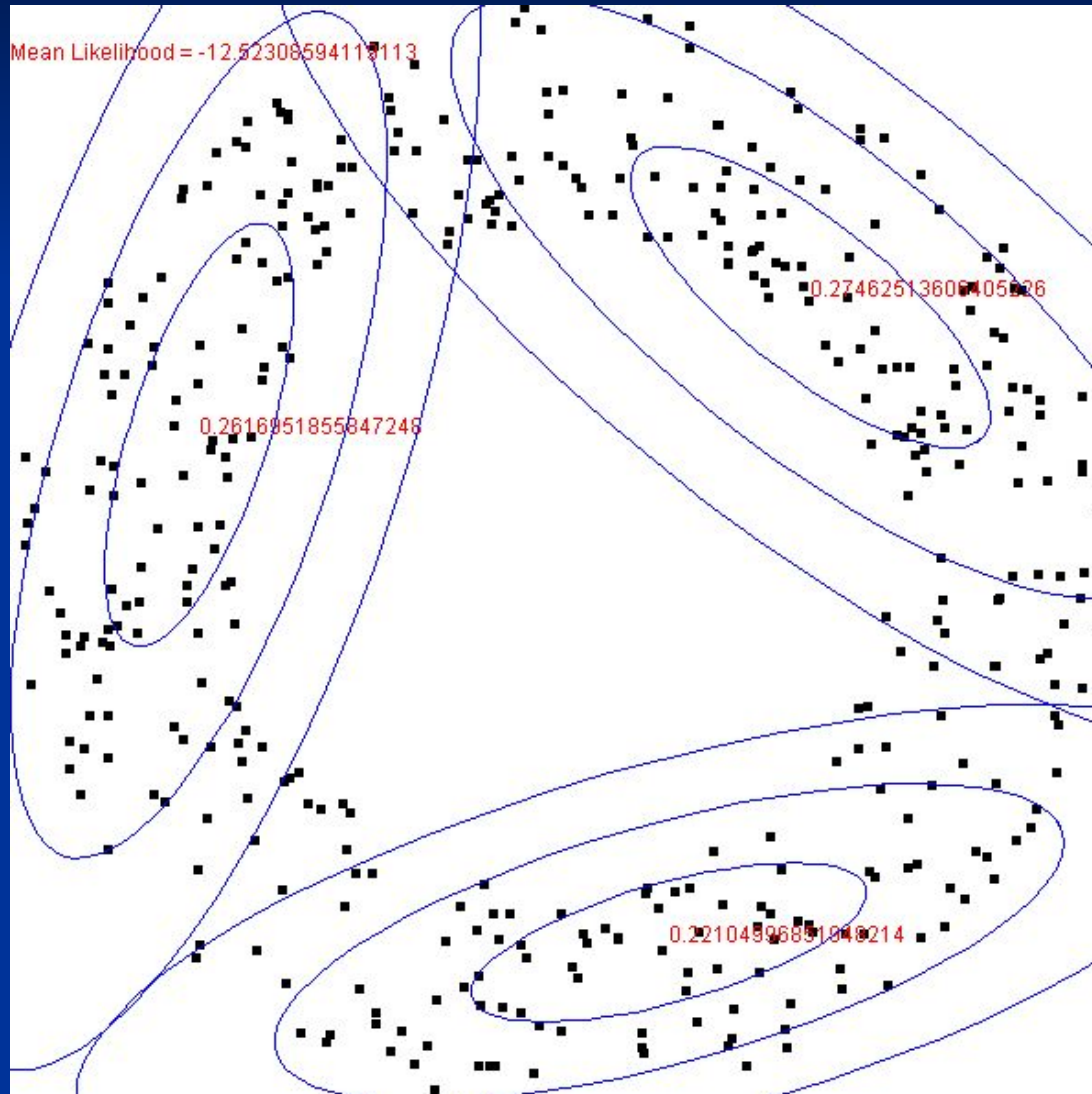
■ Плюсы

- Более полная модель кластеров (больше итоговой информации)
- Более качественная аппроксимация
- Эффективная оценка качества кластеризации (правдоподобие)

■ Минусы

- Все равно некоторая ограниченная модель со строгой «геометрией»
- Чувствительность к размерности и нормализации данных

Иллюстрация



Понижение размерности наблюдаемых данных



- Зачастую, наблюдаемые данные могут обладать высокой размерностью, но в действительности быть функцией всего нескольких скрытых (латентных) переменных
- Задачей понижения размерности является некоторое преобразование исходного пространства в пространство более низкой размерности без существенной потери информативности данных

Метод главных компонент



- Пусть имеется набор наблюдений $X^l = \{x_1, \dots, x_l\}$, $X \in R^d$
- Будем строить новый базис в пространстве R^d , таким образом что:
 - Центр координат совпадает с мат. ожиданием наблюдений (выборочным средним)
 - Первый вектор направлен таким образом, что дисперсия вдоль него была максимальной
 - Каждый последующий вектор ортогонален предыдущим и направлен по направлению максимальной дисперсии

Метод главных компонент

Расчет базиса



- Сдвинем все данные таким образом, чтобы их выборочное среднее равнялось нулю
- Рассчитаем ковариационную матрицу:

$$\Sigma = \begin{pmatrix} \text{cov}(x^1, x^1) & \text{cov}(x^1, x^2) & \dots & \text{cov}(x^1, x^d) \\ \text{cov}(x^2, x^1) & \text{cov}(x^2, x^2) & \dots & \text{cov}(x^2, x^d) \\ \dots & \dots & \dots & \dots \\ \text{cov}(x^d, x^1) & \text{cov}(x^d, x^2) & \dots & \text{cov}(x^d, x^d) \end{pmatrix}$$

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{(n-1)}$$

Выборочное среднее

Ковариация двух сл. величин

Метод главных компонент

Расчет базиса



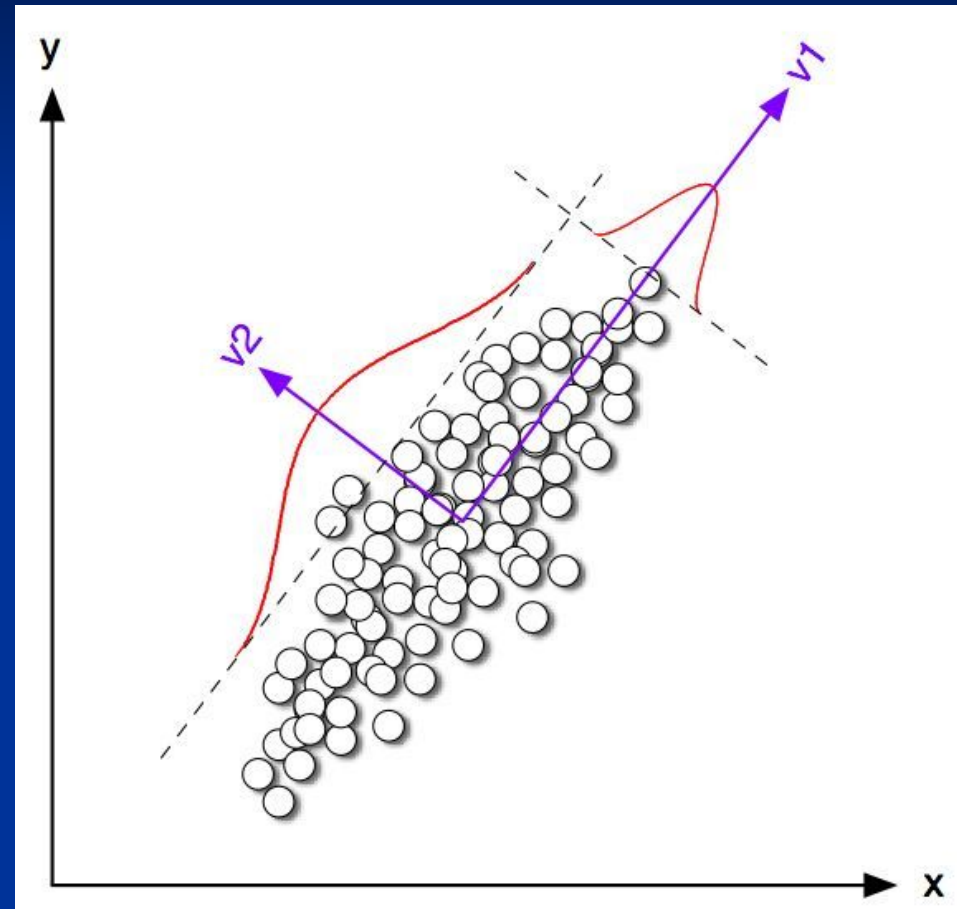
- Векторами нового базиса будут являться собственные вектора ковариационной матрицы
- Собственные числа – значениями дисперсии наблюдений вдоль них

Иллюстрация



- Убирая базисные вектора с малыми значениями мы можем сократить размерность без существенной потери информации

*Вопрос: Почему это так?



Случай нормального распределения



- Расстояние от центра распределения в новой системе координат равно:

$$\sqrt{(x - \bar{x})^T \Sigma_y^{-1} (x - \bar{x})}$$

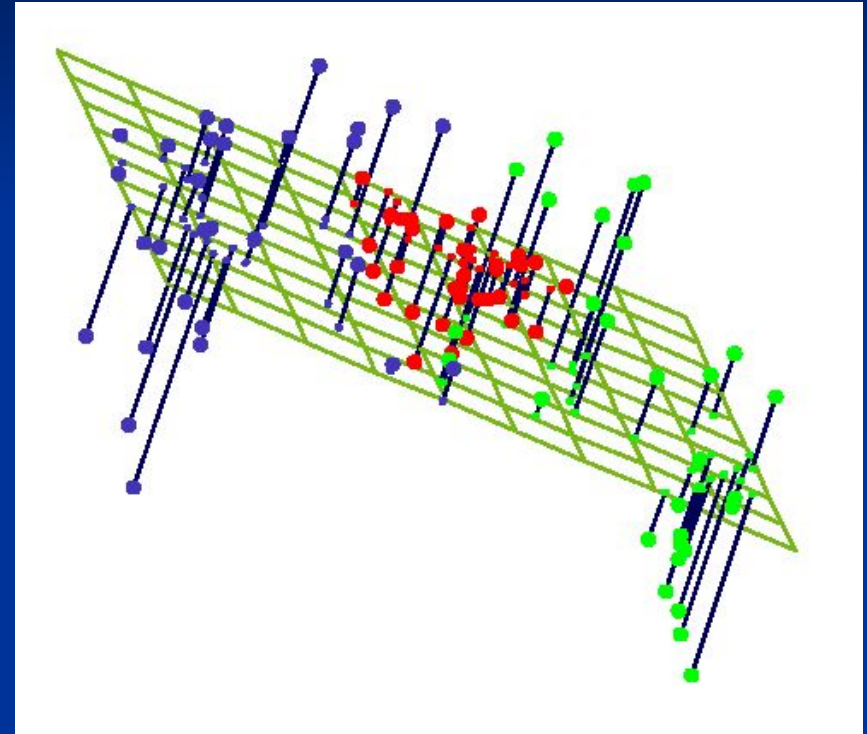
- Так называемое расстояние Махалонобиса
- Пропорционально правдоподобию наблюдения

Метод главных компонент

Связь с линейной аппроксимацией



- Если рассмотреть систему проекций данных на первые $1 \leq n \leq d$ главных компонент, то мы получим систему наилучших линейных приближений данных (в смысле среднеквадратичного отклонения)



Метод главных компонент



- Следует применять:
 - Данные распределены нормально и требуется привести их к более удобной форме
 - Или предполагается, что данные содержатся в линейном многообразии исходного пространства и требуется выделить лишь его и сократить размерность
- НЕ следует применять:
 - Распределение данных произвольно и далеко от нормального
 - Данные нелинейные

Самоорганизующиеся карты SOM (Карты Кохенена)

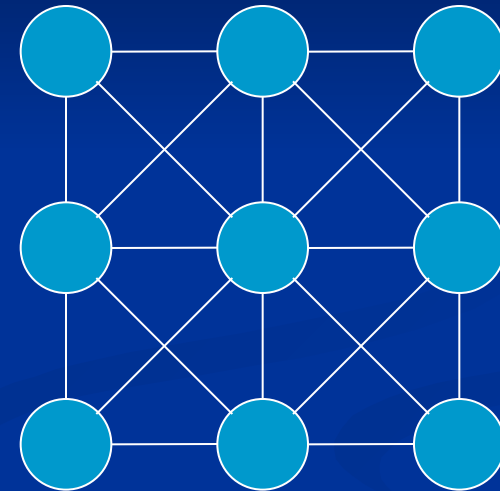


- Основная идея
 - вписать в данные сетку низкой размерности, и анализировать ее, вместо самих данных

SOM (Карты Кохенена) Модель сетки



- Матрица узлов $M^{KL} = (m_{kl})$
- Соседство - 4 или 8 СВЯЗНОСТЬ
- Каждому узлу соответствует точка в исходном пространстве



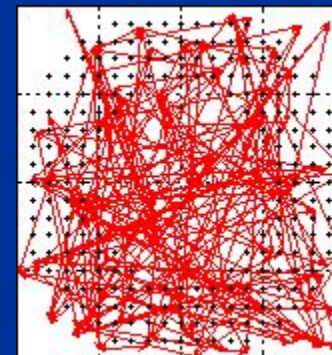
$$m_{kl} \in R^d$$

SOM (Карты Кохенена)

Алгоритм построения

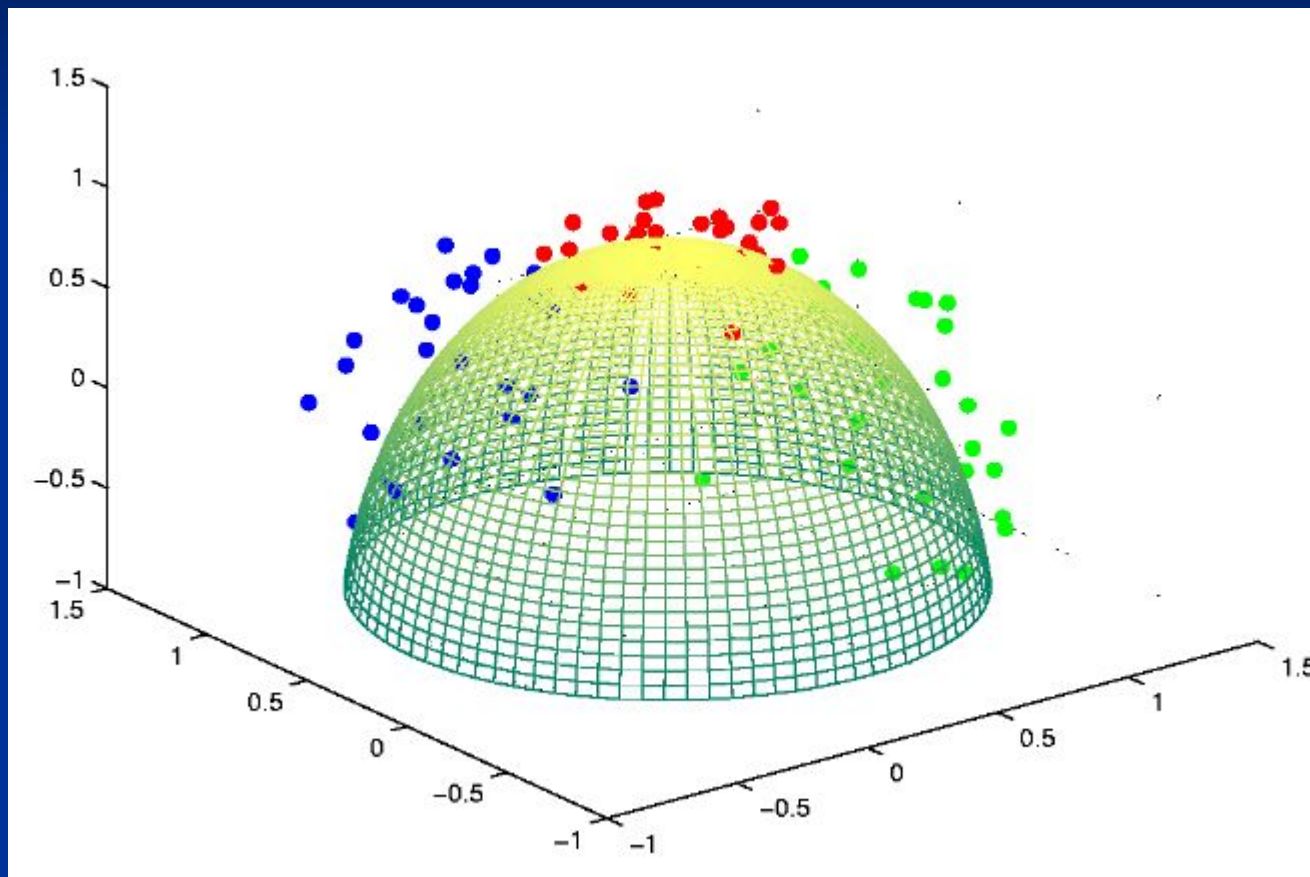


- Проинициализируем (m_{ij}) случайными значениями
- Далее, в случайном порядке будем предъявлять наблюдения и для каждого:
 - Вычисляем ближайший узел
 - Выберем множество соседей узла, такое что расстояние на сетке между ними меньше r
 - Для некоторого множества соседей узла, включая сам узел, изменяем их положения согласно: $m_{kl} \leftarrow m_{kl} + \alpha(x_i - m_{kl})$
 - Повторяем процедуру уменьшая r и α пока сеть не стабилизируется



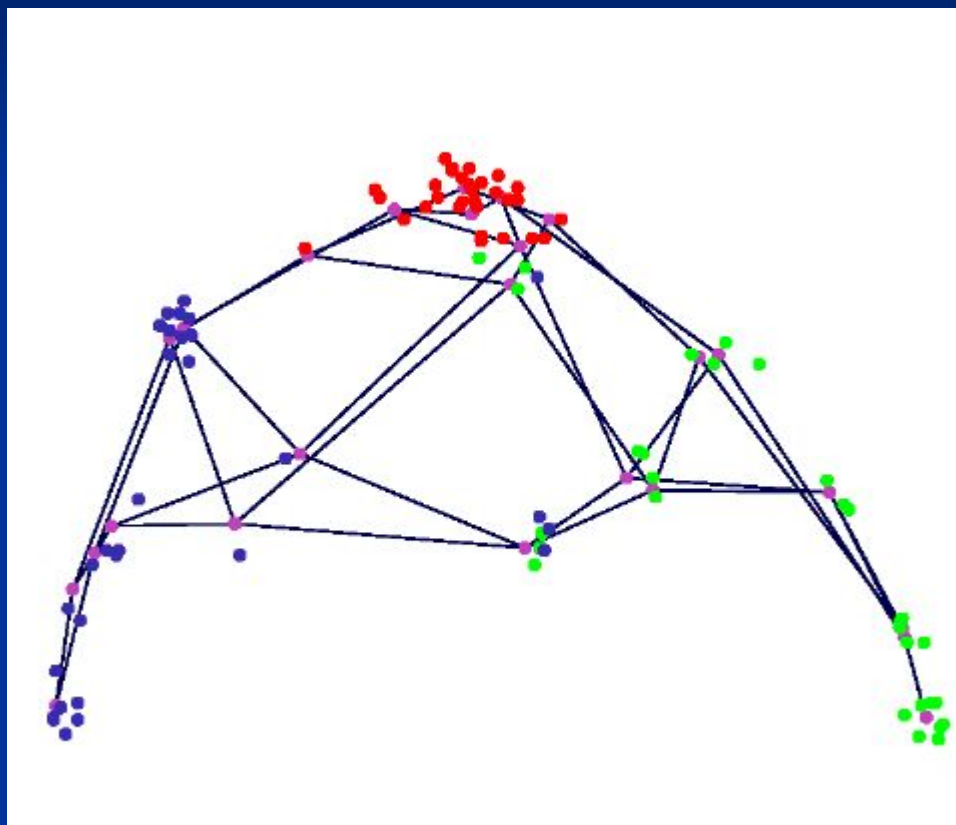
SOM (Карты Кохенена)

Иллюстрация: исходные данные



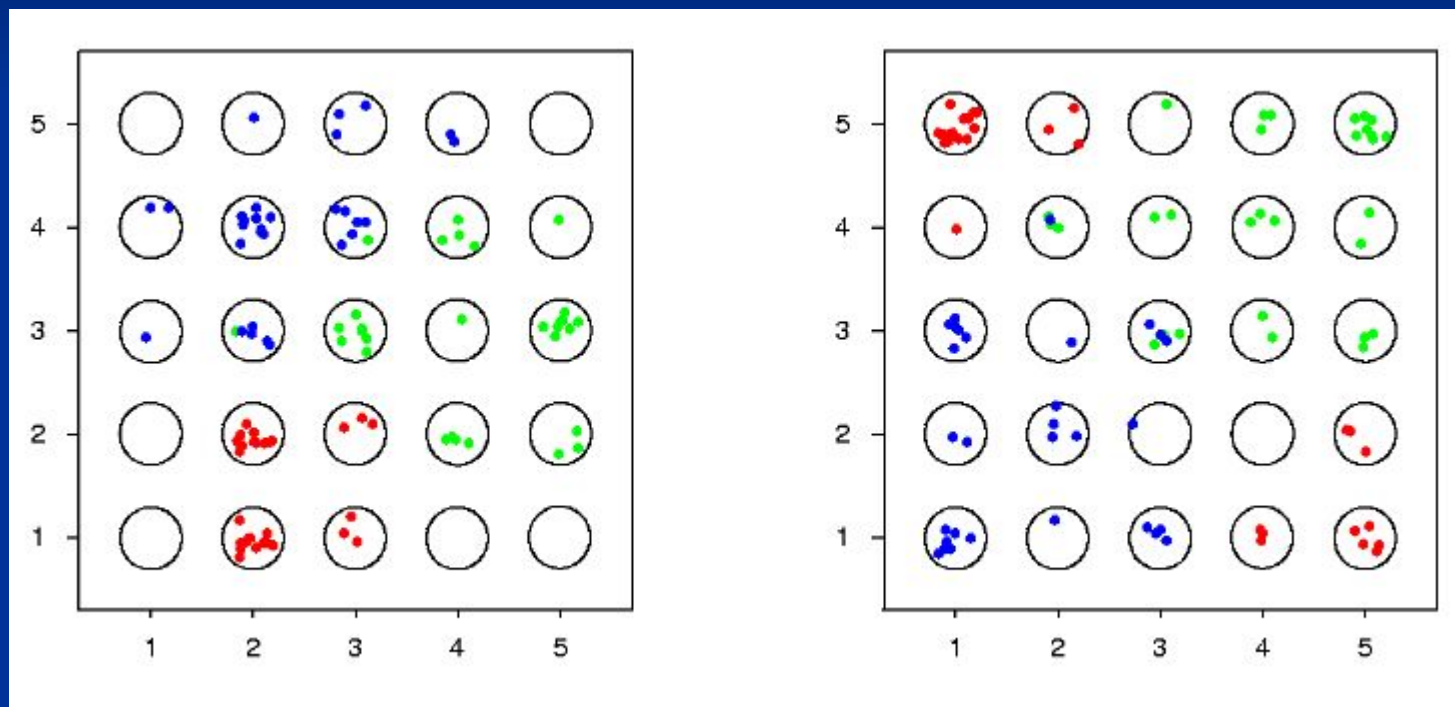
SOM (Карты Кохенена)

Иллюстрация: сетка



SOM (Карты Кохенена)

Иллюстрация: проекции на матрицу



SOM (Карты Кохенена)

Практическое использование



- Данные представляют некоторую поверхность, требуется сократить размерность
- Хорошо подходят для последующей кластеризации
- Могут работать «online»
- В случае слишком сложных данных не информативны

Задание №2



- Каждому будут выданы
 - Данные (3 набора)
 - Алгоритмы классификации, реализованные в MatLab
- Требуется
 - Для каждого из наборов натренировать наилучший возможный классификатор (из выданных вам)
 - Написать отчет (по форме лежащий на сайте) описывающий каким образом вы выбирали классификатор и настроили параметры
- Оценка складывается из:
 - Аккуратности и полноты отчета
 - Результатов работы Вами присланного классификатора на *наших* данных (часть выборки мы дали Вам, часть оставили себе)

Содержание отчета



- Применявшиеся методы
 - Список

- Алгоритм, по которому оценивались алгоритмы и выбирались параметры

- Результаты в виде
 - Графиков
 - Таблиц

- Выводы (кратко!)
 - Какой классификатор был выбран в итоге
 - Почему именно этот классификатор

Оформление решения



- Классификаторы
 - <Номер данных>_<Название метода>
 - Например 23_SVM
- Отчет
 - Файл в формате MS Word