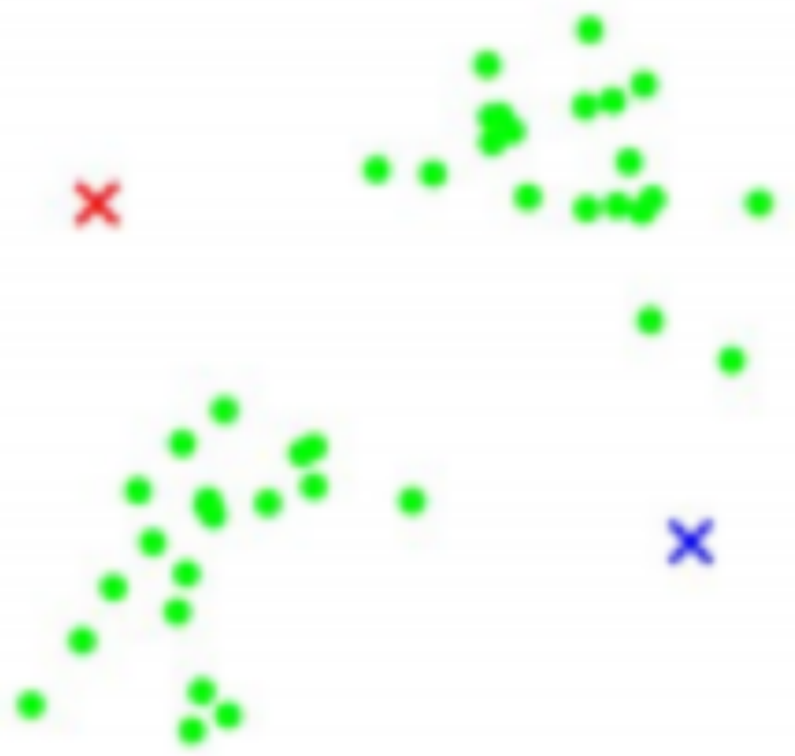
Метод К средних (K Means)

План

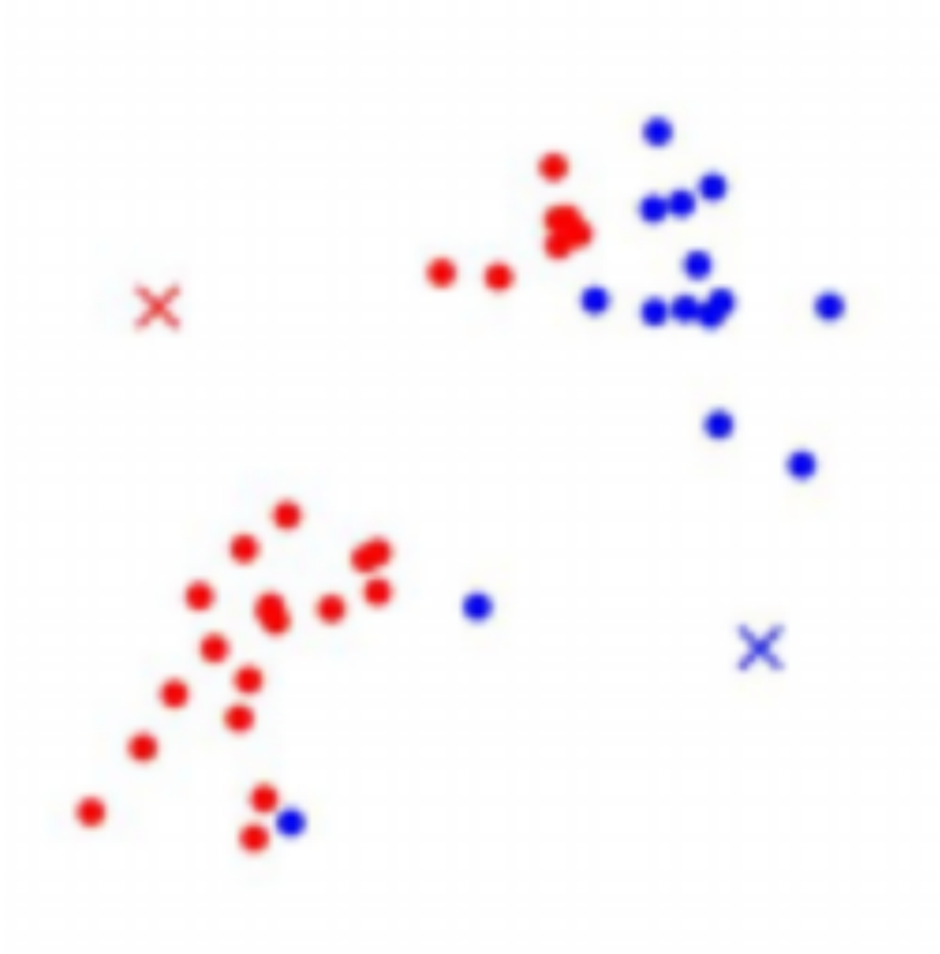
1. Как работает K Means
2. Вариации K Means
3. Что делать, когда данных много: Mini Batch K-Means
4. Что делать, когда много признаков
5. Выбор начальных приближений: Kmeans++
6. Пример: уменьшение количества цветов в изображении
7. Работа K means с разными формами кластеров
8. Пример: мешок визуальных слов (bag of visual words)
9. Что оптимизирует K means

Как работает K Means

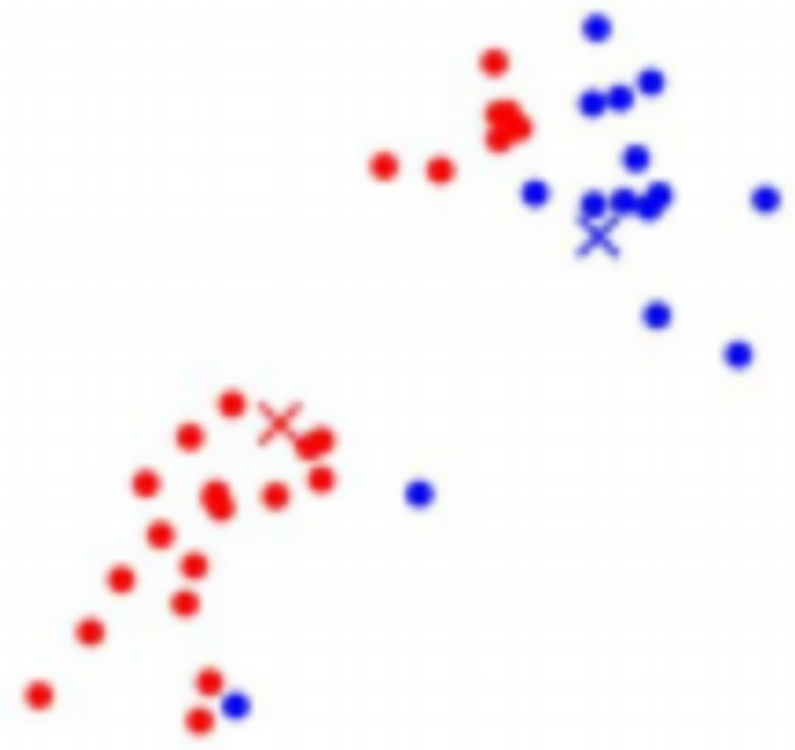
Как работает K Means



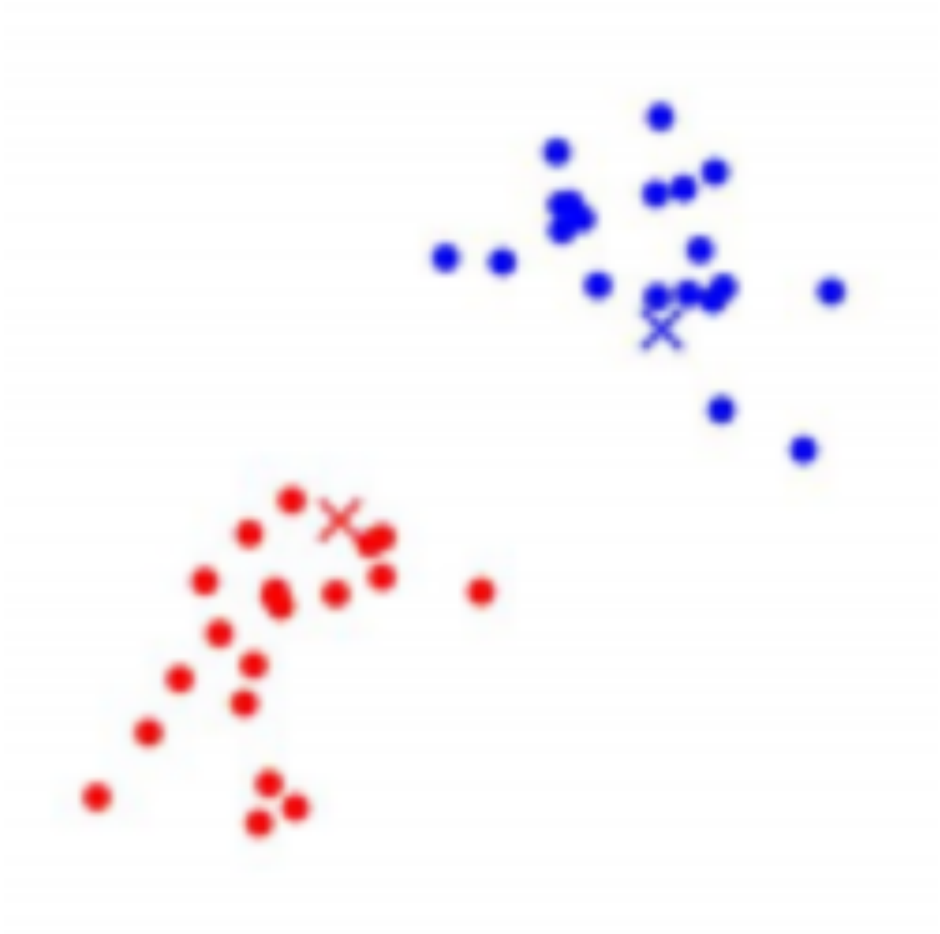
Как работает K Means



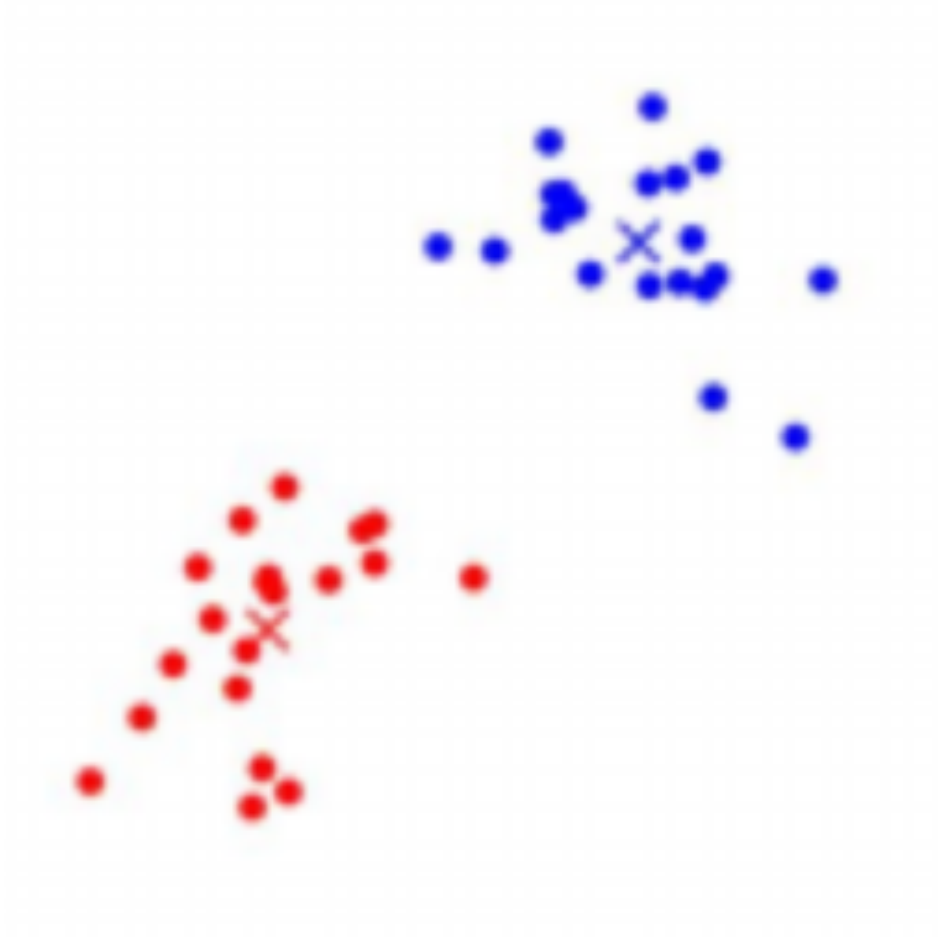
Как работает K Means



Как работает K Means



Как работает K Means



Вариации K Means

- В версии Болла Холла: уже расказанный метод
- В версии Мак Кина: каждый раз, когда объект переходит из одного кластера в другой – центры кластеров пересчитываются

Mini-Batch K Means

- Если данных много, относить объекты к кластерам и вычислять центры – достаточно долго
- Выход – на каждом шаге K Means работать со случайной подвыборкой из всех объектов
- В среднем все должно сходиться к тому же результату

Понижение размерности пространства

- Каждое вычисление расстояния обычно требует O(d) элементарных операций, где d – размерность пространства признаков
- Если признаков очень много, K Means начинает работать долго
- Решение – уменьшить число признаков
- Варианты: отбор признаков, метод главных компонент (PCA), сингулярное разложение (SVD) – об этом – далее в курсе

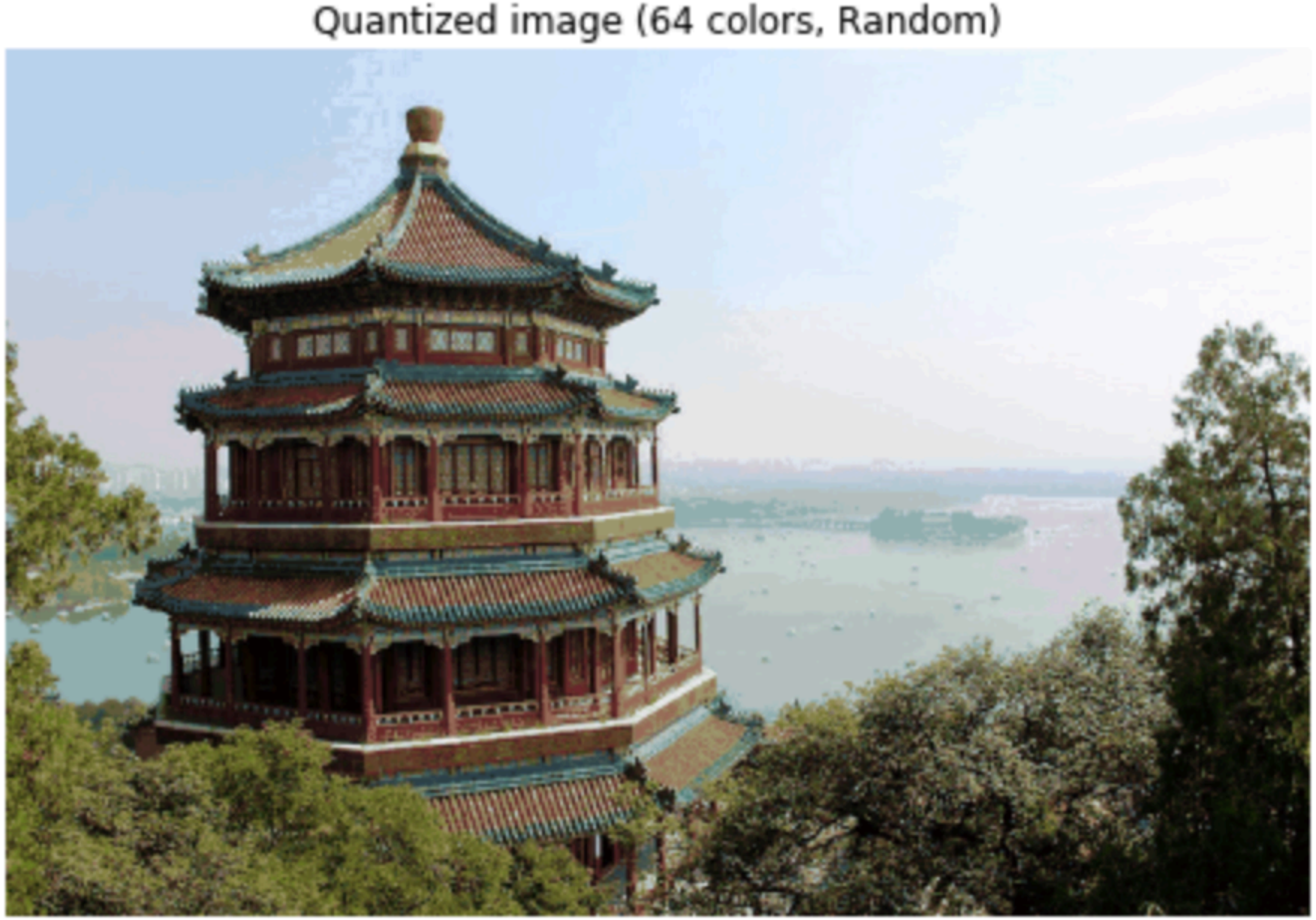
K Means++

- В зависимости от начального приближения центров кластеров может потребоваться разное время для сходимости
- Можно брать центры подальше друг от друга – для двух кластеров понятно, что это значит, а для K?
- Вариант выбора начальных приближений:
  - первый центр выбираем случано из равномерного распределения на выборке
  - Каждый следующий центр выбираем случайно из оставшихся точек так, чтобы вероятность выбрать каждую точку была пропорциональна квадрату расстояния от нее до ближайшего центра

Пример: квантизация изображений

Original image (96,615 colors)

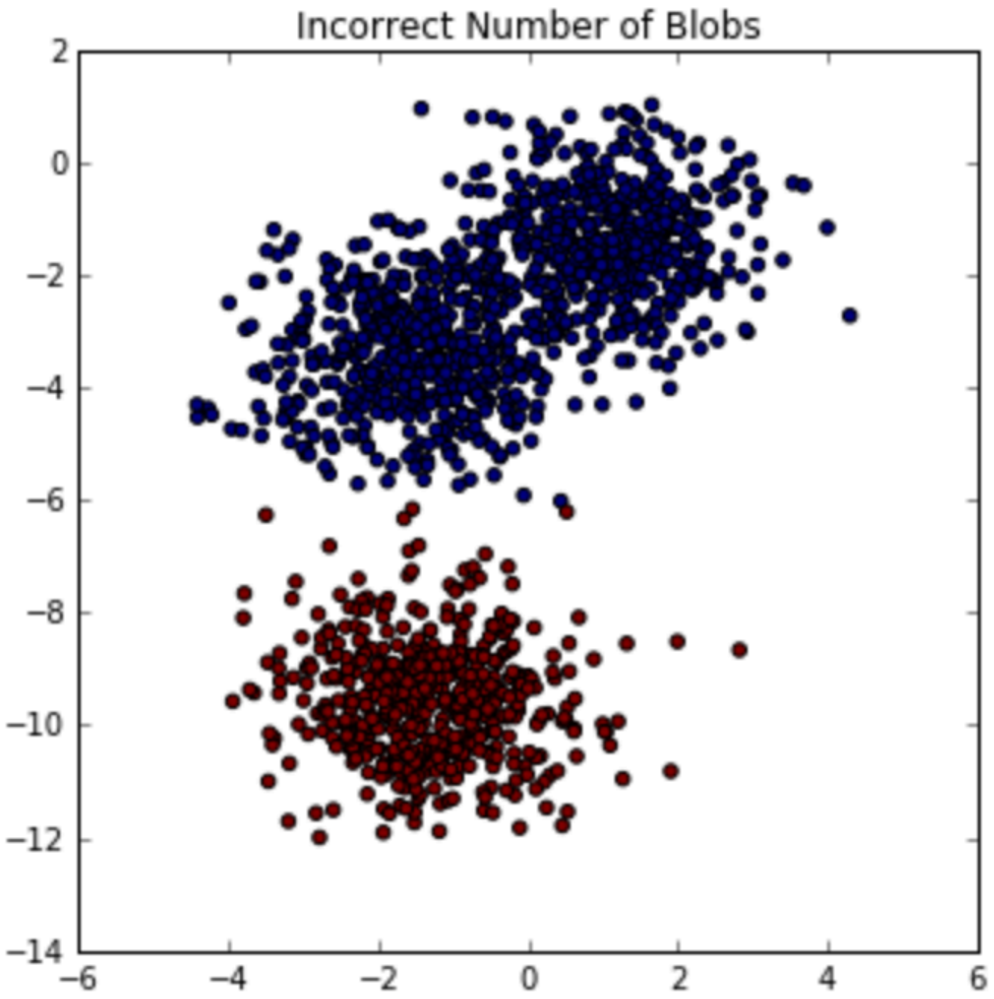
Пример: квантизация изображений



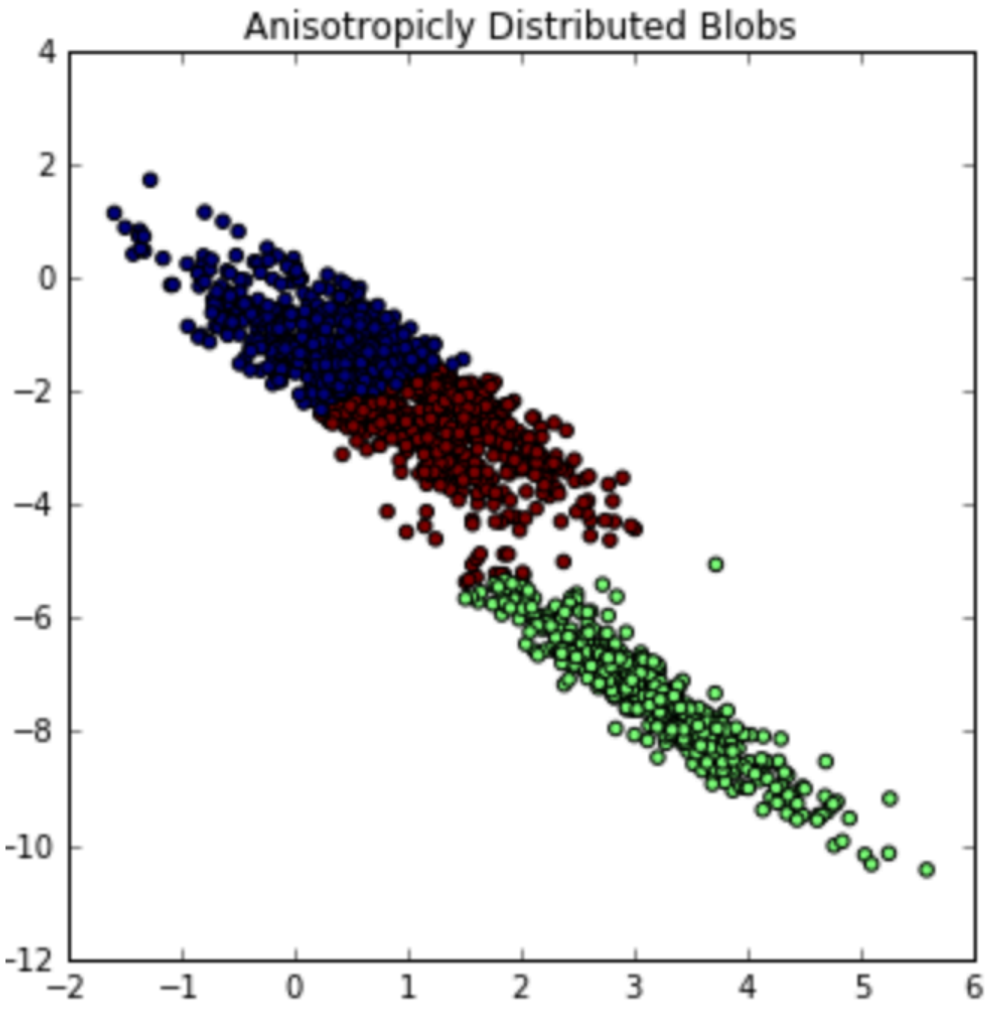
Пример: квантизация изображений

Quantized image (64 colors, K-Means)

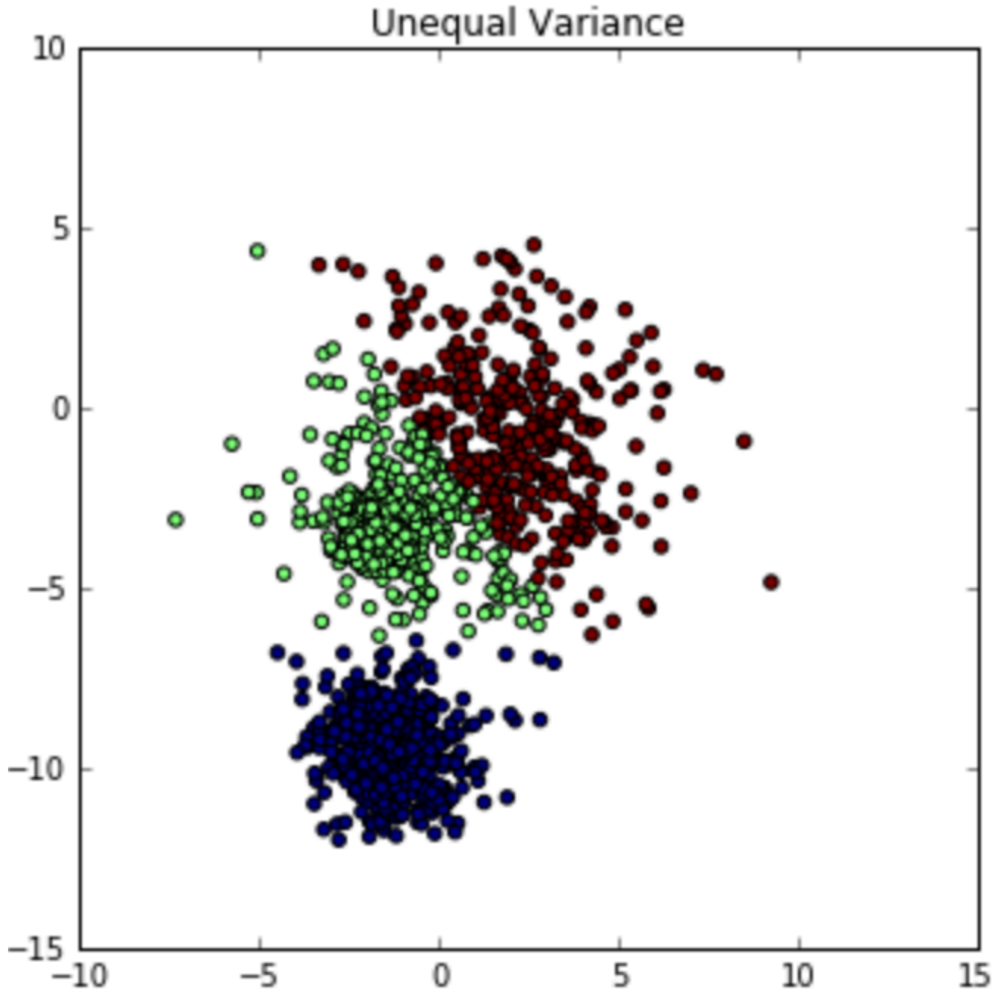
K Means и разные формы кластеров



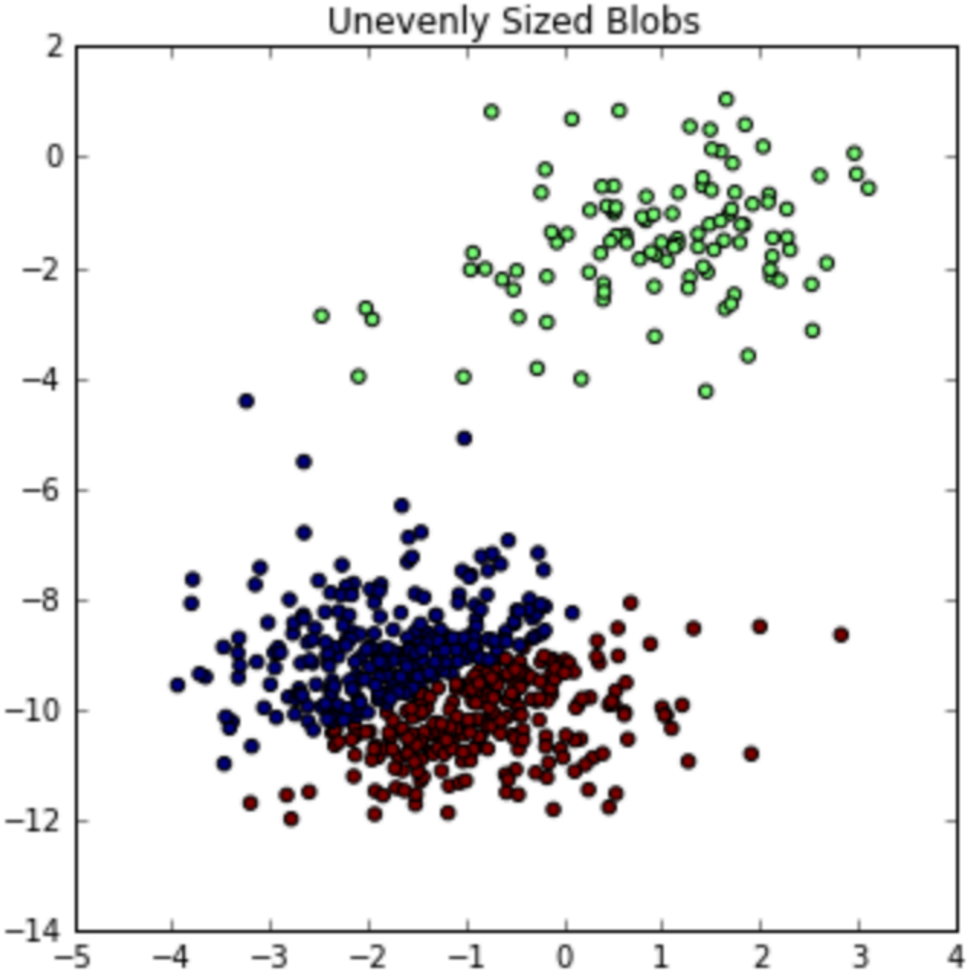
K Means и разные формы кластеров



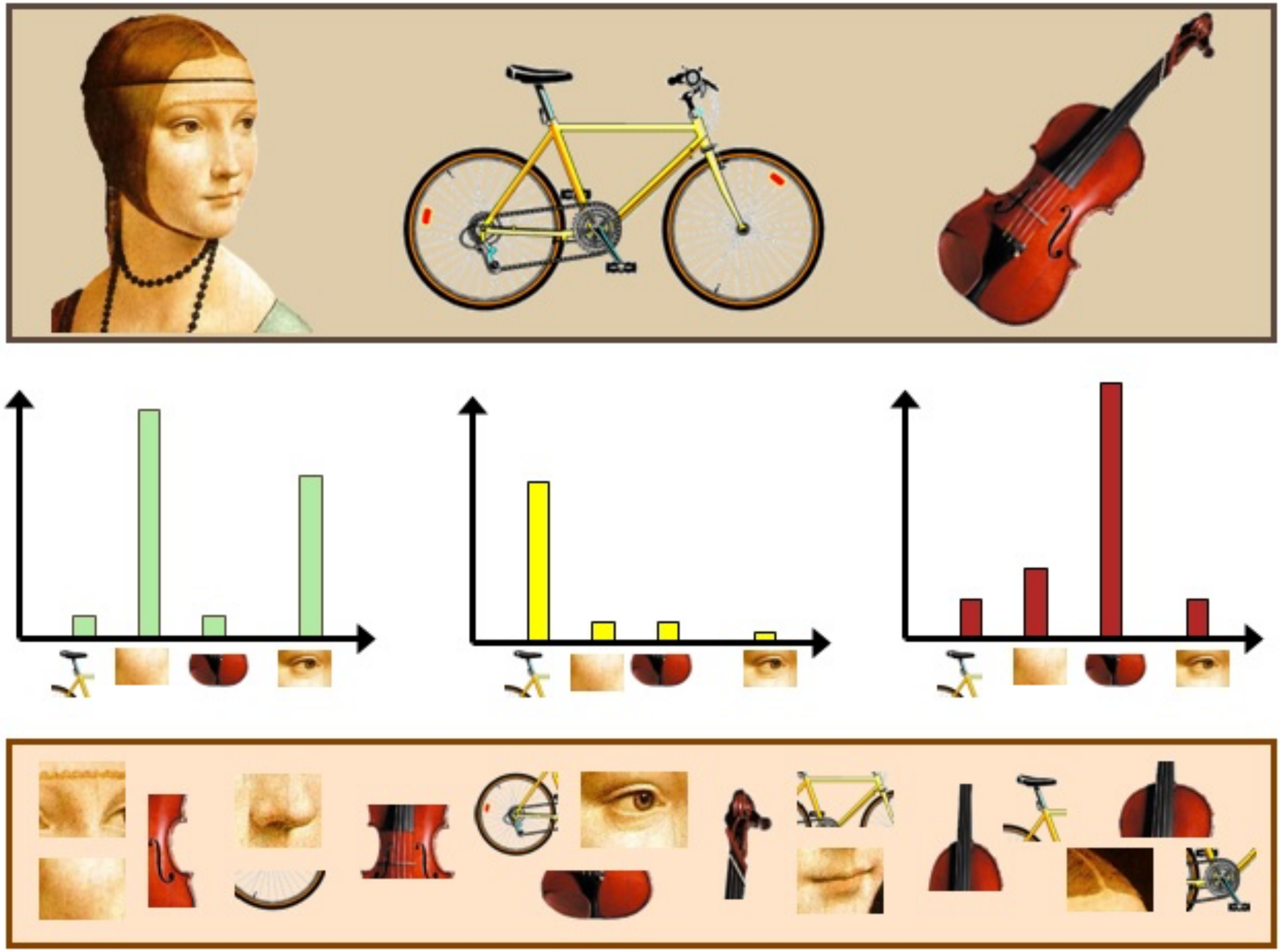
K Means и разные формы кластеров



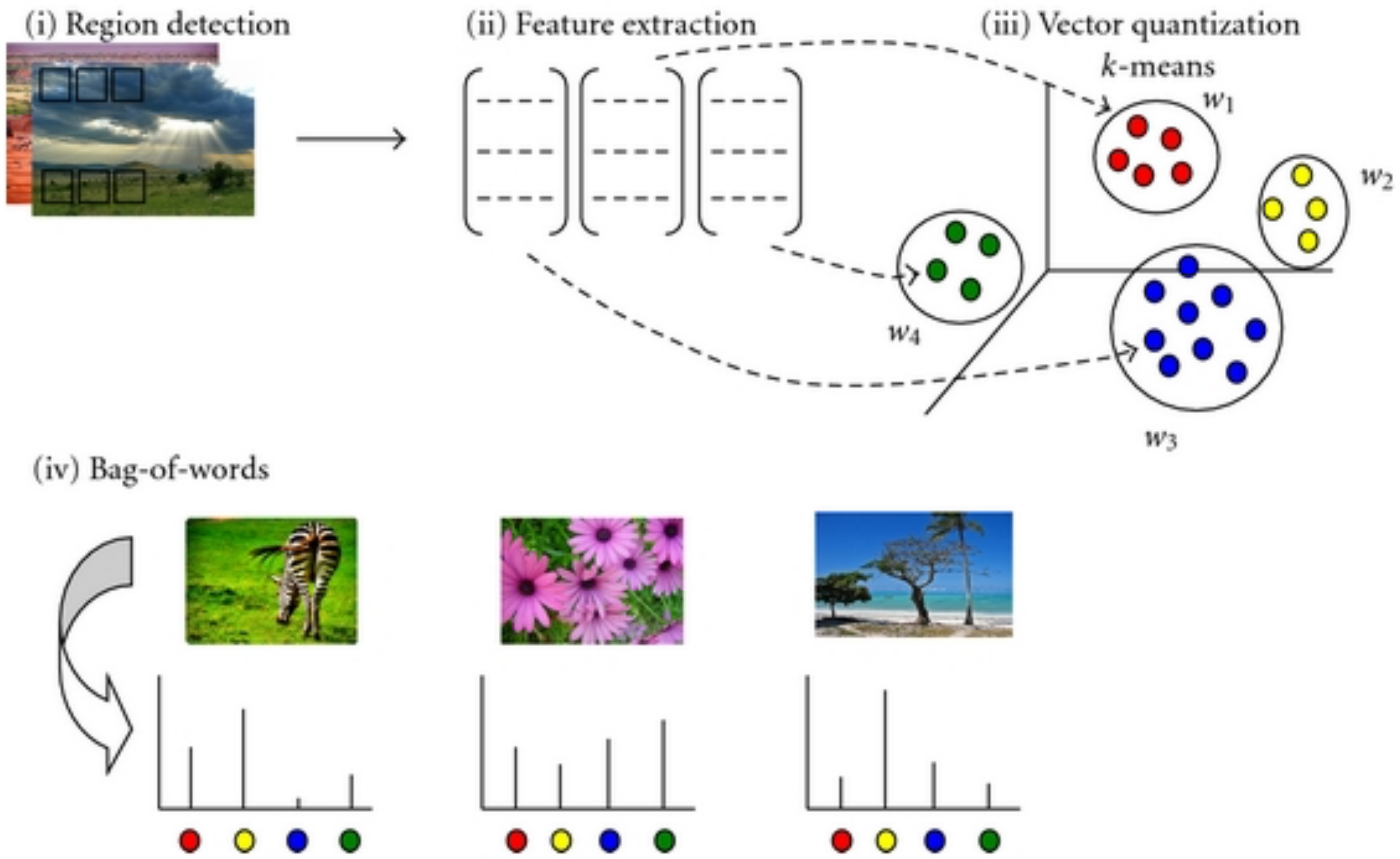
K Means и разные формы кластеров



Пример: мешок визуальных слов



Пример: мешок визуальных слов



Что оптимизирует K Means

Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to \min$$

Что оптимизирует K Means

Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i<j}[y_i = y_j]\rho(x_i, x_j)}{\sum_{i<j}[y_i = y_j]} \to \min$$

Альтернативный вариант, если есть центры кластеров:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:\, y_i = y} \rho^2(x_i, \mu_y) \to \min,$$

Что оптимизирует K Means

В 1967 году Мак Кин показал, что для его версии K Means:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i:\, y_i = y} \rho^2(x_i, \mu_y) \to \min,$$

Что оптимизирует K Means

K Means итеративно минимизирует среднее внутрикластерное расстояние:

1. Объект присваивается к тому кластеру, центр которого ближе
2. Центр кластера перемещается в среднее арифметическое векторов признаков объектов из него

Что оптимизирует K Means

K Means итеративно минимизирует среднее внутрикластерное расстояние:

1. Объект присваивается к тому кластеру, центр которого ближе
2. Центр кластера перемещается в среднее арифметическое векторов признаков объектов из него

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \underset{\mu}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}(\mu - x_i)^2$$

Что оптимизирует K Means

K Means итеративно минимизирует среднее внутрикластерное расстояние:

1. Объект присваивается к тому кластеру, центр которого ближе
2. Центр кластера перемещается в среднее арифметическое векторов признаков объектов из него

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \operatorname*{argmin}_{\mu} \frac{1}{N}\sum_{i=1}^{N}(\mu - x_i)^2$$

$$\frac{d}{d\mu}\,\frac{1}{N}\sum_{i=1}^{N}(\mu - x_i)^2 = \frac{2}{N}\sum_{i=1}^{N}(\mu - x_i) = 0 \Longrightarrow \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Итоги

1. Как работает K Means
2. Вариации: K Means Болла-Холла и Мак Кина
3. Что делать, когда данных много: Mini Batch K-Means
4. Что делать, когда много признаков: понижение размерности
5. Выбор начальных приближений: Kmeans++
6. Пример: квантизация изображений
7. Работа K means с разными формами кластеров
8. Пример: мешок визуальных слов (bag of visual words)
9. Что оптимизирует K means