Задача кластеризации

Ранее: обучение на размеченных данных (supervised learning)

Обучающая выборка:

$x_1, \ldots, x_l$ - объекты

$y_1, \ldots, y_l$ - ответы

Ранее: обучение на размеченных данных (supervised learning)

Обучающая выборка:

$x_1, \dots, x_l$ - объекты

$y_1, \dots, y_l$ - ответы

Тестовая выборка:

$x_{l+1}, \dots, x_{l+u}$

Ранее: обучение на размеченных данных (supervised learning)

Обучающая выборка:
$x_1, \dots, x_l$ - объекты
$y_1, \dots, y_l$ - ответы

Тестовая выборка:
$x_{l+1}, \dots, x_{l+u}$

В регрессии: y_i - прогнозируемая величина
В классификации: y_i - метка класса

Восстановление отображения

Считаем, что есть отображение:

$$x \mapsto y$$

Обучающая выборка – это примеры значений, по которым мы пытаемся построить $a(x)$:

$$a(x) \approx y$$

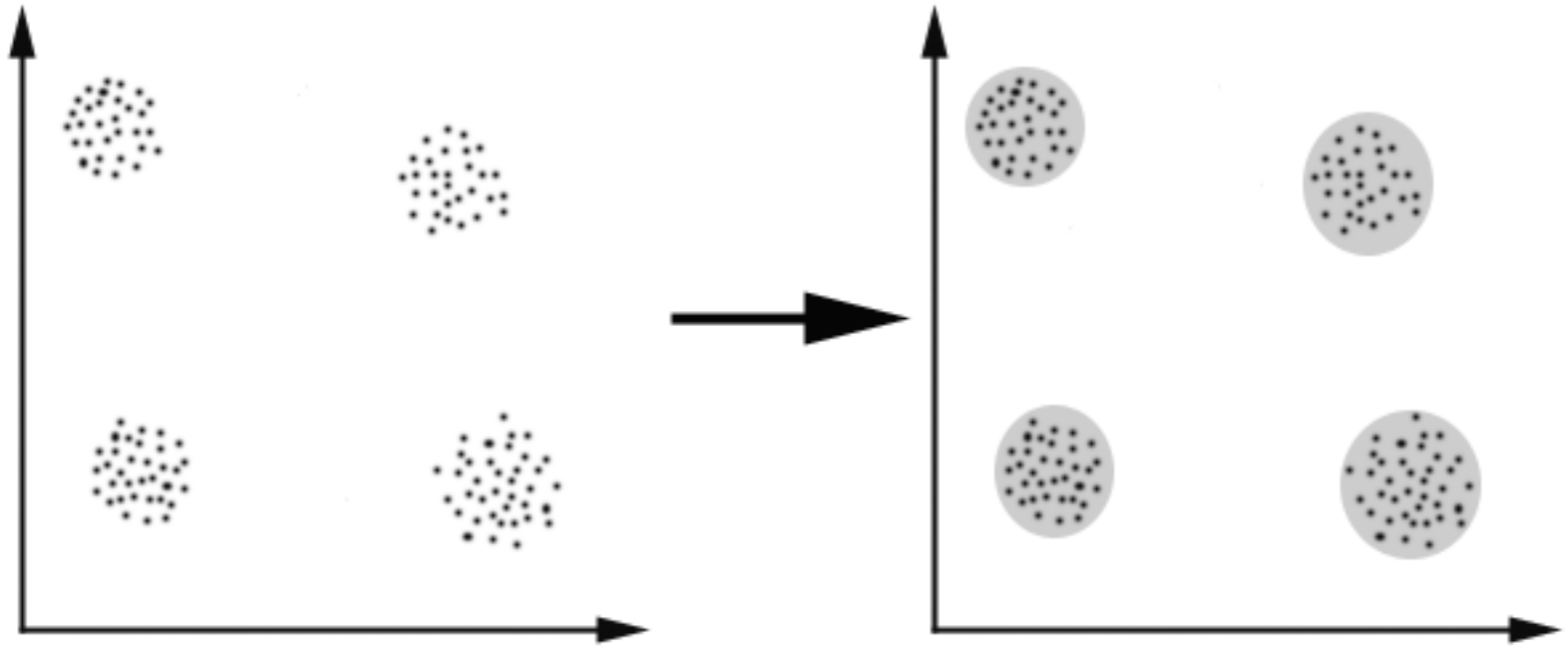
Кластеризация

«Обучающая» выборка:
$x_1, \dots, x_l$ - объекты

Она же и тестовая

Нужно поставить метки $y_1, \dots, y_l$, так, чтобы объекты с одной и той же меткой были похожи, а с разными метками – не очень похожи

Как это выглядит



Восстановление отображения в кластеризации

Считаем, что есть отображение:

$$x \mapsto y$$

Пытаемся построить $a(x)$, но примеров y теперь нет.

Нужно не приближать известные значения, а строить отображение с некоторыми хорошими свойствами.

Среднее внутрикластерное расстояние

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

Среднее межкластерное расстояние

$$F_1 = \frac{\sum\limits_{i<j}[y_i \neq y_j]\,\rho(x_i, x_j)}{\sum\limits_{i<j}[y_i \neq y_j]} \to \max$$

Придумываем метрику качества

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

$$F_1 = \frac{\sum\limits_{i<j}[y_i \neq y_j]\,\rho(x_i, x_j)}{\sum\limits_{i<j}[y_i \neq y_j]}$$

$$F_0 / F_1 \to \min$$

Резюме

- Отличия от обучения на размеченных данных
- Постановка задачи кластеризации
- Простые способы оценить качество кластеризации
- В следующем видео: о том, какими бывают задачи кластеризации