Отчёт. Кластеризация

Используемые методы

1. K-means

Входные данные: количество кластеров k, вектора с характеристиками ирисов.

Выполнение алгоритма:

1. Выбираем k начальных центроидов кластеров
2. Каждый вектор относим к тому кластеру, чей центроид является наиболее близким
3. Выполняем повторное вычисление центроидов каждого кластера
4. Повторяем, пока не достигнем условия остановки: достигнуто пороговое число итераций, центроиды кластеров больше не изменяются, достигнуто пороговое значение целевой функции

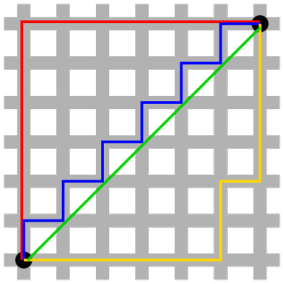
2. DBSCAN

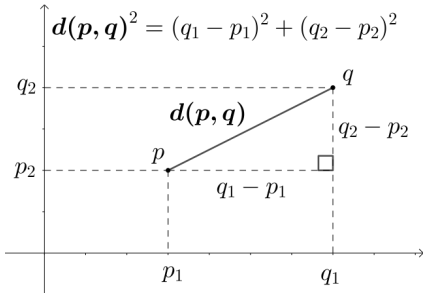
Входные данные: вектора с характеристиками ирисов

Алгоритм: DBSCAN (Density-based spatial clustering of applications with noise, плотностной алгоритм пространственной кластеризации с присутствием шума), как следует из названия, оперирует плотностью данных.

Используемые метрики

В обоих случаях была использована манхэттенская метрика и евклидова.





Отчёт. Кластеризация

Реализация

Я использовала датасет с ирисами Фишера. Так как всего компонент у каждого вектора 4, хотелось наиболее полно визуализировать данные, поэтому графики и представлены в трёхмерном формате - я не использовала 4 компоненту при визуализации, но в алгоритмах она учитывается.

k-means запрограммирован без использования sklearn, DBSCAN же с использование этой библиотеки. В втором случае я использовала функцию DBSCAN.

Функции:

- def external_eval(label)

Описание: Внешняя оценка качества кластеризации, использует средства библиотека sklearn.

Параметры: label - массив с метками классов.

Возвращаемое значение: значение в диапазоне [0,1] - качество кластеризации.

- def graph_comp(lab_ke, lab_kc, lab_de, lab_dc):

Описание: Визуализация результатов кластеризации всех методов, использованных в программе.

Параметры: lab_ke - массив с метками классов: кластеризация k-means, евклидова метрика. lab_kc - массив с метками классов: кластеризация k-means, манхэттенская метрика. lab_de - массив с метками классов, кластеризация DBSCAN, евклидова метрика. lab_dc - массив с метками классов: кластеризация DBSCAN, манхэттенская метрика

Возвращаемое значение: нету

- def internal_eval(labels)

Описание: Внутренняя оценка качества кластеризации

Параметры: labels - массив с метками классов

Отчёт. Кластеризация

Возвращаемое значение: кортеж. Первый элемент - сумма попарных расстояний между элементами одного кластера, второй элемент - сумма расстояний между элементами из разных кластеров.

- def dbscan_meth(metric='euclidean', show=False)

Описание: Кластеризации датасета при помощи DBSCAN.

Параметры: metric - используемая метрика, show - стоит ли визуализировать итерации алгоритма(может быть полезна при демонстрации правильности работы)

Возвращаемое значение: массив с метками классов

- kmeans(metric='euclidean', show=False)

Описание: Кластеризации датасета при помощи kmeans.

Параметры: metric - используемая метрика, show - стоит ли визуализировать итерации алгоритма(может быть полезна при демонстрации правильности работы)

Возвращаемое значение: массив с метками классов

Итоги

- Внешняя оценка kmeans, euclidean - 0.7386548254402864
- Внешняя оценка kmeans, cityblock - 0.7475832649918129
- Внешняя оценка DBSCAN, euclidean - 0.5989947874137124
- Внешняя оценка DBSCAN, cityblock - 0.4562659041377559
- Внутренняя оценка kmeans, euclidean - (3497.603799211151, 49877.52916031098)
- Внутренняя оценка kmeans, cityblock - (3565.0632354922777, 49742.61028774873)

Отчёт. Кластеризация

- Внутренняя оценка DBSCAN, euclidean - (5412.477948252723, 46047.78086222787)

- Внутренняя оценка DBSCAN, cityblock - (9019.707504355063, 38833.32175002311)

Библиография

- Материала курса «Анализ Неструктурированных Данных»
- https://ru.wikipedia.org/wiki/%D0%95%D0%B2%D0%BA%D0%BB%D0%B8%D0%B4%D0%BE%D0%B2%D0%B0_%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0?oldformat=true - Картинка с евклидовой метрикой
- https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D1%81%D1%82%D0%BE%D1%8F%D0%BD%D0%B8%D0%B5_%D0%B3%D0%BE%D1%80%D0%BE%D0%B4%D1%81%D0%BA%D0%B8%D1%85_%D0%BA%D0%B2%D0%B0%D1%80%D1%82%D0%B0%D0%BB%D0%BE%D0%B2 - Картинка с манхэттенской метрикой