**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра МОЭВМ**

**отчет**

**по лабораторной работе №6**

**по дисциплине «Статистические методы обработки экспериментальных данных»**

**Тема: Кластерный анализ. Метод k-средних.**

|  |  |  |
| --- | --- | --- |
| Студент гр. 5381 |  | Лянгузов А. А. |
| Преподаватель |  | Середа В. И. |

Санкт-Петербург

2019

**Цель работы.**

Освоение основных понятий и некоторых методов кластерного анализа.

**Задание**

Дано конечное множество из объектов, представленных двумя признаками (в качестве этого множества принимаем исходную двумерную выборку, сформированную ранее в лабораторной работе №4). Выполнить разбиение исходного множества объектов на конечное число подмножеств (кластеров) с использованием метода k-средних. Полученные результаты содержательно проинтерпретировать.

**Основные теоретические положения.**

*Кластерный анализ* (англ. *cluster analysis)* – это метод классификации многомерных наблюдений на основе определения сходства или близости (расстояния) между объектами. Цель кластерного анализа заключается в определении однородных в некотором смысле групп, которые называются кластерами.

Алгоритм кластерного анализа включает пять этапов.

* 1 этап. Представление исходных данных в виде матрицы (таблицы "объект – признак").
* 2 этап. Определение сходства объектов.
* 3 этап. Выбор метода объединения объектов в кластеры.
* 4 этап. Определение оптимального числа кластеров.
* 5 этап. Интерпретация кластеров и качества разбиения.

Меры расстояний:

Для того, чтобы сравнивать два объекта, необходимо иметь *критерий*, на основании которого будет происходить сравнение. Как правило, таким критерием является *расстояние* между объектами.

Есть множество мер расстояния, [рассмотрим](http://www.statsoft.ru/HOME/TEXTBOOK/modules/stcluan.html) несколько из них:  
Евклидово расстояние — наиболее распространенное расстояние. Оно является геометрическим расстоянием в многомерном пространстве:

Квадрат евклидова расстояния. Иногда может возникнуть желание возвести в квадрат стандартное евклидово расстояние, чтобы придать большие веса более отдаленным друг от друга объектам.   
Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является просто средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако отметим, что для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).   
Расстояние Чебышева. Это расстояние может оказаться полезным, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением). 

**Алгоритм k-means (k-средних)**

Наиболее простой, но в то же время достаточно неточный метод кластеризации в классической реализации. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k. Действие алгоритма таково, что он стремится минимизировать среднеквадратичное отклонение на точках каждого кластера. Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на какой-то итерации не происходит изменения кластеров.

**Экспериментальные результаты.**

Двумерная выборка: n = 107

Таблица 1

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| v | 501.00 | 369.00 | 344.00 | 473.00 | 426.00 | 528.00 | 497.00 | 467.00 | 506.00 | 431.00 | 454.00 |
| E | 130.40 | 84.30 | 86.80 | 137.90 | 121.10 | 163.40 | 147.30 | 140.50 | 158.40 | 125.00 | 131.10 |
| v | 371.00 | 482.00 | 393.00 | 441.00 | 463.00 | 440.00 | 481.00 | 340.00 | 468.00 | 397.00 | 496.00 |
| E | 89.20 | 139.90 | 103.20 | 122.80 | 129.10 | 128.50 | 135.20 | 85.10 | 142.00 | 108.60 | 143.10 |
| v | 434.00 | 541.00 | 352.00 | 438.00 | 453.00 | 423.00 | 351.00 | 525.00 | 409.00 | 469.00 | 386.00 |
| E | 122.30 | 146.80 | 87.70 | 134.90 | 119.50 | 131.10 | 89.00 | 165.90 | 121.00 | 131.50 | 95.50 |
| v | 505.00 | 436.00 | 488.00 | 449.00 | 493.00 | 512.00 | 472.00 | 423.00 | 465.00 | 351.00 | 359.00 |
| E | 137.50 | 114.30 | 134.10 | 124.50 | 129.70 | 169.90 | 134.20 | 130.80 | 140.70 | 102.90 | 71.90 |
| v | 457.00 | 467.00 | 400.00 | 418.00 | 492.00 | 434.00 | 510.00 | 392.00 | 463.00 | 459.00 | 397.00 |
| E | 126.40 | 135.10 | 114.60 | 118.60 | 137.50 | 110.50 | 140.60 | 82.70 | 125.00 | 145.40 | 106.80 |
| v | 424.00 | 436.00 | 429.00 | 398.00 | 493.00 | 522.00 | 518.00 | 463.00 | 437.00 | 386.00 | 493.00 |
| E | 119.00 | 116.70 | 112.90 | 109.00 | 154.50 | 154.50 | 144.40 | 121.20 | 121.80 | 105.80 | 151.20 |
| v | 414.00 | 480.00 | 585.00 | 562.00 | 508.00 | 421.00 | 463.00 | 422.00 | 406.00 | 544.00 | 345.00 |
| E | 113.50 | 153.90 | 177.70 | 175.90 | 159.00 | 117.80 | 136.70 | 122.90 | 110.10 | 166.70 | 95.90 |
| v | 478.00 | 393.00 | 437.00 | 448.00 | 458.00 | 422.00 | 468.00 | 430.00 | 371.00 | 543.00 | 471.00 |
| E | 126.60 | 122.80 | 115.10 | 121.90 | 121.70 | 115.70 | 144.90 | 104.30 | 91.90 | 155.40 | 143.90 |
| v | 475.00 | 521.00 | 353.00 | 437.00 | 362.00 | 490.00 | 484.00 | 459.00 | 480.00 | 482.00 | 522.00 |
| E | 132.00 | 139.60 | 98.00 | 118.40 | 111.70 | 139.90 | 140.40 | 136.70 | 153.30 | 148.20 | 143.80 |
| v | 576.00 | 390.00 | 514.00 | 442.00 | 421.00 | 443.00 | 438.00 | 429.00 |  |  |  |
| E | 166.40 | 91.40 | 153.60 | 115.40 | 107.90 | 121.90 | 126.70 | 120.90 |  |  |  |

**Обработка результатов эксперимента.**

1. Масштабирование выборки.

Масштабирование выборки осуществляется по формулам:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| v | 0.65714286 | 0.11836735 | 0.01632653 | 0.54285714 | 0.35102041 | 0.76734694 | 0.64081633 | 0.51836735 | 0.67755102 | 0.37142857 | 0.46530612 |
| E | 0.5529301 | 0.1172023 | 0.1408318 | 0.6238185 | 0.4650284 | 0.8648393 | 0.7126654 | 0.6483932 | 0.8175803 | 0.5018904 | 0.5595463 |
| v | 0.12653061 | 0.57959184 | 0.21632653 | 0.41224490 | 0.50204082 | 0.40816327 | 0.57551020 | 0.00000000 | 0.52244898 | 0.23265306 | 0.63673469 |
| E | 0.1635161 | 0.6427221 | 0.2958412 | 0.4810964 | 0.5406427 | 0.5349716 | 0.5982987 | 0.1247637 | 0.6625709 | 0.3468809 | 0.6729679 |
| v | 0.38367347 | 0.82040816 | 0.04897959 | 0.40000000 | 0.46122449 | 0.33877551 | 0.04489796 | 0.75510204 | 0.28163265 | 0.52653061 | 0.18775510 |
| E | 0.4763705 | 0.7079395 | 0.1493384 | 0.5954631 | 0.4499055 | 0.5595463 | 0.1616257 | 0.8884688 | 0.4640832 | 0.5633270 | 0.2230624 |
| v | 0.67346939 | 0.39183673 | 0.60408163 | 0.44489796 | 0.62448980 | 0.70204082 | 0.53877551 | 0.33877551 | 0.51020408 | 0.04489796 | 0.07755102 |
| E | 0.6200378 | 0.4007561 | 0.5879017 | 0.4971645 | 0.5463138 | 0.9262760 | 0.5888469 | 0.5567108 | 0.6502836 | 0.2930057 | 0.0000000 |
| v | 0.47755102 | 0.51836735 | 0.24489796 | 0.31836735 | 0.62040816 | 0.38367347 | 0.69387755 | 0.21224490 | 0.50204082 | 0.48571429 | 0.23265306 |
| E | 0.5151229 | 0.5973535 | 0.4035917 | 0.4413989 | 0.6200378 | 0.3648393 | 0.6493384 | 0.1020794 | 0.5018904 | 0.6947070 | 0.3298677 |
| v | 0.34285714 | 0.39183673 | 0.36326531 | 0.23673469 | 0.62448980 | 0.74285714 | 0.72653061 | 0.50204082 | 0.39591837 | 0.18775510 | 0.62448980 |
| E | 0.4451796 | 0.4234405 | 0.3875236 | 0.3506616 | 0.7807183 | 0.7807183 | 0.6852552 | 0.4659735 | 0.4716446 | 0.3204159 | 0.7495274 |
| v | 0.30204082 | 0.57142857 | 1.00000000 | 0.90612245 | 0.68571429 | 0.33061224 | 0.50204082 | 0.33469388 | 0.26938776 | 0.83265306 | 0.02040816 |
| E | 0.3931947 | 0.7750473 | 1.0000000 | 0.9829868 | 0.8232514 | 0.4338374 | 0.6124764 | 0.4820416 | 0.3610586 | 0.8960302 | 0.2268431 |
| v | 0.56326531 | 0.21632653 | 0.39591837 | 0.44081633 | 0.48163265 | 0.33469388 | 0.52244898 | 0.36734694 | 0.12653061 | 0.82857143 | 0.53469388 |
| E | 0.5170132 | 0.4810964 | 0.4083176 | 0.4725898 | 0.4706994 | 0.4139887 | 0.6899811 | 0.3062382 | 0.1890359 | 0.7892250 | 0.6805293 |
| v | 0.55102041 | 0.73877551 | 0.05306122 | 0.39591837 | 0.08979592 | 0.61224490 | 0.58775510 | 0.48571429 | 0.57142857 | 0.57959184 | 0.74285714 |
| E | 0.5680529 | 0.6398866 | 0.2466919 | 0.4395085 | 0.3761815 | 0.6427221 | 0.6474480 | 0.6124764 | 0.7693762 | 0.7211720 | 0.6795841 |
| v | 0.96326531 | 0.20408163 | 0.71020408 | 0.41632653 | 0.33061224 | 0.42040816 | 0.40000000 | 0.36326531 |  |  |  |
| E | 0.8931947 | 0.1843100 | 0.7722117 | 0.4111531 | 0.3402647 | 0.4725898 | 0.5179584 | 0.4631380 |  |  |  |

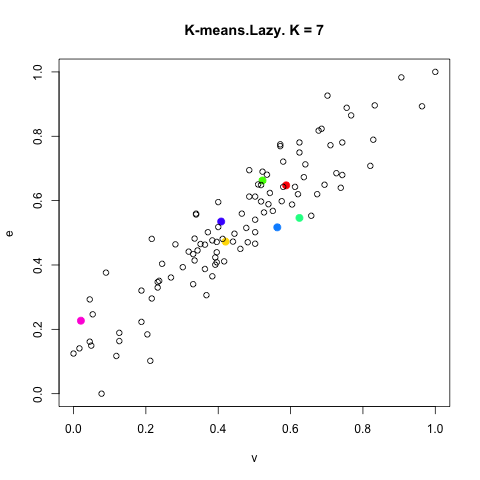
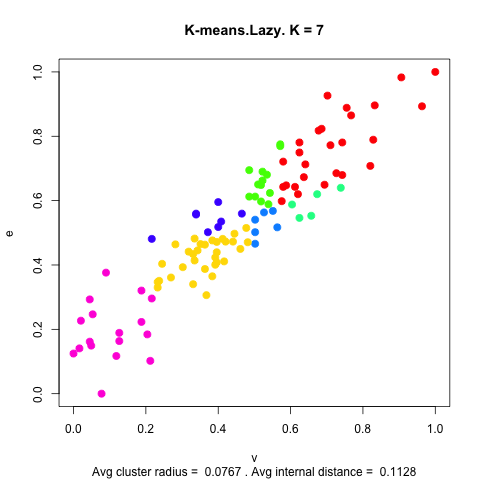
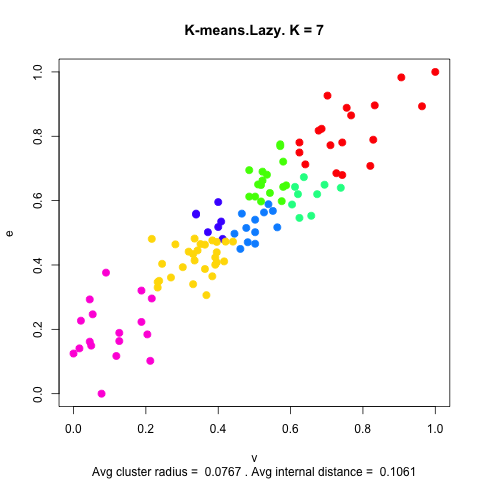
1. K-means кластеризация.

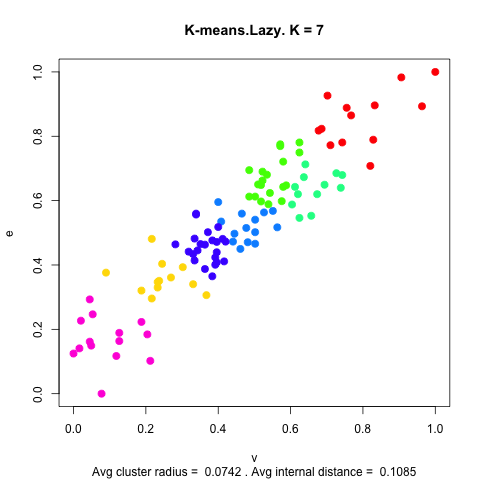
Количество кластеров взяли равным 7.

В качестве начальных центров кластеров взяли рандомные пары .

Пересчет центров кластеров после распределения всех пар осуществляется по формулам:

Результаты представим графически.

1. Случайные центры:
2. Кластеризация по данным центрам:
3. Пересчет центров и кластеризация:
4. Пересчет центров осуществляется пока центры не перестанут меняться.

В данном случае конечный вариант:

**Вывод.**

В ходе выполнения лабораторной работы ознакомились кластерным анализом, в частности, k-means - методом *k*-средних.

К недостаткам k-means можно отнести:

* необходимость заранее знать количество кластеров;
* чувствительность к выбору начальных центров кластеров.

В отчете использовался вариант k-means, когда пересчет центров кластеров осуществляется после распределения всех пар по кластерам, пока центры не перестанут меняться (кодовое название: “Lazy”). Однако, разработанный скрипт также содержит другой вариант метода k-средних, где пересчет центров происходит после каждого добавления точки (кодовое название: “Long”).

В результате работы алгоритма k-средних, выборка была разделена на 7 кластеров.