|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | **Министерство науки и высшего образования Российской Федерации**  Калужский филиал  федерального государственного бюджетного  образовательного учреждения высшего образования  ***«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»***  ***(КФ МГТУ им. Н.Э. Баумана)*** |

|  |  |
| --- | --- |
| **ФАКУЛЬТЕТ** | **ИУК «Информатика и управление»** |
| **КАФЕДРА** | **ИУК5 «Информатика и вычислительная техника»** |

**Лабораторная работа №2**

**«Метод главных компонент и кластеризация»**

**ДИСЦИПЛИНА: «Проектирование программного обеспечения»**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил: студент гр. ИУК4-11М | |  |  | ( | Сафронов Н.С. | ) |
|  |  |  | (подпись) |  | (Ф.И.О.) |  |
| Проверил: | |  |  | ( | Потапов А.Е. | ) |
|  |  |  | (подпись) |  | (Ф.И.О.) |  |

|  |  |
| --- | --- |
| Дата сдачи (защиты):  Результаты сдачи (защиты): | |
|  | - Балльная оценка:  - Оценка: |

Калуга, 2024

**Цель работы:** формирование практических навыков создания интеллектуальных систем с обучением без учителя (unsupervised learning).

**Задачи:** подготовить данные для эксперимента. Выполнить анализ при помощи метода главных компонент. Выполнить кластеризацию при помощи методов K-Means и агломеративного. Проанализировать эффективность вариации гиперпараметров моделей и применения композиции методов.

**Результаты выполнения работы**

Отмасштабируем выборку с помощью StandardScaler с параметрами по умолчанию и понизим размерность с помощью PCA.





**Рисунок 1** – Масштабирование и понижение размерности выборки

**Вопрос 1:**

Какое минимальное число главных компонент нужно выделить, чтобы объяснить 90% дисперсии исходных (отмасштабированных) данных?



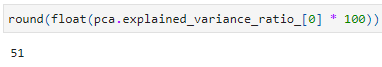
**Рисунок 2** – Форма результата понижения размерности методом главных компонент

Таким образом, в результате понижения размерности было выделено 65 главных компонент.

**Ответ:** 65.

**Вопрос 2:**

Сколько процентов дисперсии приходится на первую главную компоненту? Округлите до целых процентов.



**Рисунок 3** – Результат получения дисперсии первой главной компоненты

**Ответ:** 51.



**Рисунок 4** – Визуализация данных в проекции на первые две главные компоненты

**Вопрос 3:**

Если все получилось правильно, Вы увидите сколько-то кластеров, почти идеально отделенных друг от друга. Какие виды активности входят в эти кластеры?

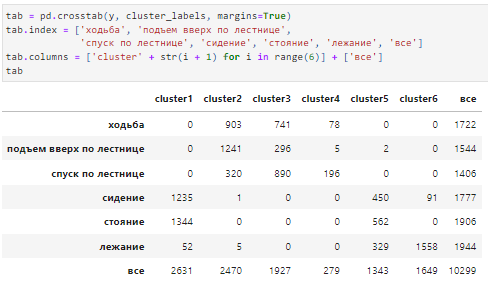
**Ответ:** 2 кластера: (ходьба, подъем вверх по лестнице, спуск по лестнице) и (сидение, стояние, лежание).

****

**Рисунок 5** – Кластеризация методом KMeans

****

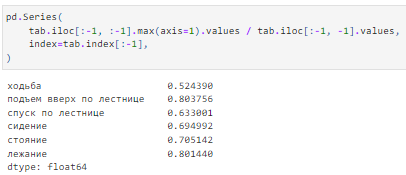
**Рисунок 6** – Визуализация кластеризации методом KMeans

****

**Рисунок 7** – Составы полученных кластеров

**Вопрос 4:**

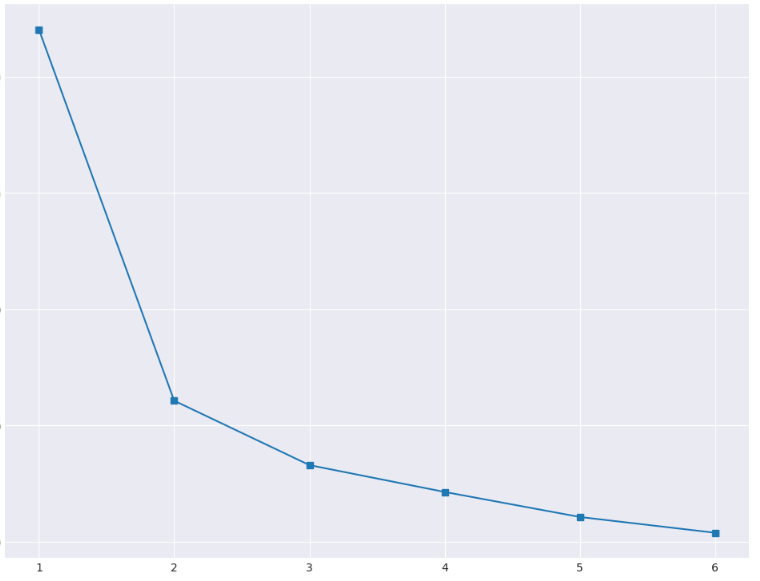
Какой вид активности отделился от остальных лучше всего в терминах простой метрики, описанной выше?



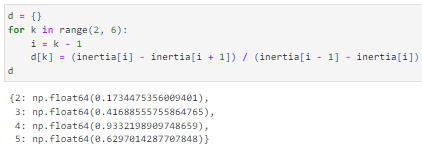
**Рисунок 8** – Максимальная доля объектов в классе, отнесенных к кластеру

**Ответ:** перечисленные варианты не подходят.

Воспользуемся методом локтя:

****

**Рисунок 9** – Величина инерции для различного количества кластеров



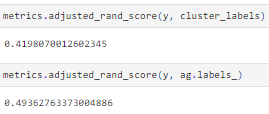
**Рисунок 10** – Значение метрики разности для различных значений количества кластеров

**Вопрос 5:**

Какое количество кластеров оптимально выбрать, согласно методу локтя?

Выбираем наименьшее значение метрики разности.

**Ответ:** 2.

****

**Рисунок 11** – Значение ARI для KMeans и агломеративной кластеризации

**Вопрос 6:**  
Отметьте все верные утверждения.

KMeans имеет меньшую метрику, следовательно справился хуже.

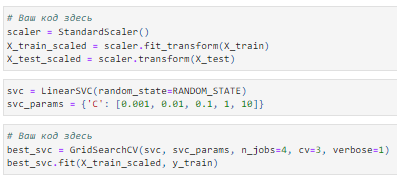
ARI опирается на пары объектов и оценивает, находятся ли они в одном кластере или в разных кластерах в двух сравниваемых разбиениях (истинном и предсказанном).

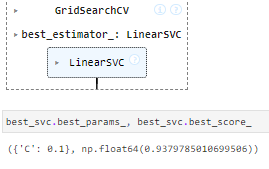
Если разбиение произведено случайным образом, вероятность того, что одна и та же пара объектов попадет в один и тот же кластер в двух разных разбиениях, очень низка. Поэтому вклад этих пар в итоговый индекс также будет низким, что и приводит к значению ARI, близкому к нулю.

**Ответ:**

* Согласно ARI, KMeans справился с кластеризацией хуже, чем Agglomerative Clustering
* Для ARI не имеет значения какие именно метки присвоены кластерам, имеет значение только разбиение объектов на кластеры
* В случае случайного разбиения на кластеры ARI будет близок к нулю

Воспользуемся методом опорных векторов:

****

****

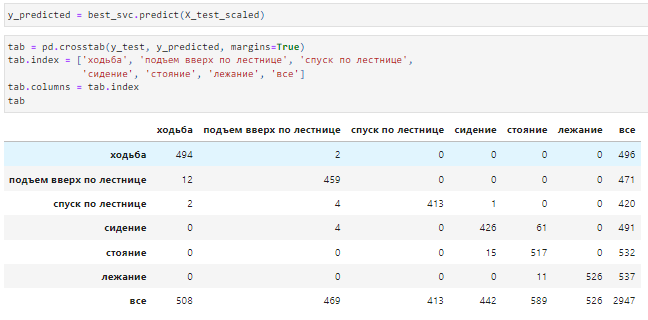
**Рисунок 12** – Метод опорных векторов

**Вопрос 7**

Какое значение гиперпараметра C было выбрано лучшим по итогам кросс-валидации?

**Ответ:** 0.1.

Определим вид активности для тестовой выборки и сравним с заданными значениями.

****

**Рисунок 13** – Сравнение результата метода опорных векторов и исходных данных (матрица неточностей)

**Вопрос 8:**

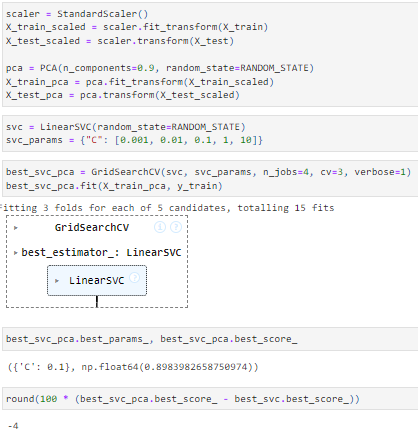
Какой вид активности SVM определяет хуже всего в терминах точности? Полноты?

Точность равняется отношению соответствующего диагонального элемента матрицы неточностей и суммы всей строки класса. Полнота – отношению диагонального элемента матрицы и суммы всего столбца класса

Таким образом, наименьшее значения точности у сидения:

Наименьшее значения полноты у стояния:

**Ответ:** по точности – у сидения, по полноте – у стояния.



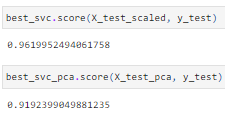
**Рисунок 14** – Обучение PCA на обучающей выборке и применение к основной, настройка гиперпараметра C

**Вопрос 9:**

Какова разность между лучшим качеством (долей верных ответов) на кросс-валидации в случае всех 561 исходных признаков и во втором случае, когда применялся метод главных компонент? Округлите до целых процентов.

**Ответ:** 4.

**Вопрос 10:**  
Выберите все верные утверждения.



**Рисунок 15 –** Значение точности для PCA и простой модели

Как мы можем заметить, значения отличаются менее чем на 10%.

PCA используется для визуализации многомерных данных, так как сводит их к двум или трем измерениям, что удобно для отображения на графиках. PCA основан на вычислении собственных значений и собственных векторов ковариационной матрицы данных, что делает его линейным методом. Это вычислительно менее затратный процесс по сравнению с t-SNE.

PCA строит новые признаки (главные компоненты) как линейные комбинации исходных признаков. Каждая главная компонента — это взвешенная сумма исходных признаков, где веса (коэффициенты) определяются как направления максимальной вариации в данных. Эти линейные комбинации могут не иметь явного физического смысла или легко интерпретируемого значения. Например, если в исходных данных есть признаки, такие как «вес» и «рост», PCA может создать компоненту, которая представляет собой нечто вроде 0.5 × "вес" + 0.7 × "рост", что не всегда легко объяснить в реальном мире.

**Ответ:**

* PCA можно использовать для визуализации данных, однако для этой задачи есть и лучше подходящие методы, например, tSNE. Зато PCA имеет меньшую вычислительную сложность
* PCA строит линейные комбинации исходных признаков, и в некоторых задачах они могут плохо интерпретироваться человеком

**Вывод**: в ходе выполнения лабораторной работы были получены практические навыки создания интеллектуальных систем с обучением без учителя (unsupervised learning).