Практическое задание № 6

КЛАСТЕРИЗАЦИЯ С ИСПОЛЬЗОВАНИЕ АЛГОРИТМА K-СРЕДНИХ

Цель работы: получить практику анализа статистических данных с использованием алгоритма K-средних.

Содержание задания

1. Общие сведения

1. Ознакомиться с материалами лекции № 6.
2. Установить необходимое программное обеспечение. При выполнении задания наверняка понадобятся **Python 3**, **NumPy**, **SciPy**, и **Matplotlib**.
3. Ознакомиться с содержимым папки с заданием, которая включает в себя файлы, представленные ниже.

`main.py` – «основной» модуль, необходимый для выполнения задания, который поможет выполнить его поэтапно. Настоящий программный код не требует какой-либо коррекции!

`data.mat` – база данных для выполнения задания.

`bird_small.png` – тестовое изображение для выполнения задания.

`findClosestCentroids.py` – модуль, содержащий функцию `findClosestCentroids`, которая позволяет выполнить поиск ближайших средних для объектов, заключенных в матрице объекты-признаки.

`computeCentroids.py` – модуль, содержащий функцию `computeCentroids`, которая необходима для вычисления новых средних для точек данных, попавших в соответствующие кластеры.

`kMeansInitCentroids.py` – модуль, содержащий функцию `kMeansInitCentroids`, которая необходима для выполнения случайной инициализации средних.

`runkMeans.py` – модуль, содержащий функцию `runkMeans`, которая необходима для выполнения алгоритма K-средних. Данный модуль не требует коррекции!

4. Поэтапно выполнить задание, связанное с реализацией и исследованием нейронной сети прямого распространения.
5. Ответить на вопросы, необходимые для составления отчета по данному практическому заданию. Отчет сдается на проверку в печатной или письменной форме в указанные сроки.

2. Алгоритм K-средних

При выполнении данного задания требуется заполнить пустые места программного кода в блоках с комментарием «Ваш код здесь». Данную процедуру необходимо выполнить для следующих функций: `findClosestCentroids`, `computeCentroids`, `kMeansInitCentroids`.

1. Загрузите базу данных, находящуюся в файле **`data.mat`**. Она представляет собой некоторое искусственное множество объектов, описываемых двумерным вектором признаков. В отличие от предыдущих практических заданий база данных не является размеченной, а сам алгоритм K-средних, который потребуется применить для анализа структуры данных, является разновидностью методов обучения без учителя. Метод K-средних представляет итерационную процедуру, которая начинается со случайного определения начальных значений средних, а затем сопровождается итерационным назначением примеров ближайшим средним и пересчетом этих средних с учетом объединения данных в группы.

2. Завершите код в модуле **`findClosestCentroids.py`**, который позволит выполнить поиск ближайших средних для объектов, заключенных в матрице объекты-признаки. Выполнение данной процедуры было представлено в лекции № 6. При выполнении данной части задания могут понадобиться функции из библиотеки **NumPy**, представленные ниже.

`sum` – позволяет вычислить сумму элементов вдоль определенной размерности двумерного массива и сумму всех элементов для одномерного массива.

`zeros` – позволяет сформировать массив заданной формы и типа, состоящий из нулевых значений.

`repmat` – выполняет повторение массивов размерности 0, 1 и 2 вдоль определенной размерности.

`argmin` – выполняет поиск индексов минимальных значений вдоль определенной размерности.

Также полезным может оказаться оператор поэлементное возведение компонентов двумерного и одномерного массивов в квадрат: `** 2`.

3. Завершите код в модуле **`computeCentroids.py`**, который позволит вычислить новые средние для точек данных, попавших в соответствующие кластеры. Выполнение данной процедуры было представлено в лекции № 6. При выполнении данной части задания могут понадобиться следующие функции из библиотеки **NumPy**: `sum` и `where`.

`where` – позволяет возвратить номера элементов массива, удовлетворяющие определенному условию.

4. После завершения вышеуказанных пунктов выполните файл **`main.py`** для анализа результата кластеризации с использованием алгоритма K-средних. Результат должен быть идентичным тому, что представлен на рис. 1, если число средних равно 3, а количество итераций алгоритма равно 10. Обратите внимание, что поочередный вызов функций `findClosestCentroids` и `computeCentroids` из соответствующих модулей выполнен внутри функции `runkMeans`, которая находится в модуле **`runkMeans.py`**. Выполните самостоятельный анализ программного кода данной функции.

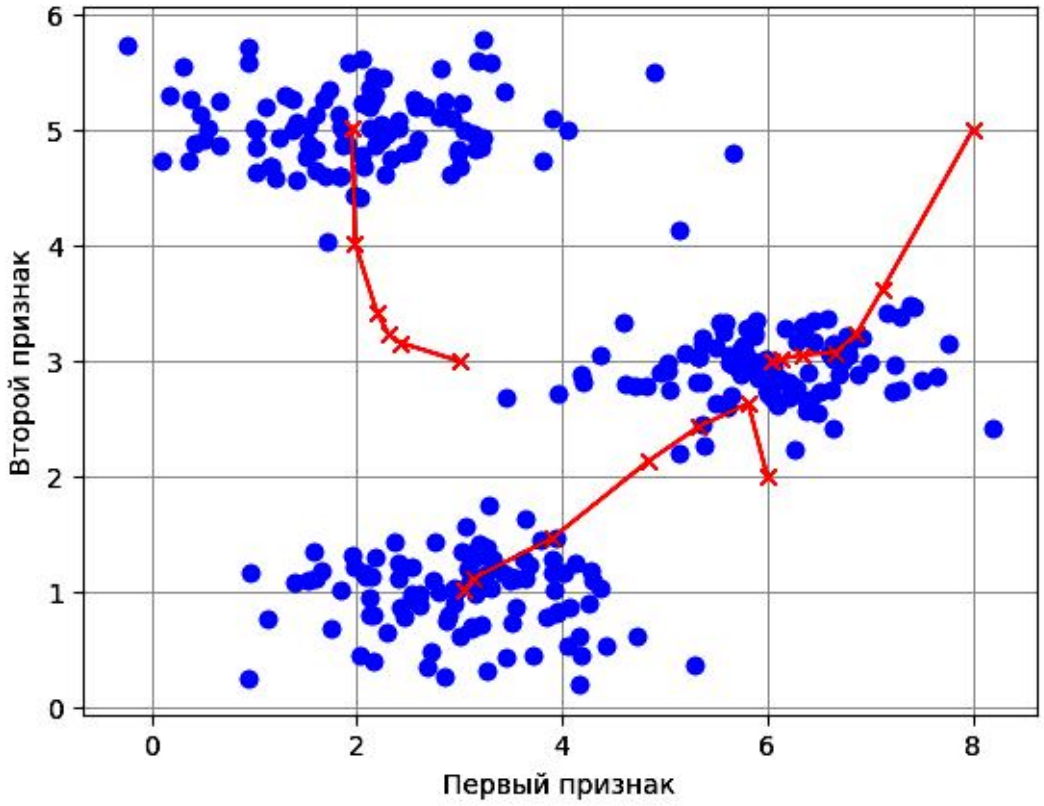

Рис. 1. Результат кластеризации данных с использованием алгоритма K-средних

5. Завершите код в модуле **`kMeansInitCentroids.py`**, который позволит выполнить случайную инициализацию средних, выбирая случайным образом K примеров из матрицы объекты-признаки. Выполнение данной процедуры было представлено в лекции № 6. При выполнении данной части задания может понадобиться функция `choice` из библиотеки **NumPy**.

`choice` – позволяет сгенерировать случайную выборку из заданного одномерного массива.

6. После выполнения предыдущего пункта выполните применение алгоритма K-средних к задаче сжатия цифрового изображения **bird_small.png**. Теоретический материал по данному вопросу был рассмотрен в рамках лекции № 6. При выполнении данного пункта достаточно выполнить файл **main.py** и пронаблюдать полученный результат. Пример сжатия цифрового изображения с использованием алгоритма K-средних представлен на рис. 2.

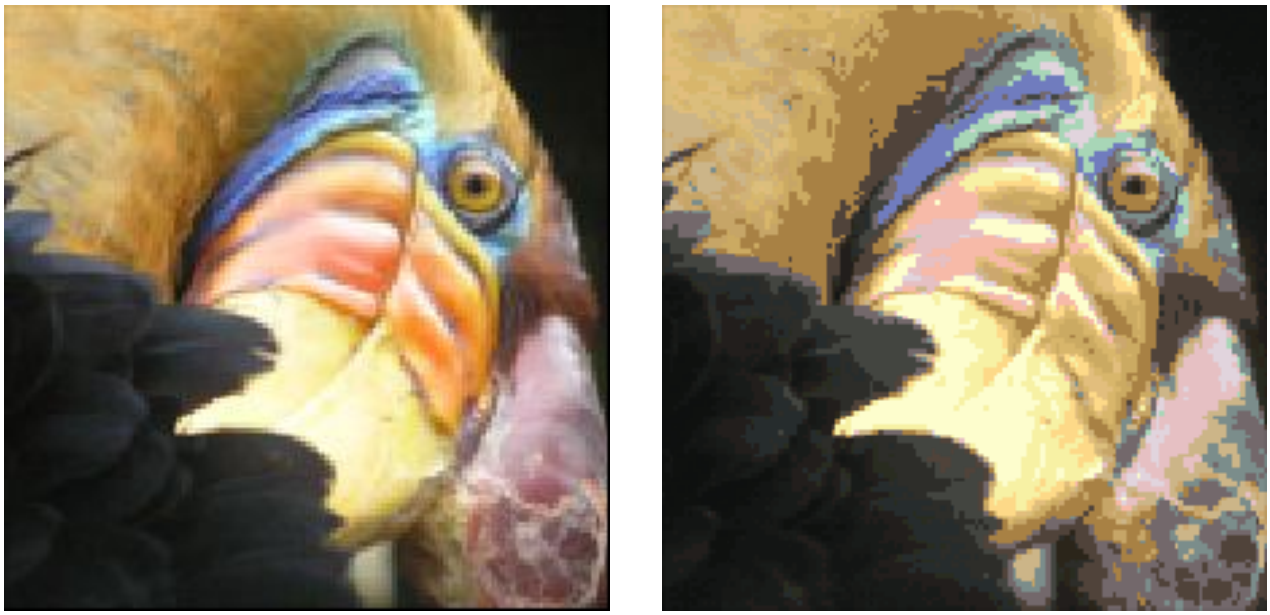

Рис. 2. Исходное изображение (слева) и сжатое изображение с использованием алгоритма K-средних (справа). Число средних было выбрано равным 32

3. Вопросы для составления отчета

1. Зададим значения трех средних в векторном виде: первое среднее – $[3, 3]^T$, второе среднее – $[6, 2]^T$, третье среднее – $[8, 5]^T$. К каким ближайшим средним относятся первые три примера из матрицы объекты-признаки, которая содержится в файле **data.mat** (**50 баллов**)?

2. Для заданных в вопросе 1 векторов средних определите принадлежность всех элементов в матрице объекты-признаки к соответствующим средним. Выполните пересчет новых средних значений с учетом принадлежности элементов к соответствующим кластерам. Чему равны эти средние (**50 баллов**)?