Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Кафедра Штучного інтелекту

Звіт з лабораторної роботи №3

з курсу Інтелектуальний аналіз даних

по темі “КОМПРЕСІЯ (СТИСК) ІНФОРМАЦІЇ МЕТОДОМ

ГОЛОВНИХ КОМПОНЕНТ (РСА) І НА ОСНОВІ СИСТЕМИ Е. ОЯ.”

Виконали:

студенти гр. ІТШI 17-1

Гура А. О.

Перевірила:

доц. Дейнеко А.О.

Харків 2019

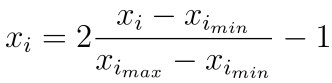
1. Цель работы: Приобретение навыков реализации компрессии данных методом PCA.

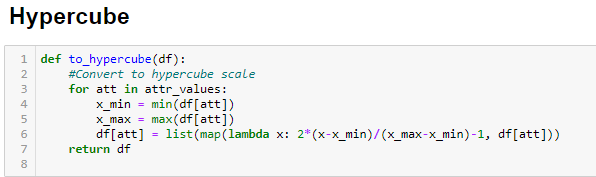
2. Ход работы:

2.1: Предварительная подготовка данных

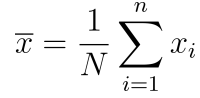
2.1.1. Массив исходных данных разделить на атрибуты. Каждый атрибут кодировать на гиперкуб [-1; 1]:





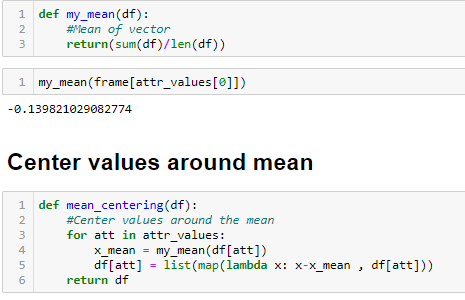


2.1.2. Найти среднее значение



2.1.3. Отцентрировать все значения относительно среднего:

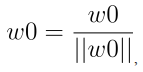




2.1.4. Снова собрать атрибуты в массив X.

2.2. Вычисление:

Необходимо вычислить *c* векторов *w* (матрица *W*) и *с* компонент *y* (матрица *Y*). Векторы и компоненты вычисляются рекуррентно, то есть на основе предыдущих значений. «Стартовый» вектор *w0* весовых коэффициентов состоит из *с* случайных чисел в диапазоне [-1, 1]. Стартовый вектор *w0* нужно нормировать:

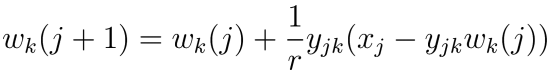


Компонента *у* состоит из *r* скаляров. Каждый следующий скаляр вычисляется из вектора весовых коэффициентов *w* и значений *х* по формуле:

,

где j = 1:r, k = (1,c), x(j) - образец из массива данных (строка длины *с*).

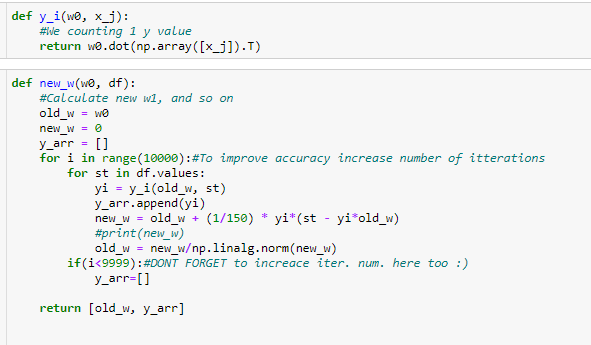
Обратите внимание, что в данной формуле используется матричное умножение. Поэтому необходимо, чтобы w был представлен строкой, а x(j) - столбцом либо вектором(векторное умножение). Тогда результатом умножения будет число. На основе полученного y(j,k) и предыдущего значения w(j) вычислим wk (j+1) по правилу Оя:



После каждого пересчета вектор нормируем. Обратите внимание: w(k) пересчитывается r раз, но нам необходимо только последнее значение - оно является k-тым вектором, остальные хранить не нужно. При этом компонента у содержит все r значений. Для стабилизации результата компоненту y(k) и вектор w(k) следует вычислить 10^k раз. Далее после того как будет получен каждый следующий собственный вектор и главные компоненты необходимо производить процедуру вычитания главных компонент и собственных векторов из выборки:

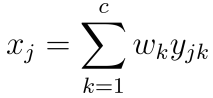


Реалзиация:



2.3. Восстановление данных:

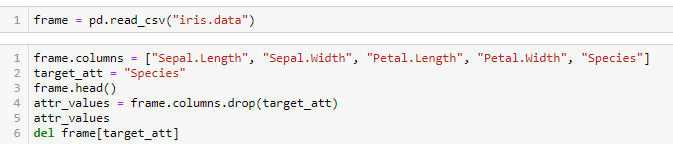
Для проверки правильности сжатия (компресии) необходимо воспользоваться соотношением вида:

,

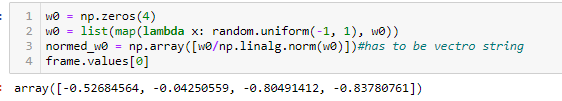
где x(j) - строка исходного массива данных.

Также в качестве проверки можно сравнить полученный результат с результатом встроенного метода PCA. Сравнивать стоит по матрице W. При этом знаки могут не совпадать.

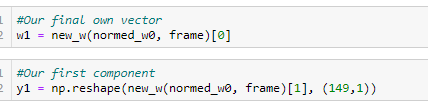
3. Вызов данных методов для выборки Ирисы Фишера:



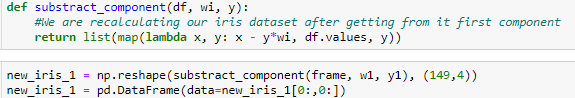
Создание первого вектора со случайными числами в заданном диапазоне:

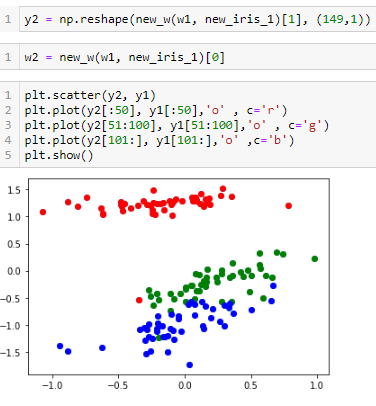


1я Главная компонента и первый собственный вектор:

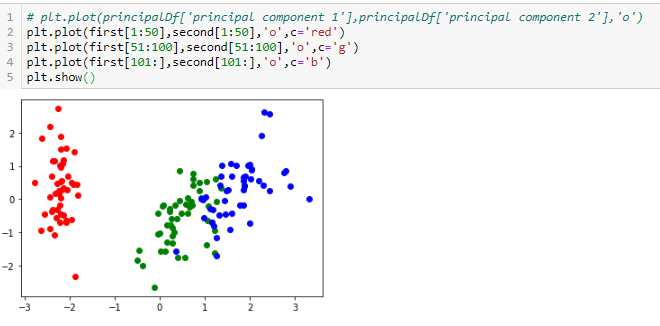


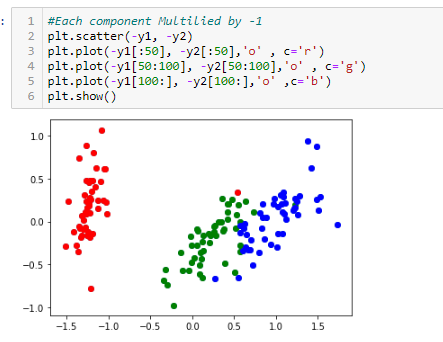
Функция для вычитания компоненты:



Вторая компонента и график зависимости 1й и 2й: 

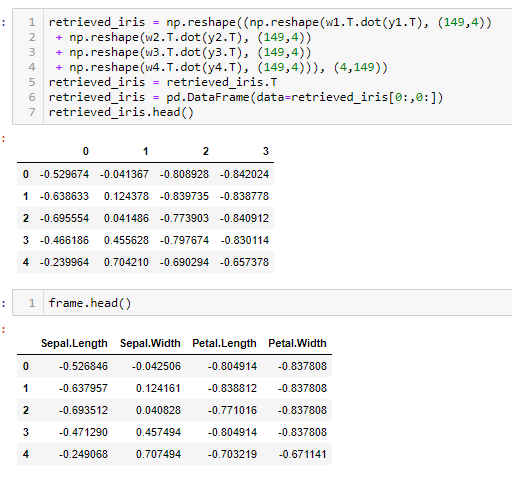
А теперь сравним с встроенным PCA:



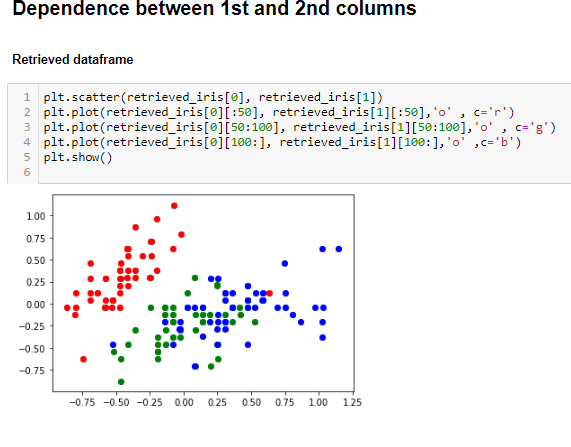
А теперь помножим на -1 наши 1ю и 2ю главные компоненты: 

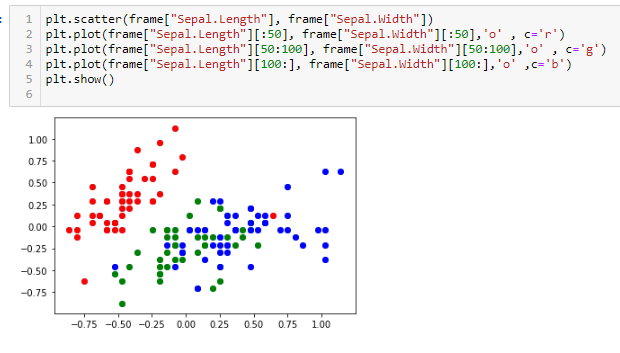
Теперь если сравнить со встроенным методом то практически идентично, не считая масштаба.

Восстановление датасета, и сравнение его с изначальным ирисом:



Сравнение плота 1го и 2го атрибута для восстановленной выборки, и для оригинальной:





4. Выводы: В ходе данной лабораторной работы был освоен один из основных методов компрессии данных PCA. Было реализовано правило Оя и применена компрессия для выборки Ирисы Фишера.