МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №2

по дисциплине «Машинное обучение»

Тема: Понижение размерности пространства признаков

Студент гр. 6304 _____ Ястребков А. С.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург
2020

Цель работы:

Ознакомиться с методами понижения размерности данных из библиотеки Scikit Learn.

Ход работы

Загрузка данных. Загружен требуемый датасет, на рис. 1 приведён фрагмент исходных данных.

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
0	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.0	0.0	1
1	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.0	0.0	1
2	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.0	0.0	1
3	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.0	0.0	1
4	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.0	0.0	1

Рис. 1. Фрагмент исходных данных.

Числовые данные были отеделены от меток класса (колонка Type) и масштабированы к интервалу [0, 1]. Для полученного набора были построены диаграммы рассеяния, приведённые на рис. 2.

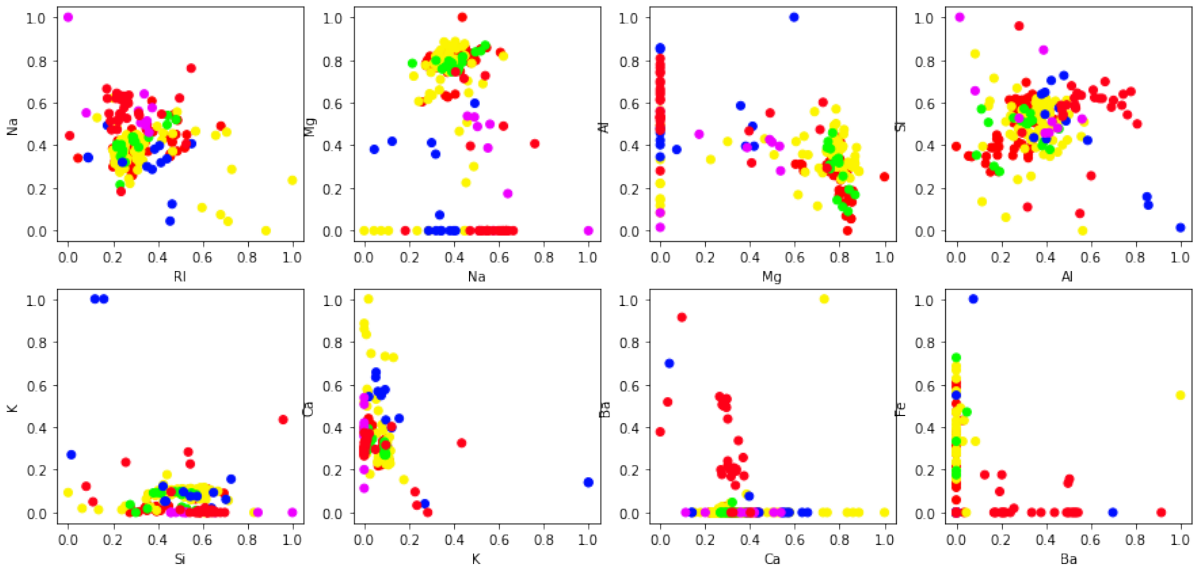

Рис. 2. Диаграммы рассеяния для исходных пар признаков.

Метод галвных компонент

С помощью метода главных компонент (PCA) было проведено понижение размерности пространства признаков до 2, диаграмма рассеяния приведена на

рис. 3. Из свойств объекта PCA были получены значения объяснённой дисперсии и собственные числа, соответствующие компонентам:

- Объяснённая дисперсия: [0.454, 0.1799];
- Собственные числа: [5.105, 3.212].

Значения объяснённой дисперсии показывают, что для двух компонент метод главных компонент объясняет не более 63% общей дисперсии. Диаграмма рассеяния не показывает явной линейной связи признаков, так что применение метода для используемого набора данных нецелесообразно.

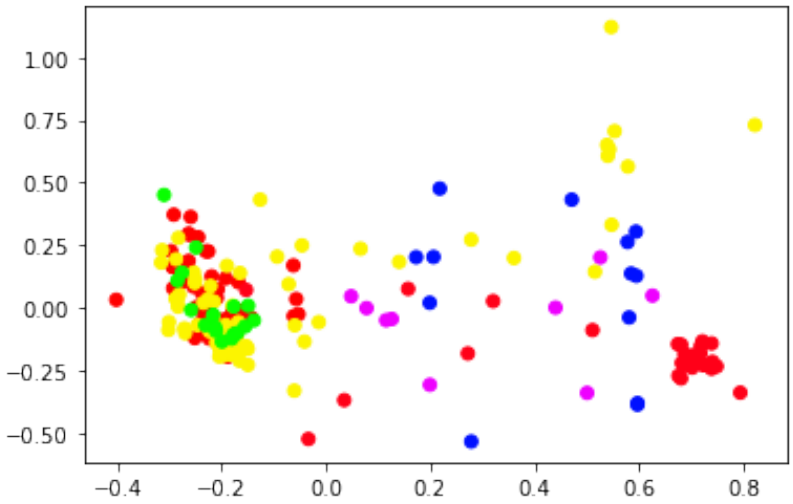

Рис. 3. Диаграмма рассеяния после использования PCA с двумя компонентами.

Было исследовано поведение метода главных компонент для различного количества компонент, график объяснённой дисперсии приведён на рис. 4. Видно, что при количестве компонент от 4 и больше метод объясняет более 85% исходной дисперсии.

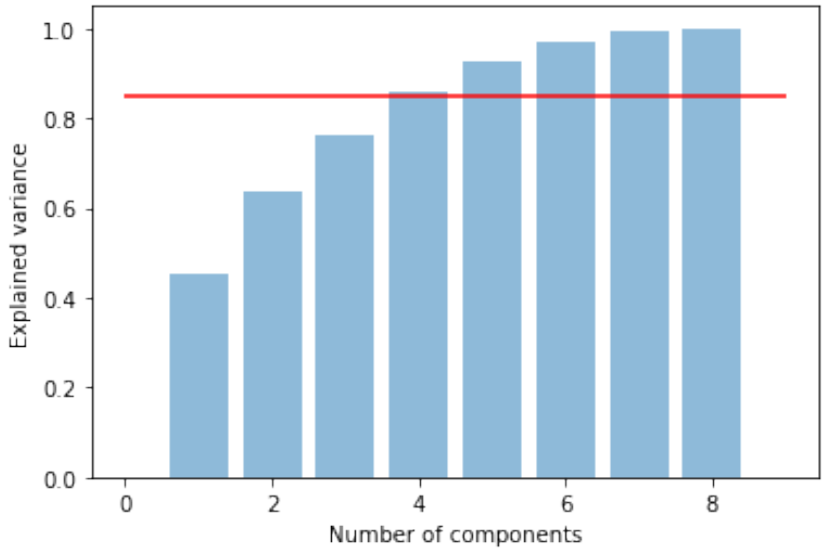

Рис. 4. Объяснённая дисперсия для разного количества компонент.

С помощью метода inverse_transform было проведено обратное преобразование данных. В таблицу 1 сведены средние значения и среднеквадратичные отклонения исходных данных и полученных обратным преобразованием. Видно, что значение математического ожидания практически не меняется, при этом значения СКО искажаются достаточно сильно (что соотносится с предыдущими выводами о малом значении объяснённой дисперсии для 2-компонентного преобразования).

Таблица 1. Матожидание и СКО исходных и восстановленных признаков.

	Среднее значение		Среднеквадратичное отклонение		
	исходное	восстановленное	исходное	восстановленное	потеря, %
RI	0.317	0.317	0.133	0.113	14.8
Na	0.403	0.403	0.123	0.058	52.5
Mg	0.598	0.598	0.320	0.318	0.7
Al	0.360	0.360	0.155	0.116	25.48
Si	0.507	0.507	0.138	0.051	63.24
K	0.080	0.080	0.105	0.028	72.89
Ca	0.328	0.328	0.132	0.121	8.47
Ba	0.056	0.056	0.157	0.100	36.59
Fe	0.112	0.112	0.191	0.106	44.57

Было исследовано поведение метода главных компонент для различных значений параметра svd_solver. Библиотека Sklearn предлагает три решателя: full, arpack, randomized. Диаграммы рассеяния приведены на рис. 5. Визуальных различий между разными решателями не заметно. Скорее всего, для явных различий необходима выборка большего объёма.

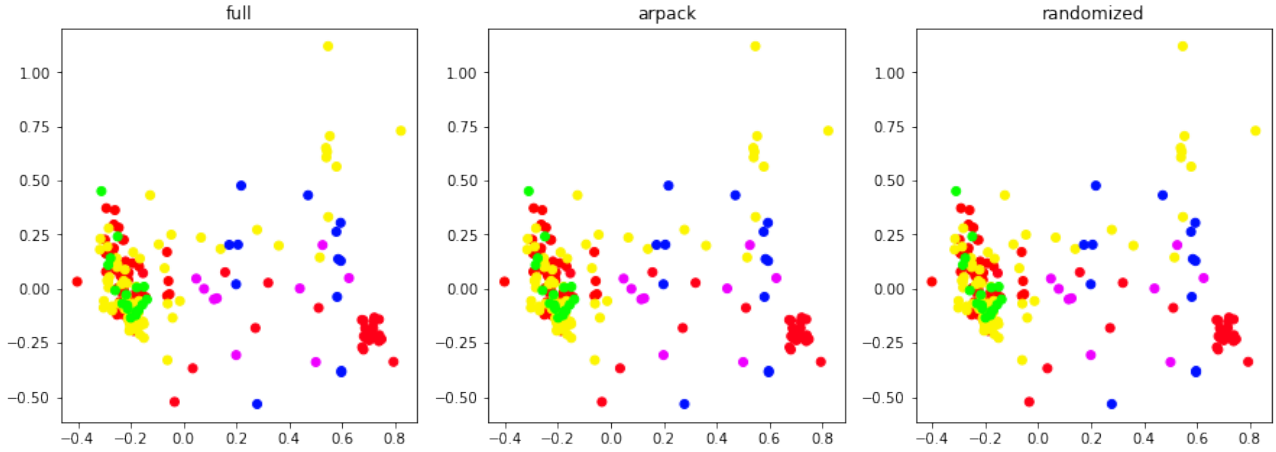

Рис. 5. Диаграммы рассеяния для различных значений svd_solver.

Модицикации метода главных компонент

KernelPCA предлагает модификацию метода главных компонент с использованием различных ядерных фукнций:

- линейное ядро: $k(x,y)=x^T y+c$
- полиномиальное ядро: $k(x,y)=(\alpha x^T y+c)^d$
- радиально-базисная функция (RBF): $k(x,y)=\exp(-\gamma\|x-y\|^2)$
- сигмоида: $\tanh(\alpha x^T y+c)$
- косинус: $k(x,y)=\frac{x y^T}{\|x\|\|y\|}$
- precompiled — ядро, заданное пользователем.

KernelPCA ведёт себя аналогично линейному PCA при использовании линейного ядра (значение 'linear' параметра kernel). На рис. 6 показаны диаграммы рассеяния при использовании различных ядер. Размер выборки не позволяет установить визуальные различия между ними.

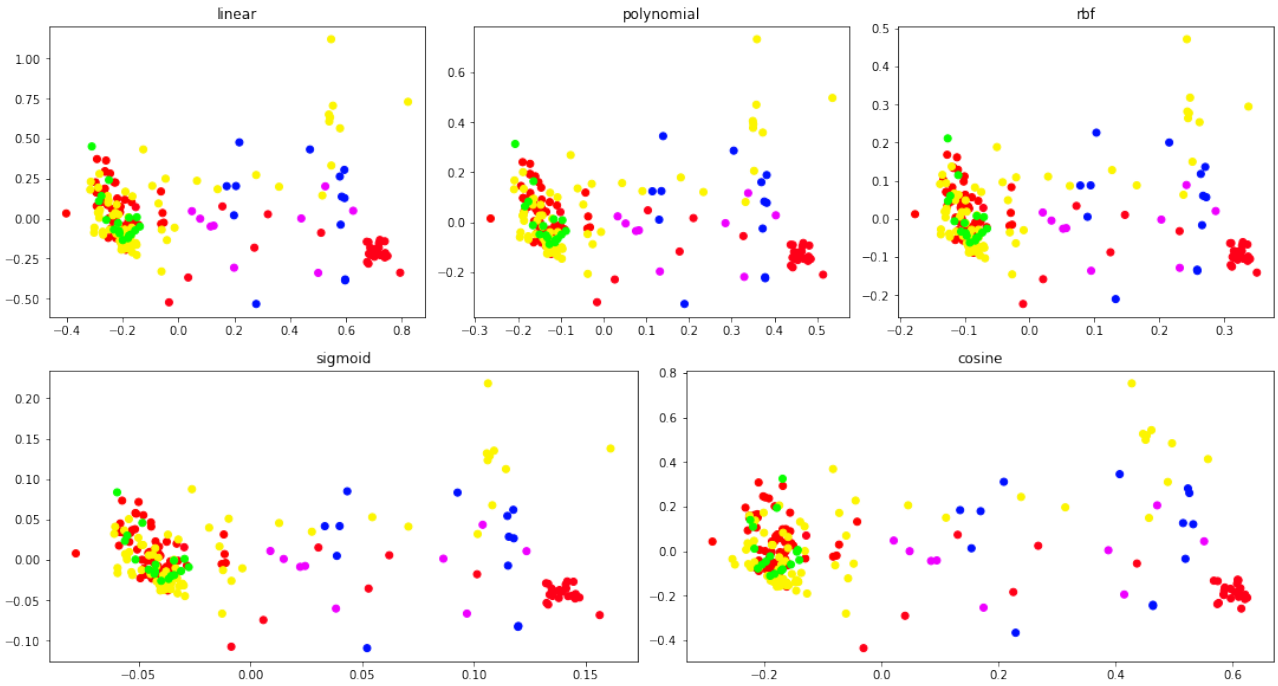

Рис. 6. Диаграммы рассеяния для различных ядер.

В таблицу 2 сведены данные по значениям атрибута lamdas_, описывающего собственные числа, и объяснённой дисперсии при разном количестве ядер для количества компонентов 4. Как видно из таблицы, для всех ядер объяснённая дисперсия различается незначительно.

Таблица 2. Собственные числа и объяснённая дисперсия для различных ядер KernelPCA.

ядро	lambdas_[0]	lambdas_[1]	lambdas_[2]	lambdas_[3]	об. дисп.
linear	26.06	10.32	7.26	5.62	0.857
polynomial	10.92	4.32	3.12	2.37	0.849
RBF	5.35	2.02	1.47	1.11	0.850
sigmoid	1.01	0.4	0.27	0.22	0.861
cosine	18.31	6.48	4.97	3.58	0.86

SparsePCA находит набор разреженных компонент, по которым можно наилучшим образом восстановить исходные данные. Разреженность можно контролировать параметрами объекта. Для сравнения были использованы два способа вычисления компонент, наименьшая угловая регрессия и координатный

спуск. При одинаковых параметрах изменение способа вычисления не даёт видимых результатов (рис. 7). Изменение же параметра, регулирующего разреженность данных, заметно, что показано на рис. 8. В таблице 3 приведены главные компоненты для различных значений alpha.

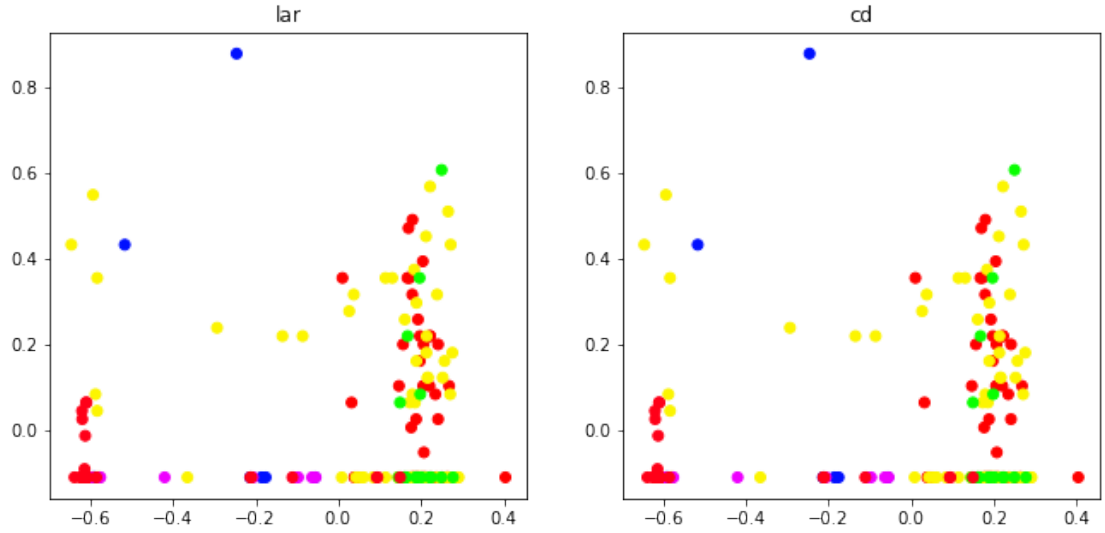

Рис. 7. Сравнение различных методов для SparsePCA.

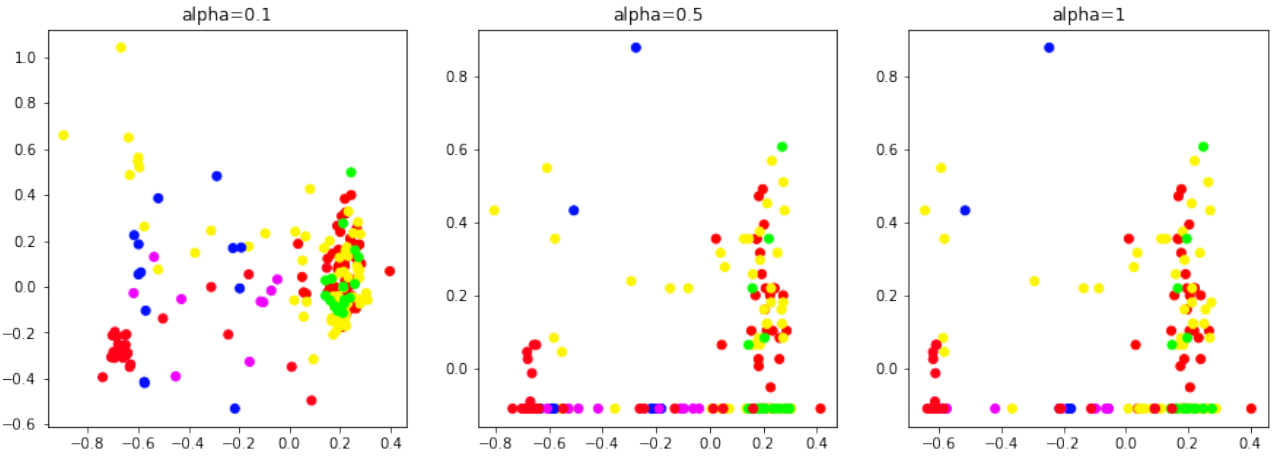

Рис. 8. Сравнение работы SparsePCA при различных коэффициентах alpha.

Таблица 3. Главные компоненты SparsePCA при различных alpha.

α=0.1	PC1	-0.067	-0.071	0.928	-0.202	-0.010	0	-0.176	-0.238	0
	PC2	0.505	-0.199	0	-0.365	-0.212	-0.095	0.476	-0.169	0.513
α=0.5	PC1	0	-0.001	0.965	-0.173	0	0	-0.061	-0.186	0
	PC2	0	0	0	0	0	0	0	0	1
α=1	PC1	0	0	0.998	-0.037	0	0	0	-0.050	0
	PC2	0	0	0	0	0	0	0	0	1

Факторный анализ

Было выполнено понижение размерности данных в помощью факторного анализа. Сравнение диаграмм рассеяния приведено на рис. 9. В отличие от PCA, факторный анализ объясняет не линейную дисперсию, а ковариацию данных, следовательно, и используется для поиска скрытых переменных (факторов). Факторный анализ не требует ортогональности (независимости) компонент, в отличие от PCA; компоненты PCA являются линейной комбинацией наблюдаемой переменной, факторный анализ представляет наблюдаемые переменные линейной комбинацией фактора.

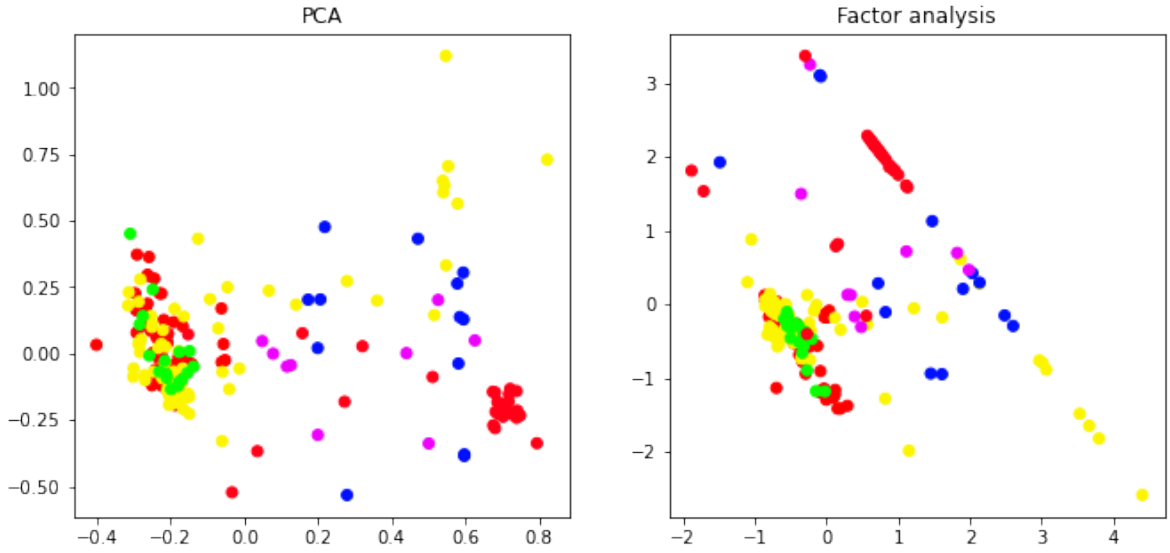

Рис. 9. Диаграммы рассеяния для PCA и факторного анализа.

Вывод:

В результате выполнения лабораторной работы были изучены методы понижения размерности пространства признаков, предлагаемые библиотекой Scikit-Learn.

Для метода главных компонент (PCA) было установлено, что количество компонент напрямую влияет на процент объяснённой дисперсии. KernelPCA может использоваться для поиска нелинейных зависимостей в данных, поскольку предлагает набор различных ядерных функций, однако на

имеющейся выборке смена ядра не сказалась. SparsePCA разумно использовать на данных большого объёма.

Были изучены отличия факторного анализа от метода главных компонент.