Data Mining in Action

Лекция 4

Преобразование признаков

Виктор Кантор

План

1. Задача понижения размерности
2. Метод главных компонент и SVD
3. Manifold learning

Как выглядит обучающая выборка

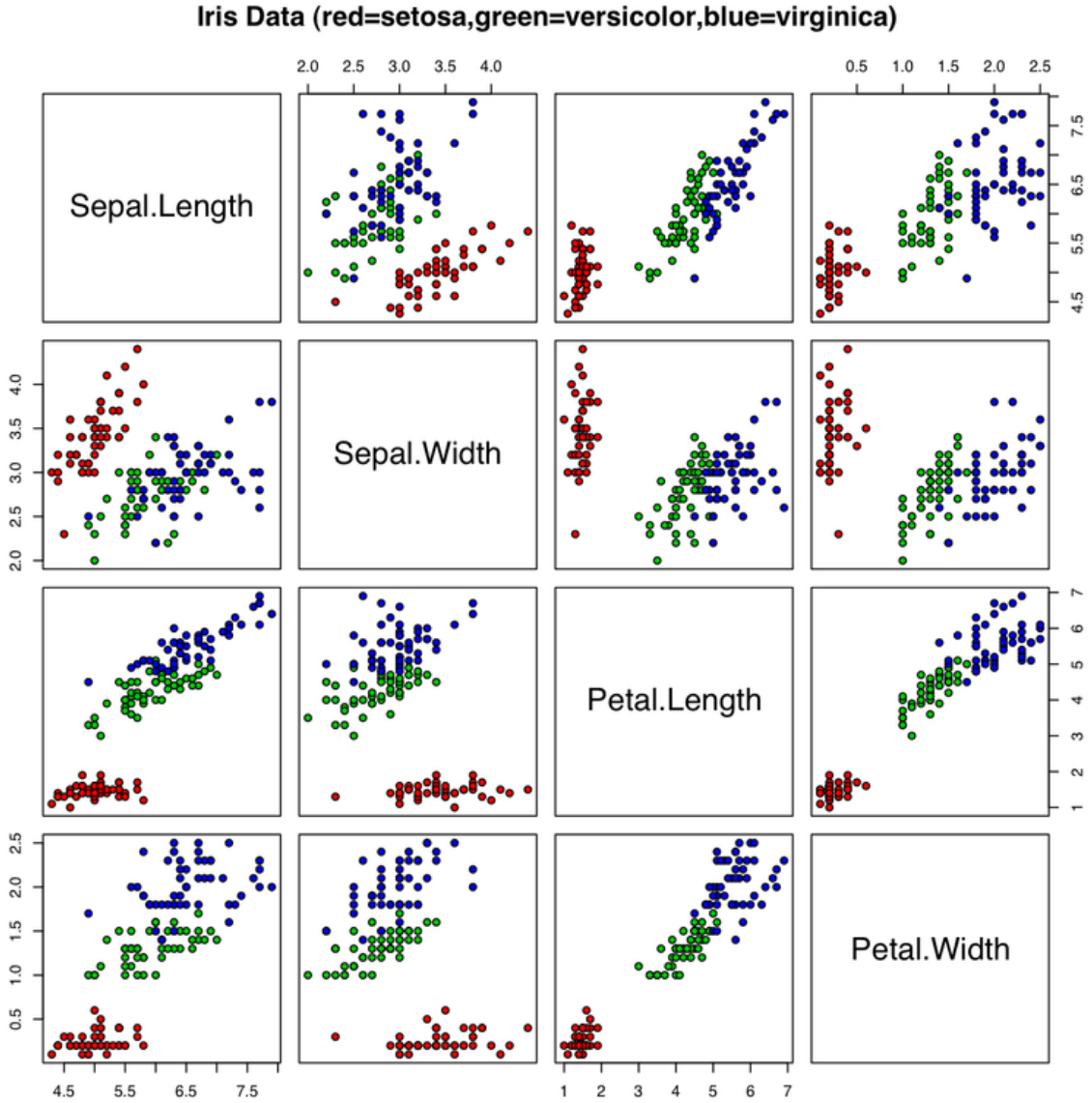
Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
5.7	2.5	5.0	2.0	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*
6.3	2.5	4.9	1.5	*I. versicolor*
4.9	2.5	4.5	1.7	*I. virginica*

Что хотелось бы уметь

- Визуализировать обучающую выборку, когда признаков больше трёх
- Уменьшать количество признаков, переходя к новым, более информативным

Визуализируем выборку

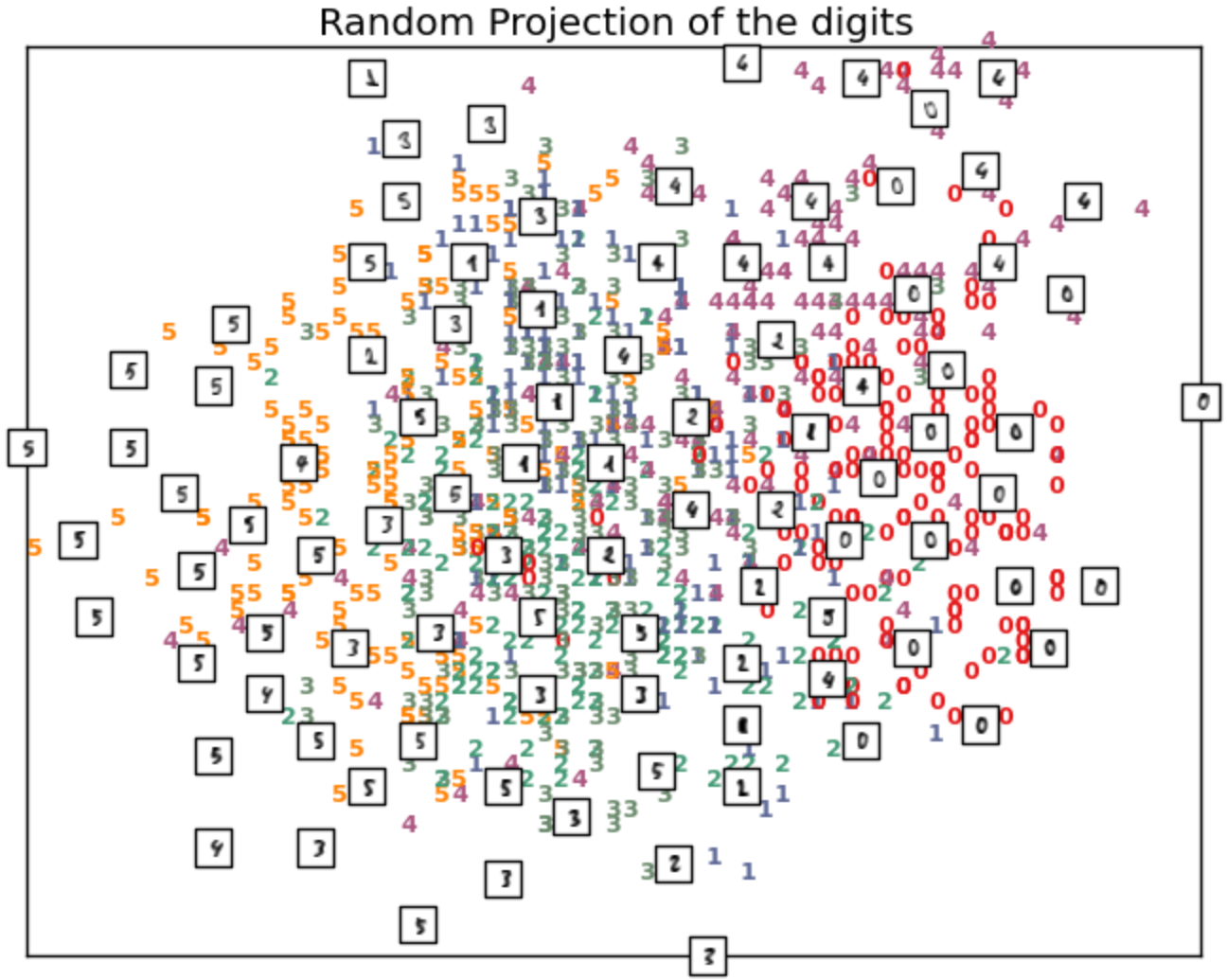


Более сложный случай

Что делать, если признаков еще больше?

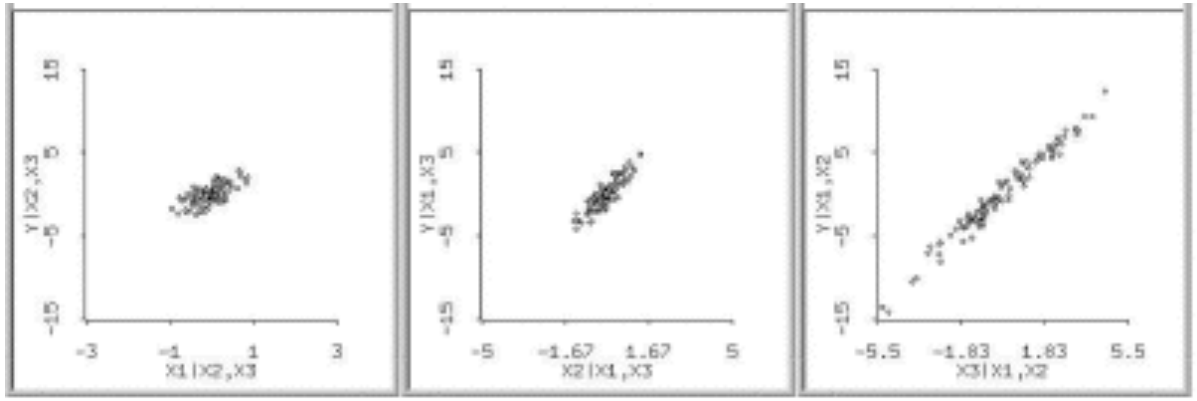
1	6	4	12	5	5	3	4	1	67	3	2	1	2	1	0	0	1	0	0	1	0	0	1	1
2	48	2	60	1	3	2	2	1	22	3	1	1	1	1	0	0	1	0	0	1	0	0	1	2
4	12	4	21	1	4	3	3	1	49	3	1	2	1	1	0	0	1	0	0	1	0	1	0	1
1	42	2	79	1	4	3	4	2	45	3	1	2	1	1	0	0	0	0	0	0	0	0	1	1
1	24	3	49	1	3	3	4	4	53	3	2	2	1	1	1	0	1	0	0	0	0	0	1	2
4	36	2	91	5	3	3	4	4	35	3	1	2	2	1	0	0	1	0	0	0	0	1	0	1
4	24	2	28	3	5	3	4	2	53	3	1	1	1	1	0	0	1	0	0	1	0	0	1	1
2	36	2	69	1	3	3	2	3	35	3	1	1	2	1	0	1	1	0	1	0	0	0	0	1
4	12	2	31	4	4	1	4	1	61	3	1	1	1	1	0	0	1	0	0	1	0	1	0	1
2	30	4	52	1	1	4	2	3	28	3	2	1	1	1	1	0	1	0	0	1	0	0	0	2
2	12	2	13	1	2	2	1	3	25	3	1	1	1	1	1	0	1	0	1	0	0	0	1	2
1	48	2	43	1	2	2	4	2	24	3	1	1	1	1	0	0	1	0	1	0	0	0	1	2
2	12	2	16	1	3	2	1	3	22	3	1	1	2	1	0	0	1	0	0	1	0	0	1	1
1	24	4	12	1	5	3	4	3	60	3	2	1	1	1	1	0	1	0	0	1	0	1	0	2
1	15	2	14	1	3	2	4	3	28	3	1	1	1	1	1	0	1	0	1	0	0	0	1	1
1	24	2	13	2	3	2	2	3	32	3	1	1	1	1	0	0	1	0	0	1	0	1	0	2
4	24	4	24	5	5	3	4	2	53	3	2	1	1	1	0	0	1	0	0	1	0	0	1	1
1	30	0	81	5	2	3	3	3	25	1	3	1	1	1	0	0	1	0	0	1	0	0	1	1
2	24	2	126	1	5	2	2	4	44	3	1	1	2	1	0	1	1	0	0	0	0	0	0	2
4	24	2	34	3	5	3	2	3	31	3	1	2	2	1	0	0	1	0	0	1	0	0	1	1
4	9	4	21	1	3	3	4	3	48	3	3	1	2	1	1	0	1	0	0	1	0	0	1	1
1	6	2	26	3	3	3	3	1	44	3	1	2	1	1	0	0	1	0	1	0	0	0	1	1
1	10	4	22	1	2	3	3	1	48	3	2	2	1	2	1	0	1	0	1	0	0	1	0	1
2	12	4	18	2	2	3	4	2	44	3	1	1	1	1	0	1	1	0	0	1	0	0	1	1
4	10	4	21	5	3	4	1	3	26	3	2	1	1	2	0	0	1	0	0	1	0	0	1	1
1	6	2	14	1	3	3	2	1	36	1	1	1	2	1	0	0	1	0	0	1	0	1	0	1
4	6	0	4	1	5	4	4	3	39	3	1	1	1	1	0	0	1	0	0	1	0	1	0	1
3	12	1	4	4	3	2	3	1	42	3	2	1	1	1	0	0	1	0	1	0	0	0	1	1
2	7	2	24	1	3	3	2	1	34	3	1	1	1	1	0	0	0	0	0	1	0	0	1	1
1	60	3	68	1	5	3	4	4	63	3	2	1	2	1	0	0	1	0	0	1	0	0	1	2
2	18	2	19	4	2	4	3	1	36	1	1	1	2	1	0	0	1	0	0	1	0	0	1	1
1	24	2	40	1	3	3	2	3	27	2	1	1	1	1	0	0	1	0	0	1	0	0	1	1

Пример случайной проекции для рукописных цифр



Проблемы «лишних признаков»

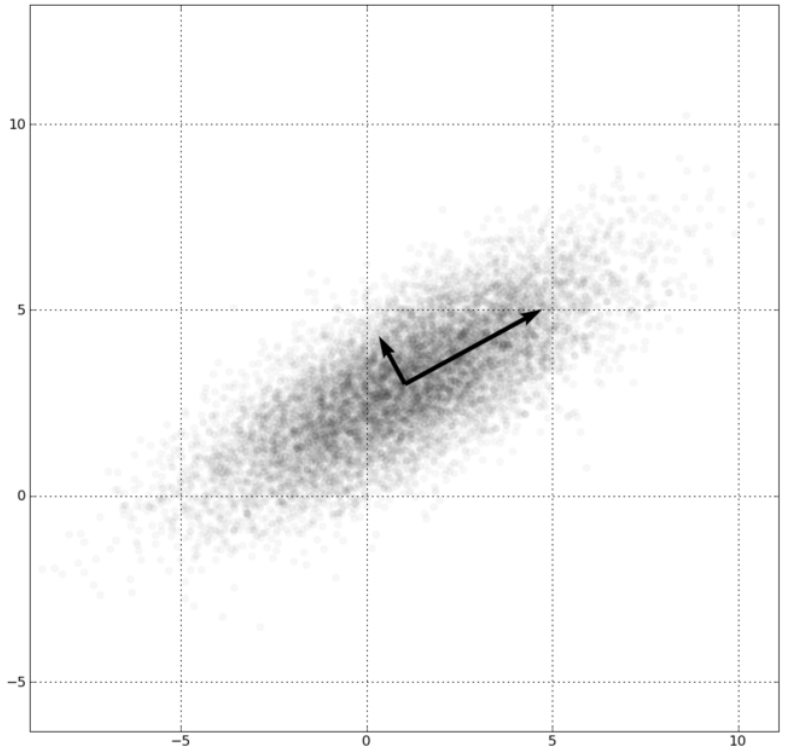
Если признаки сильно коррелированы, то у многих методов машинного обучения будут проблемы (например, из-за неустойчивости обращения матрицы ковариаций, где это нужно)

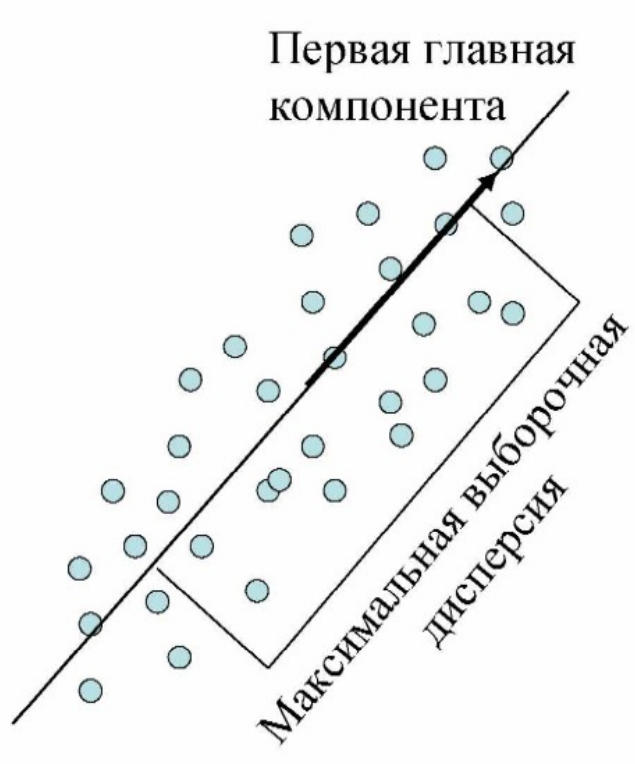


Principal Component Analysis 1

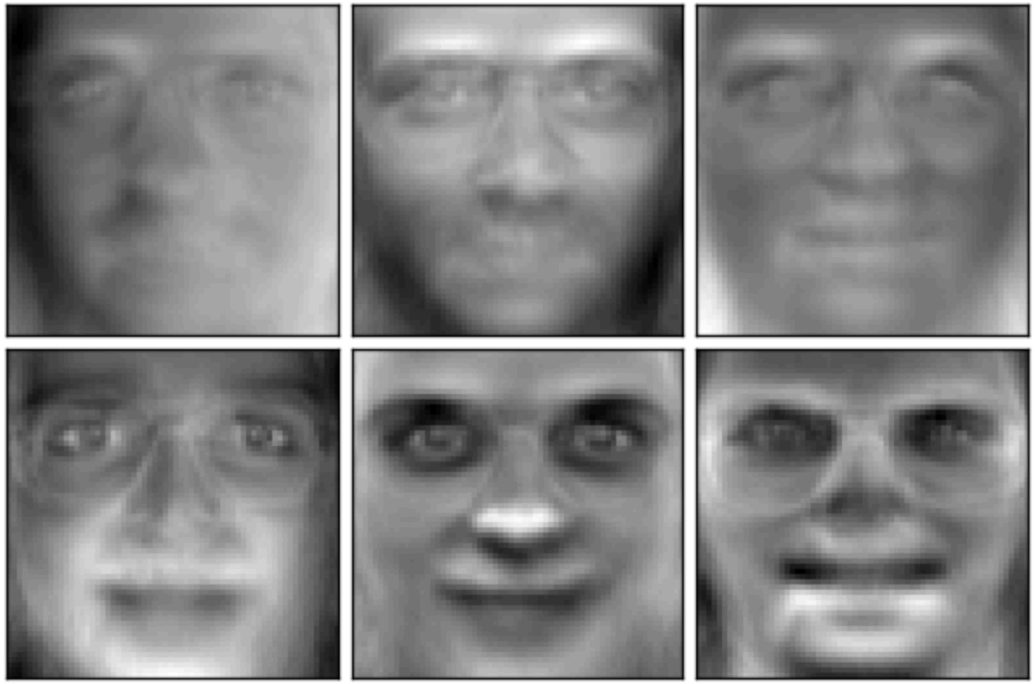
- Идея 1: давайте выделять в пространстве признаков направления, вдоль которых разброс точек наибольший (они кажутся наиболее информативными)

PCA (интерпретация 1)

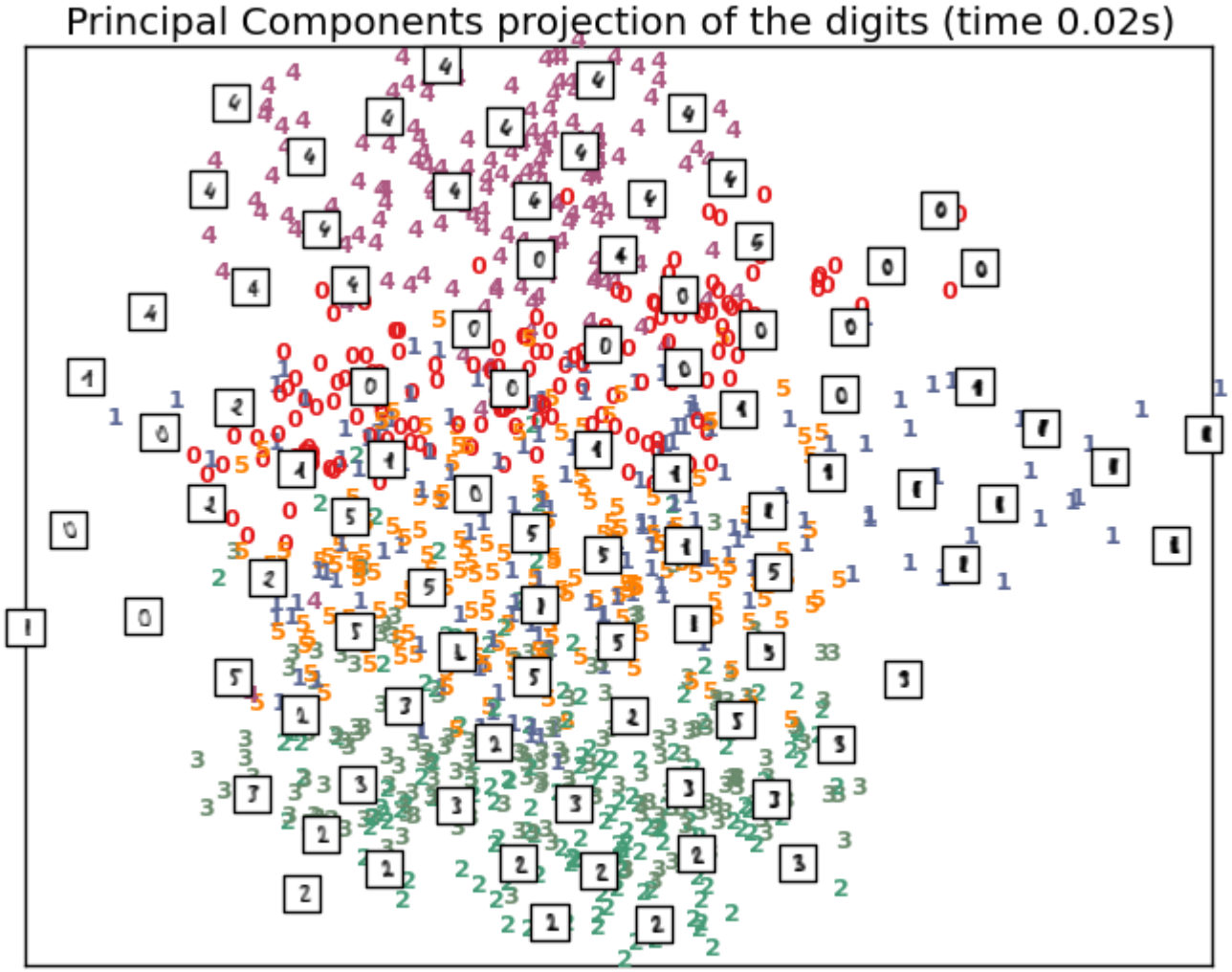




Пример: eigenfaces



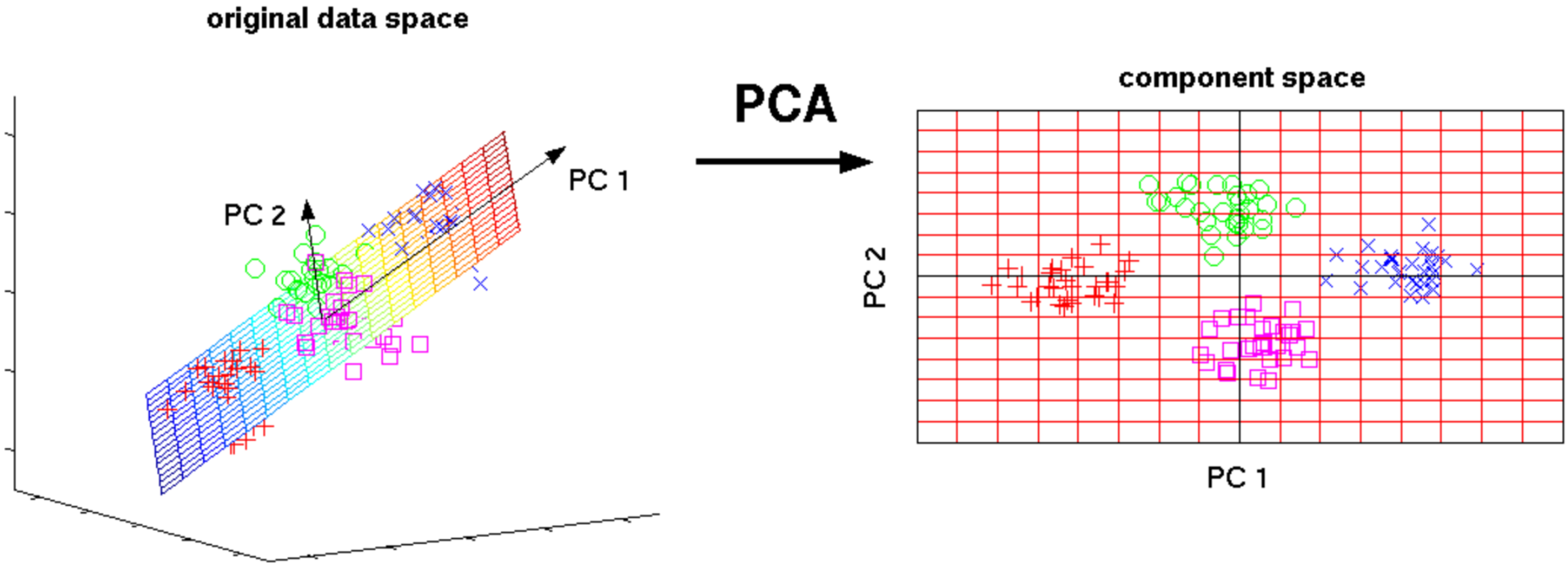
Рукописные цифры: проекция на главные компоненты



PCA (интерпретация 2)

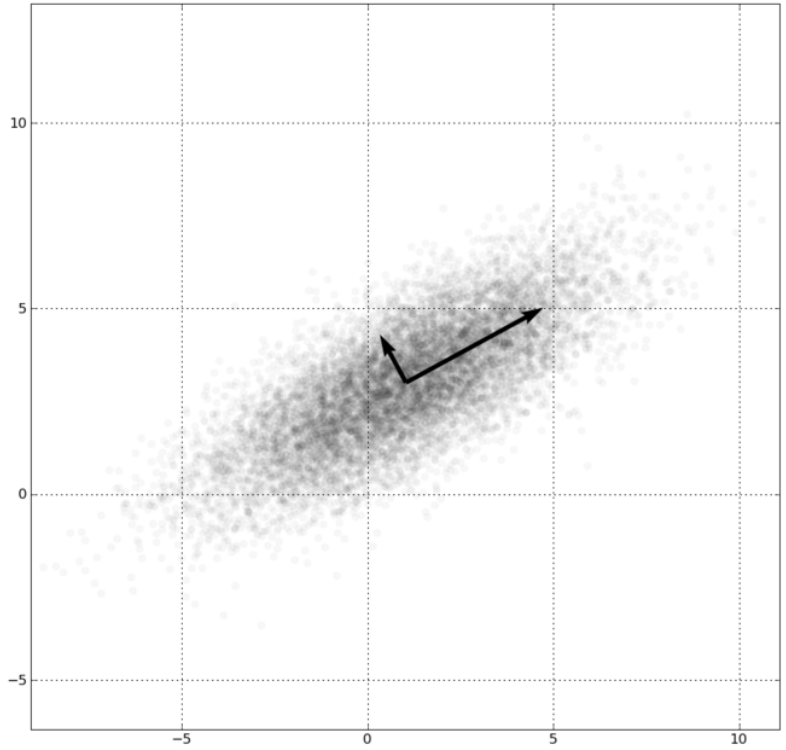
- Вторая идея: давайте строить проекцию выборки на линейное подпространство меньшей размерности. А выбирать его так, чтобы квадраты отклонений точек от проекций были минимальны.

PCA (интерпретация 2)



PCA (интерпретация 3)

Пусть точки получены из многомерного нормального распределения. Перейдем в такой базис, в котором матрица ковариаций станет диагональной. И оставим те направления, для которых больше дисперсия.



PCA (интерпретация 4)

Приблизим исходную матрицу признаков произведением двух матриц:

$$\underset{l\times n}{X} \approx \underset{l\times k}{U} \cdot \underset{k\times n}{V^T}$$

$$\left|\left|X - U \cdot V^T\right|\right| \to min$$

PCA: как сделать?

- Центрируем выборку (из каждого признака вычитаем среднее значение), получаем матрицу X с новыми значениями признаков
- Делаем SVD-разложение матрицы X:

$$X \approx A \cdot \Lambda \cdot B^T$$

Выбираем $U = A \cdot \Lambda, \quad V = B$

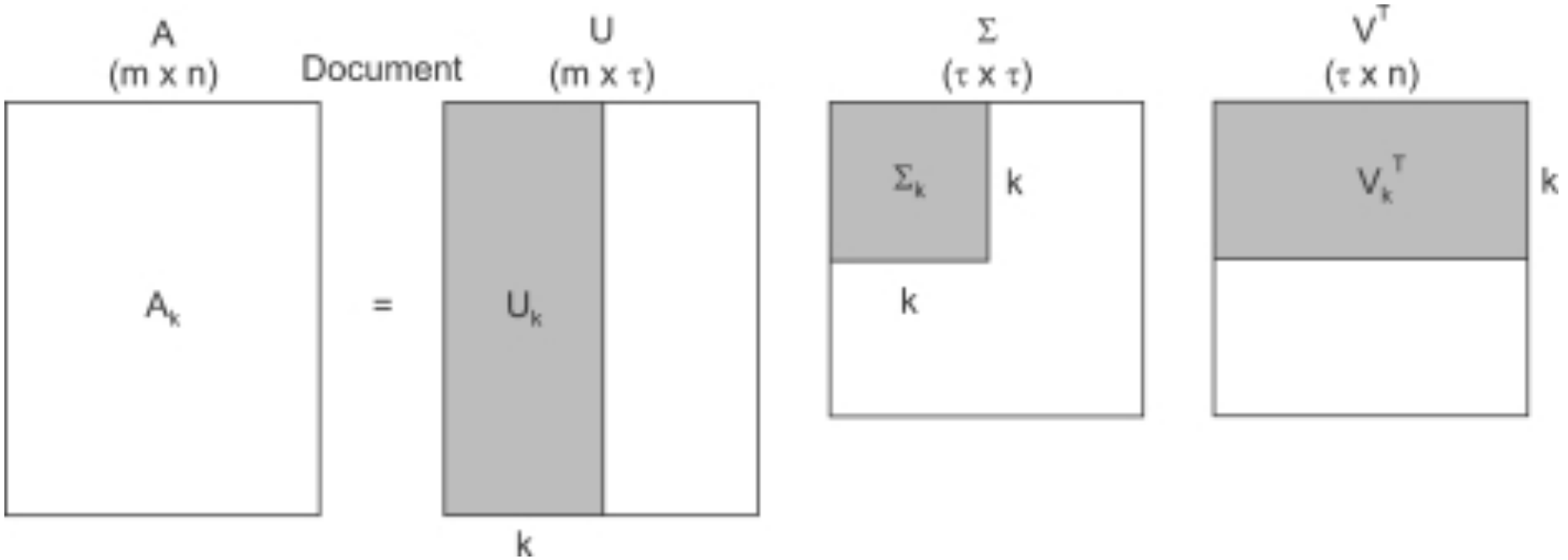
SVD

SVD = Singular Vector Decomposition (сингулярное разложение матриц)

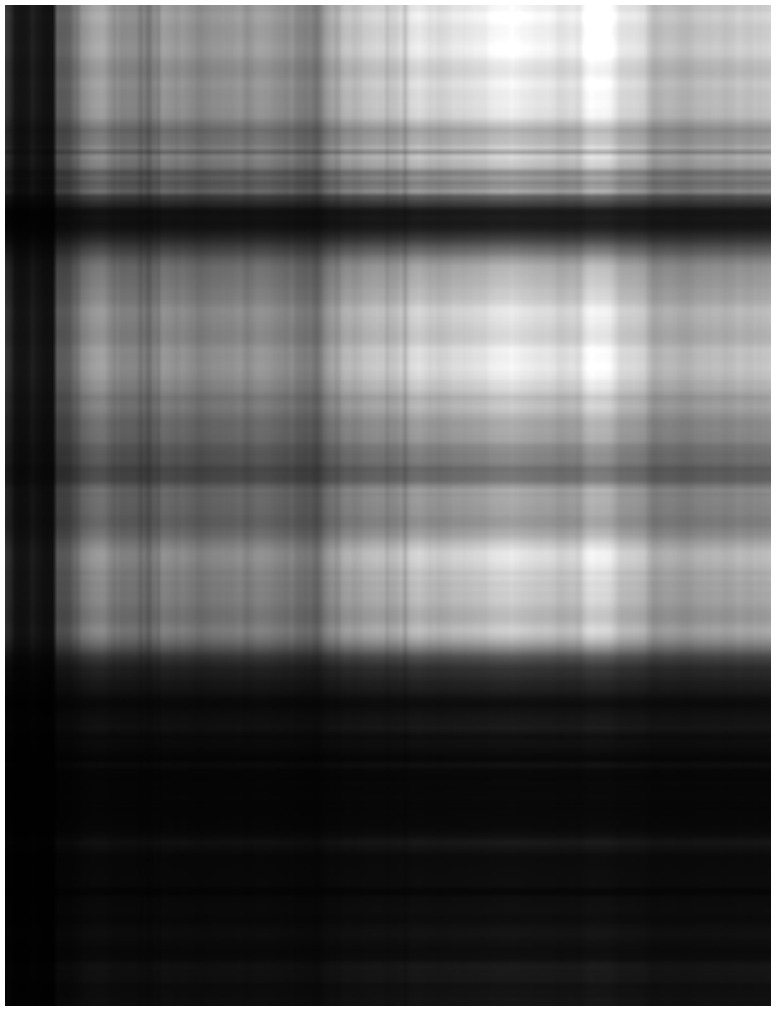
Позволяет получить наилучшее приближение исходной матрицы X матрицей X’ ранга k.

Применяется для снижения размерности пространства признаков.

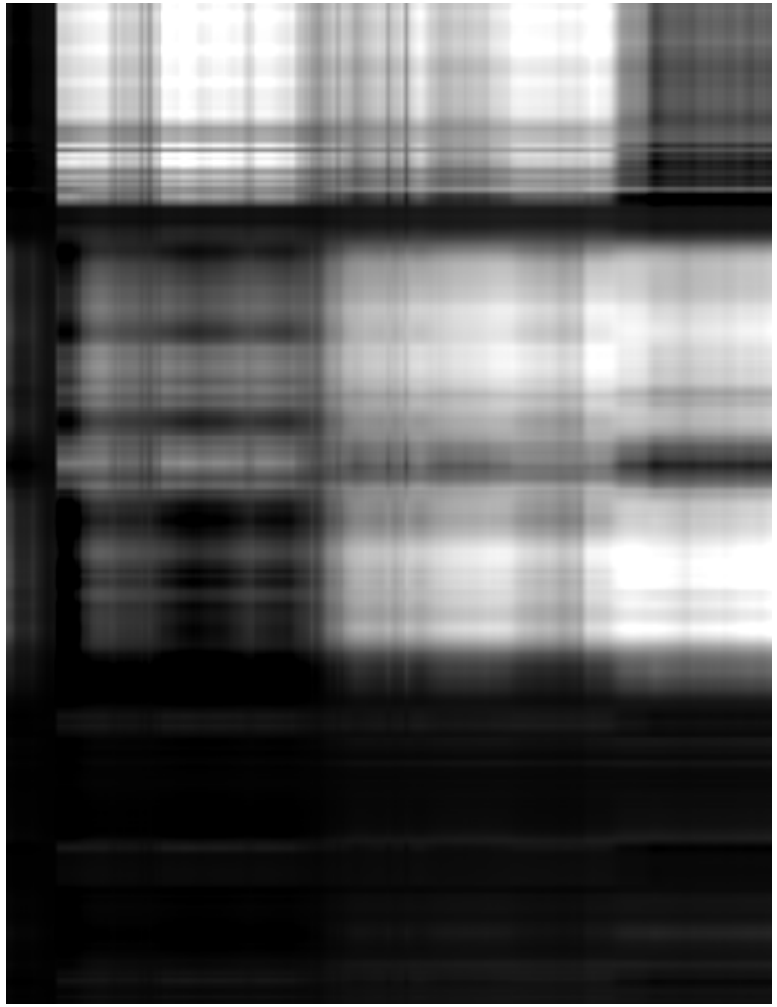
SVD



SVD: пример

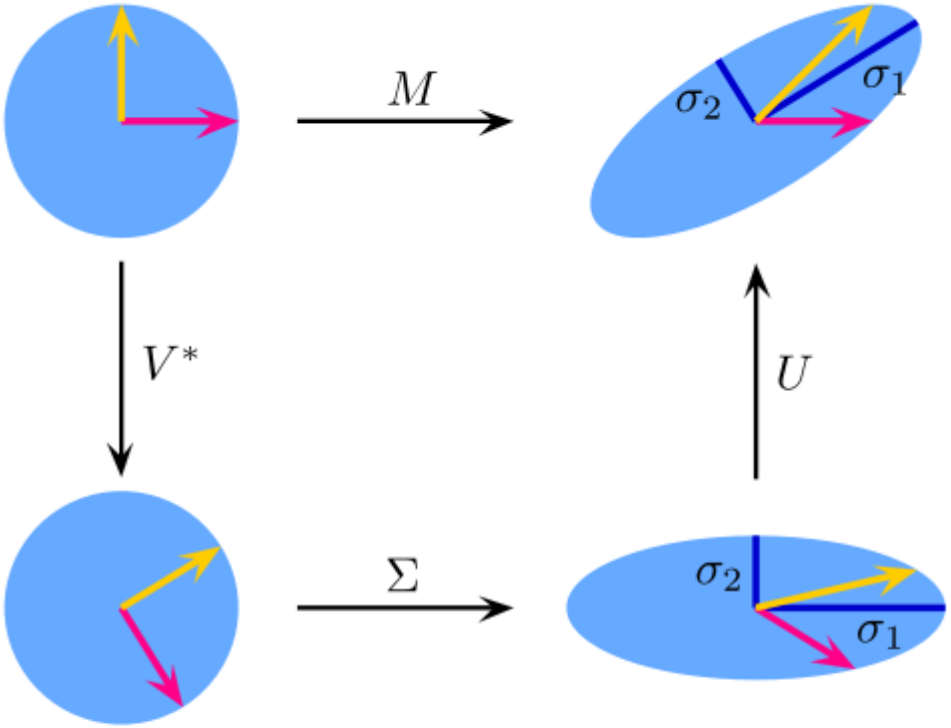


SVD: пример



SVD: пример

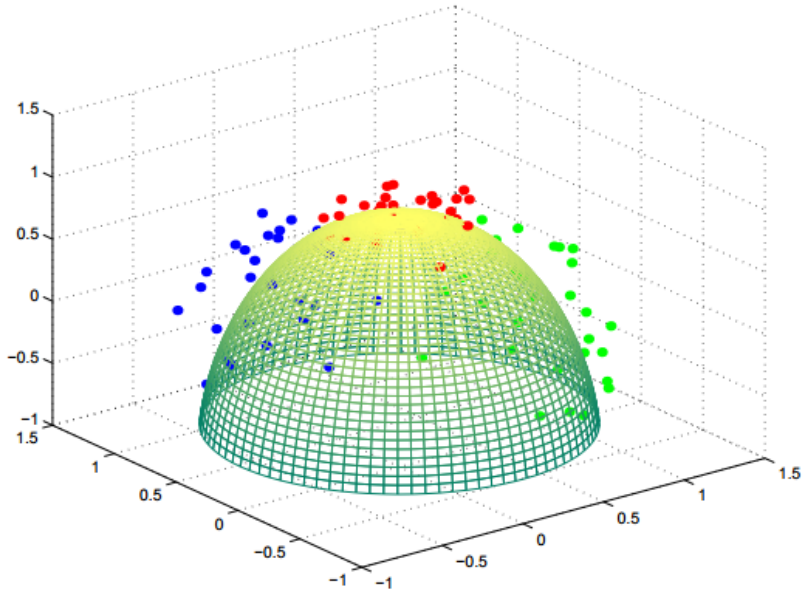
Геометрический смысл SVD



$$M = U \cdot \Sigma \cdot V^*$$

А что, если линейных преобразований признаков мало?

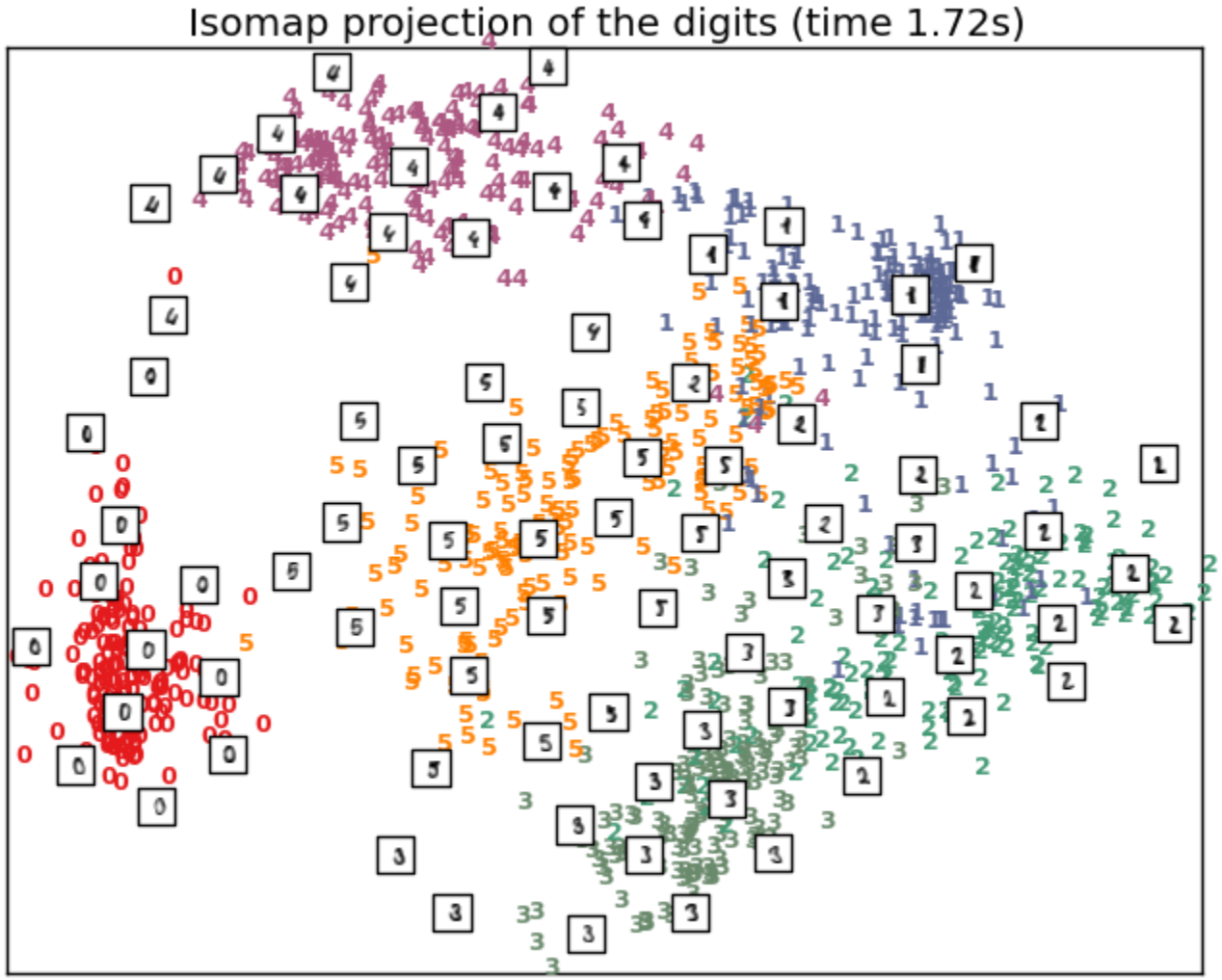
- Идея 1: объекты могут лежать в пространстве признаков на поверхности малой размерности.
- Идея 2: эта поверхность может быть нелинейной.



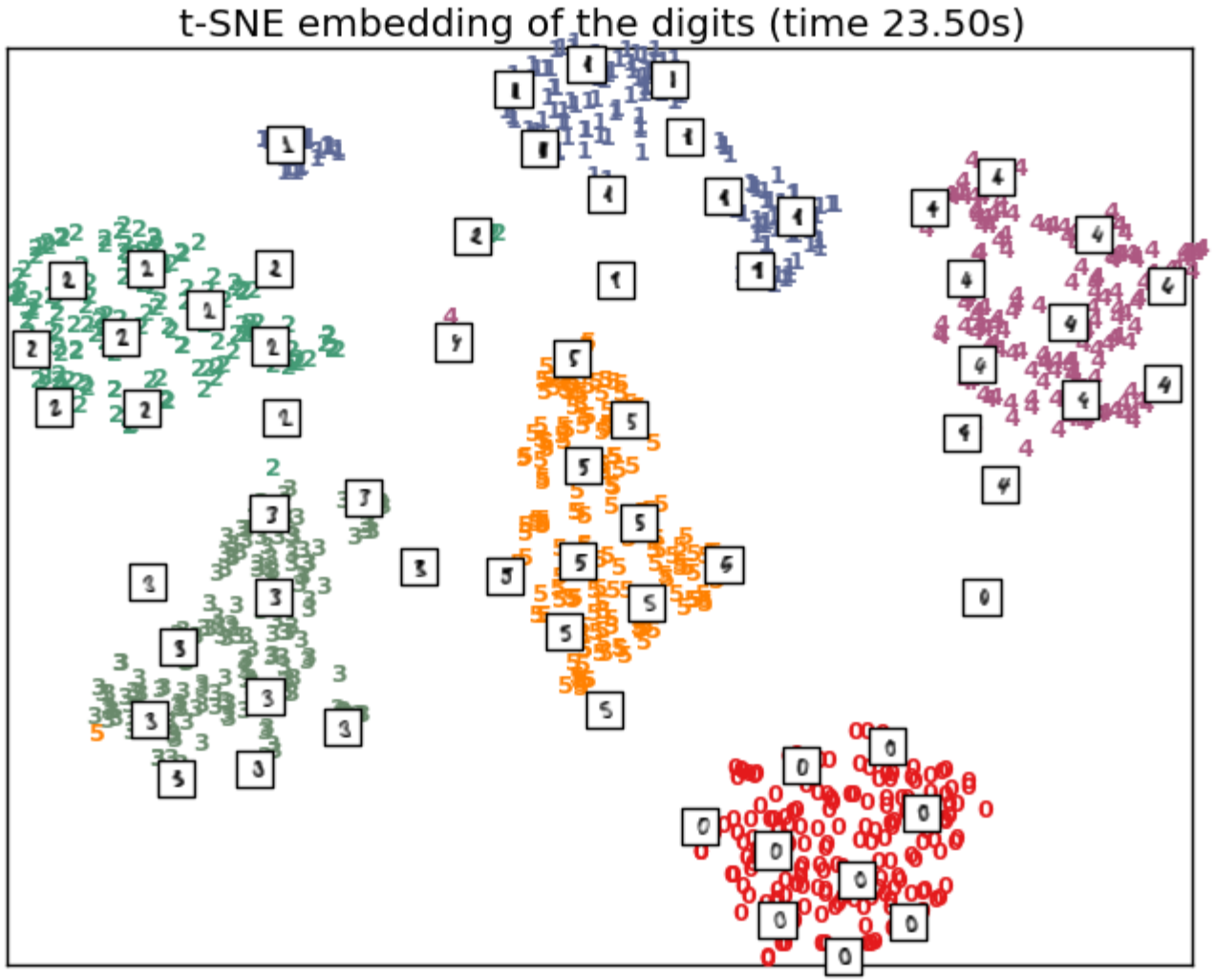
Нелинейное преобразование признаков

- SOM (Self-Organizing Maps) – самоорганизующиеся карты Кохонена. Не самый новый алгоритм, но идейно очень прост.
- Есть целое направление Manifold Learning

Manifold learning: Isomap



Manifold learning: t-SNE



Резюме

1. Задача понижения размерности
2. Метод главных компонент и SVD
3. Manifold learning

На следующей лекции: анализ текстов