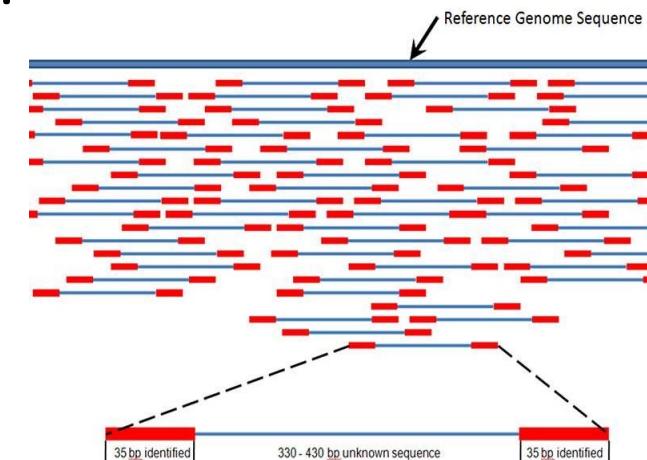


Понижение размерности

Отбор признаков и понижение размерности

Биоинформатика

- Задачи анализа генома человека
- Признаки: характеристики генов (более 20.000)
- Маленькие выборки (расшифровка генома — сложная и дорогостоящая процедура)
- Признаков существенно больше, чем объектов!



Категориальные признаки

- Пример: предсказать, понравится ли пользователю фильм
- Объект: пара «пользователь-фильм»
- Признаки: ID пользователя, ID фильма, ID жанра, ID режиссёра, ID главных актёров, ID композитора, ...
- Как много фильмов снято за всю историю?
- IMDB: >330 тысяч
- После бинарного кодирования получим миллионы признаков

Анализ текстов

- Пример: предсказание популярности фильма по тексту его сценария
- Признаки: количество вхождений каждого слова из словаря
- Сколько слов в словаре?
- Сотни тысяч признаков
- Если учитывать n -граммы, то десятки миллионов признаков

Анализ данных ЭЭГ

- Энцефалограф: 64 датчика, частота сигнала 256 Гц
- Объект: результаты измерений для одного пациента
- За 5 секунд измерений: $64 * 256 * 5 = 81\ 920$ признаков

Электроэнцефалография, ЭЭГ — раздел электрофизиологии, изучающий закономерности суммарной электрической активности мозга



UCI Machine Learning Repository



Browse Through: 557 Data Sets

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
URL Reputation	Multivariate, Time-Series	Classification	Integer, Real	2396130	3231961	2009
KASANDR	Multivariate	Causal-Discovery	Integer	17764280	2158859	2017
Gas sensor arrays in open sampling settings	Multivariate, Time-Series	Classification	Real	18000	1950000	2013
YouTube Multiview Video Games Dataset	Multivariate, Text	Classification, Clustering	Integer, Real	120000	1000000	2013
Twin gas sensor arrays	Multivariate, Time-Series, Domain-Theory	Classification, Regression	Real	640	480000	2016
Deepfakes: Medical Image Tamper Detection	Multivariate	Classification	Real	20000	200000	2020
Gas sensor array exposed to turbulent gas mixtures	Multivariate, Time-Series	Classification, Regression	Real	180	150000	2014
ElectricityLoadDiagrams20112014	Time-Series	Regression, Clustering	Real	370	140256	2015

<https://archive.ics.uci.edu/ml/index.html>

<https://archive.ics.uci.edu/ml/datasets.php>

Онлайн-ресурс, который предоставляет доступ к большому количеству наборов данных, используемых для исследований в области машинного обучения и статистики. Он был создан в 1987 году в Университете Калифорнии в Ирвайне (UCI) и с тех пор стал важным инструментом для исследователей, студентов и практиков в области анализа данных.

Репозиторий включает в себя разнообразные наборы данных, охватывающие различные области, такие как медицина, финансы, биология, социальные науки и многие другие. Каждый набор данных обычно сопровождается описанием, метаданными и иногда примерами использования, что делает его удобным для обучения и тестирования алгоритмов машинного обучения.

Задача понижения размерности

- Дано: матрица «объекты-признаки» X размера $\ell \times D$
- Найти: новую матрицу «объекты-признаки» Z размера $\ell \times d$
- $d < D$

Но зачем?

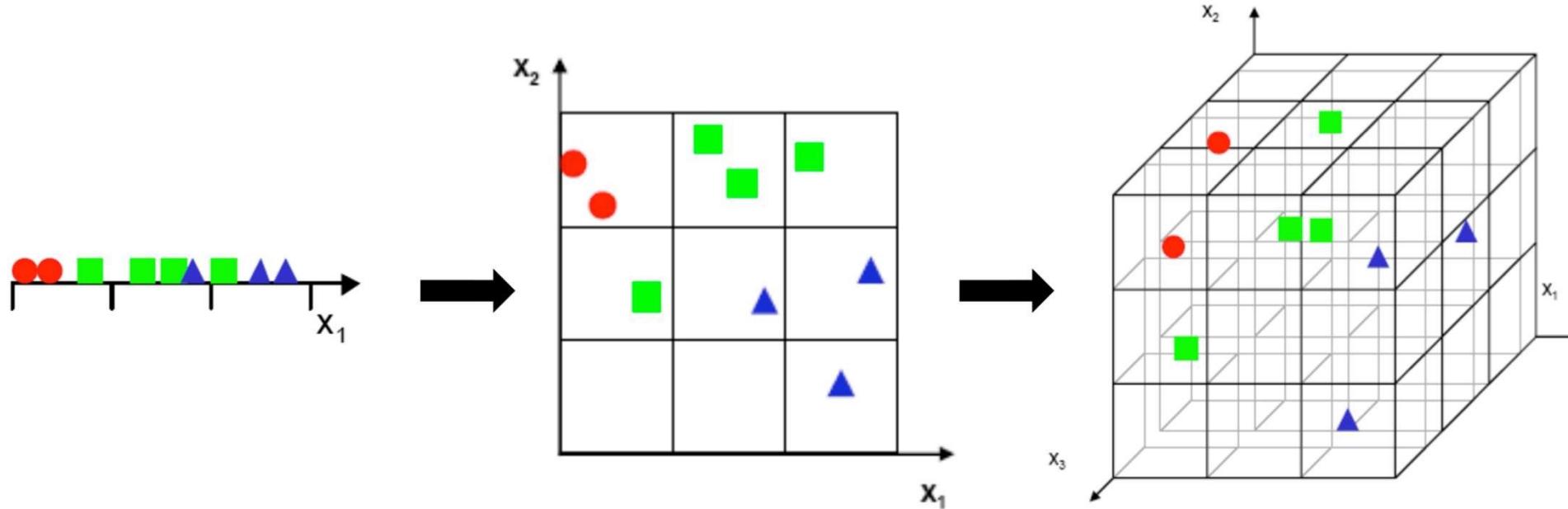
- Проклятие размерности
- Шумовые признаки
- Переобучение
- Интерпретируемость модели
- Скорость работы модели
- Визуализация данных

Проклятие размерности

- Задача: классификация пончиков на вкусные и невкусные
- 100 объектов
- Цвет: 10 вариантов
- Цвет + размер: $10 * 4 = 40$ вариантов
- Цвет + размер + форма: $10 * 4 * 4 = 160$ вариантов



Проклятие размерности



Проклятие размерности (curse of dimensionality) — это термин, используемый в статистике и машинном обучении для описания различных явлений, которые возникают, когда анализируемые данные имеют высокую размерность (то есть много признаков или переменных). С увеличением размерности данные становятся разреженными, и это может привести к нескольким проблемам:

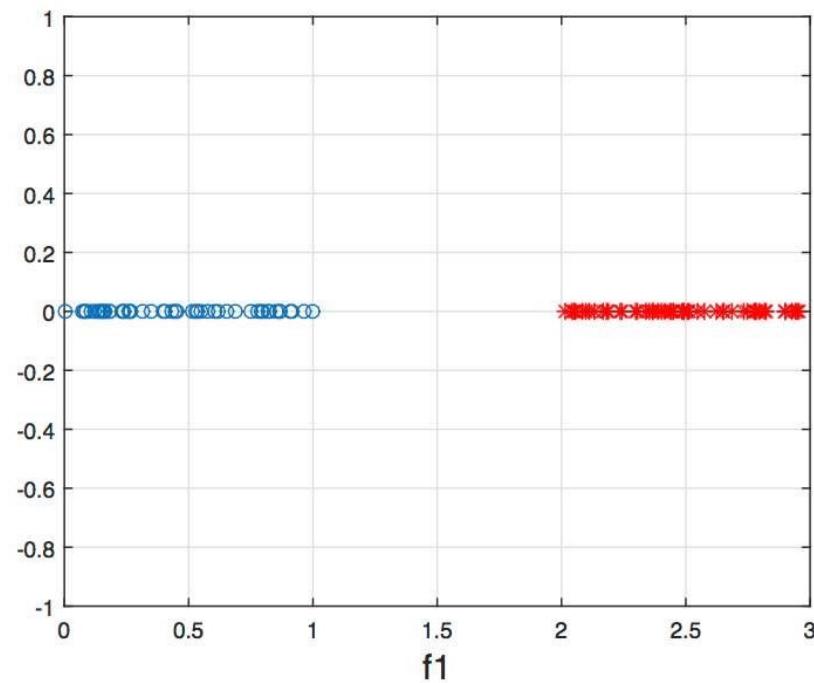
Разреженность данных: В высоких размерностях данные становятся разреженными, что затрудняет выявление закономерностей. Например, в двухмерном пространстве точки могут быть близки друг к другу, но в многомерном пространстве они могут оказаться далеко друг от друга.

Увеличение объема данных: Для адекватного представления данных в высоких размерностях требуется значительно больше данных. Это может привести к необходимости в больших объемах вычислительных ресурсов и времени для обработки.

Проблемы с переобучением: В высоких размерностях модели могут легко подстраиваться под шум в данных, что приводит к переобучению. Это означает, что модель будет хорошо работать на обучающем наборе данных, но плохо обобщать на новых, невидимых данных.

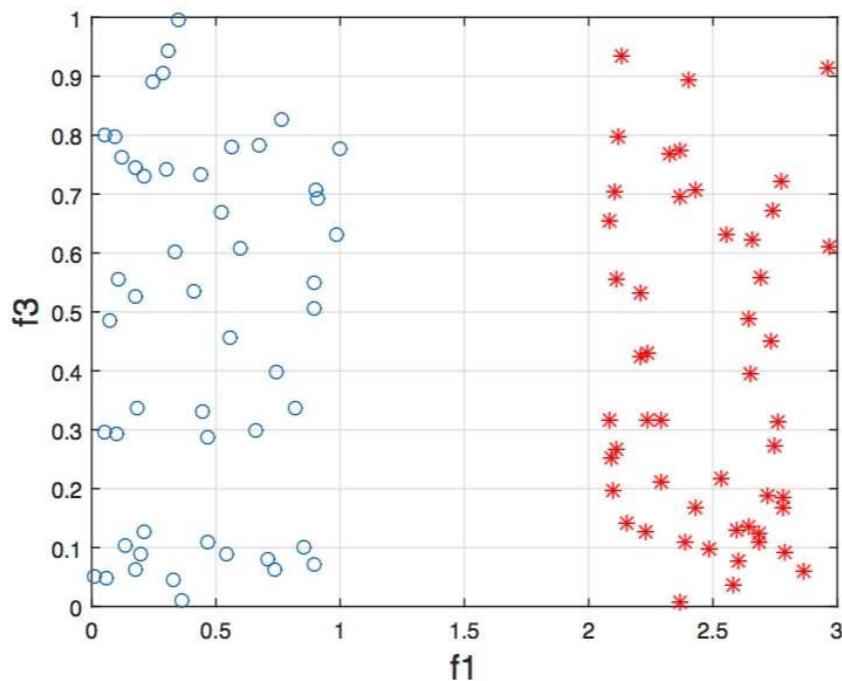
Плохие признаки

Информативный
признак



Плохие признаки

f3 — шумовой
признак



Шумовой признак (или шумовая переменная) — это переменная в наборе данных, которая не содержит полезной информации для предсказания целевой переменной и может даже ухудшать качество модели. В контексте машинного обучения и анализа данных, шумовые признаки могут вносить путаницу и мешать алгоритмам выявлять истинные закономерности.

Признак f3, упомянутый вами, может быть шумовым признаком, если:

Не имеет корреляции с целевой переменной: Если f3 не связан с целевой переменной, то его наличие в модели может привести к ухудшению предсказаний.

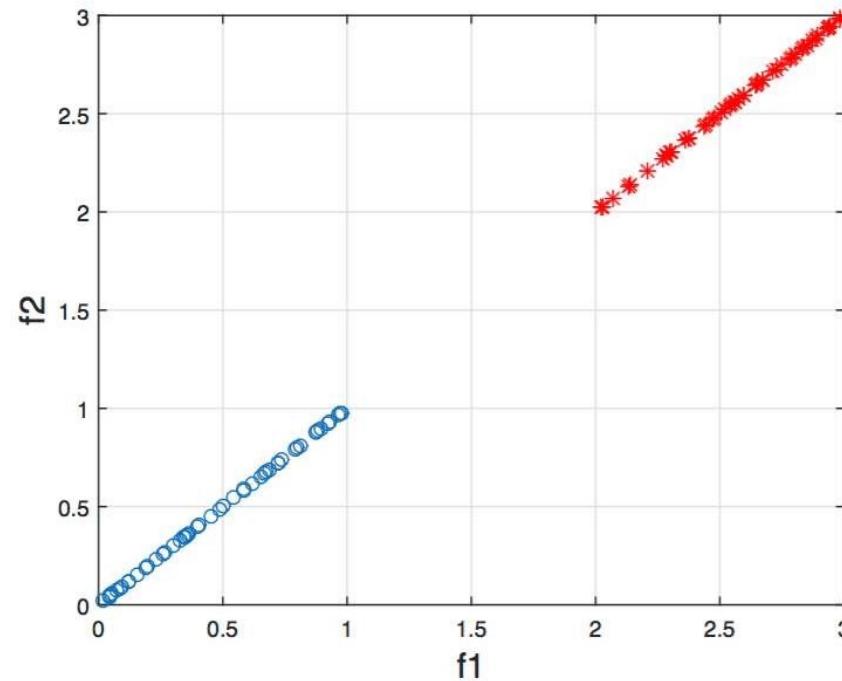
Вносит случайные колебания: Если значения f3 случайны и не имеют никакой логической связи с другими признаками или целевой переменной, это также делает его шумовым признаком.

Увеличивает сложность модели: Наличие шумовых признаков может привести к переобучению, так как модель может попытаться подстроиться под случайные колебания, а не под истинные закономерности.

Плохие признаки

Коррелирующие
признаки

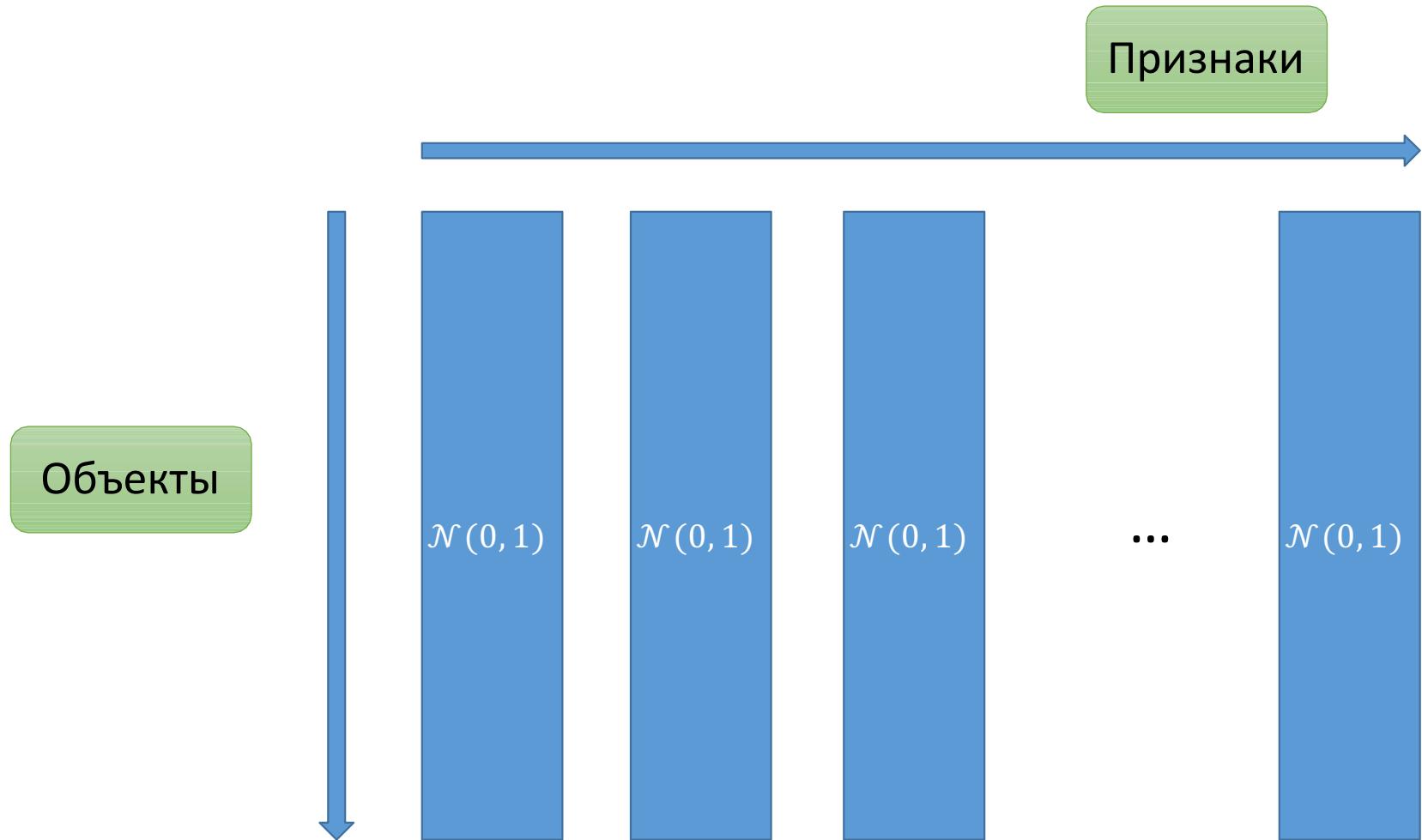
f_2 — избыточный
признак



Шумовые признаки

- Признаки, которые никак не связаны с целевой переменной
- Но по обучающей выборке это не всегда можно понять

Шумовые признаки



Шумовые признаки

- Генерируем случайные признаки
- Если их много, то некоторые будут хорошо коррелировать с ответами

y	x_1	x_2	x_3	x_4
-1	1.11	-0.5	0.42	0.33
-1	1.22	-0.46	-1.98	-0.55
1	-1.56	0.04	0.39	-1.67
1	-0.48	1.32	0.88	-0.27

Ускорение моделей

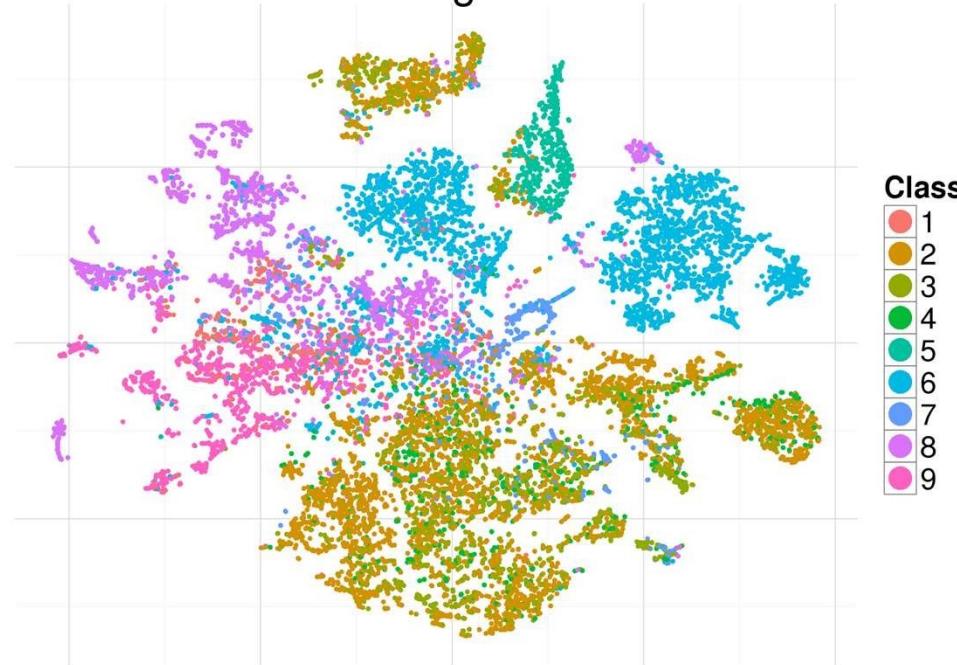
- Чем больше признаков, тем дольше обучаются модели
- Чем дольше обучаются модели, тем меньше экспериментов удаётся провести
- Чем сложнее модели, тем дольше они вычисляют прогнозы
- Могут быть жёсткие ограничения на скорость
- Пример: рекомендательные системы

Визуализация

-99.99	-99.99	315.7	317.45	317.5	317.26	315.86	314.93	313.2	312.44	313.33	314.67	-99.99
315.62	316.38	316.71	317.72	318.29	318.16	316.54	314.8	313.84	313.26	314.8	315.58	315.98
316.43	316.97	317.58	319.02	320.03	319.59	318.18	315.91	314.16	313.84	315	316.19	316.91
316.93	317.7	318.54	319.48	320.58	319.77	318.57	316.79	314.8	315.38	316.1	317.01	317.64
317.94	318.56	319.68	320.63	321.01	320.55	319.58	317.4	316.25	315.42	316.69	317.7	318.45
318.74	319.08	319.86	321.39	322.24	321.47	319.74	317.77	316.21	315.99	317.12	318.31	318.99
319.57	-99.99	-99.99	-99.99	322.24	321.89	320.44	318.7	316.7	316.79	317.79	318.71	-99.99
319.44	320.44	320.89	322.13	322.16	321.87	321.39	318.8	317.81	317.3	318.87	319.42	320.04
320.62	321.59	322.39	323.87	324.01	323.75	322.39	320.37	318.64	318.1	319.79	321.08	321.38
322.06	322.5	323.04	324.42	325	324.09	322.55	320.92	319.31	319.31	320.72	321.96	322.16
322.57	323.15	323.89	325.02	325.57	325.36	324.14	322.03	320.41	320.25	321.31	322.84	323.05
324	324.42	325.64	326.66	327.34	326.76	325.88	323.67	322.39	321.78	322.85	324.12	324.63
325.03	325.99	326.87	328.14	328.07	327.66	326.35	324.69	323.1	323.16	323.98	325.13	325.68
326.17	326.68	327.18	327.78	328.92	328.57	327.34	325.46	323.36	323.57	324.8	326.01	326.32
326.77	327.63	327.75	329.72	330.07	329.09	328.05	326.32	324.93	325.06	326.5	327.55	327.45
328.55	329.56	330.3	331.5	332.48	332.07	330.87	329.31	327.51	327.18	328.16	328.64	329.68
329.35	330.71	331.48	332.65	333.09	332.25	331.18	329.4	327.43	327.37	328.46	329.57	330.25
330.4	331.41	332.04	333.31	333.96	333.6	333.91	330.06	328.56	328.34	329.49	330.76	331.15
331.75	332.56	333.5	334.58	334.87	334.34	333.05	330.94	329.3	328.94	330.31	331.68	332.15
332.93	333.42	334.7	336.07	336.74	336.27	334.93	332.75	331.59	331.16	332.4	333.85	333.9
334.97	335.39	336.64	337.76	338.01	337.89	336.54	334.68	332.76	332.55	333.92	334.95	335.51
336.23	336.76	337.96	338.89	339.47	339.29	337.73	336.09	333.91	333.86	335.29	336.73	336.85
338.01	338.36	340.08	340.77	341.46	341.17	339.56	337.6	335.88	336.02	337.1	338.21	338.69
339.23	340.47	341.38	342.51	342.91	342.25	340.49	338.43	336.69	336.86	338.36	339.61	339.93
340.75	341.61	342.7	343.57	344.13	343.35	342.06	339.81	337.98	337.86	339.26	340.49	341.13
341.37	342.52	343.1	344.94	345.75	345.32	343.99	342.39	339.86	339.99	341.15	342.99	342.78
343.7	344.5	345.28	347.08	347.43	346.79	345.4	343.28	341.07	341.35	342.98	344.22	344.42
344.97	346	347.43	348.35	348.93	348.25	346.56	344.68	343.09	342.8	344.24	345.55	345.9
346.3	346.96	347.86	349.55	350.21	349.54	347.94	345.9	344.85	344.17	345.66	346.9	347.15
348.02	348.47	349.42	350.99	351.84	351.25	349.52	348.1	346.45	346.36	347.81	348.96	348.93
350.43	351.73	352.22	353.59	354.22	353.79	352.38	350.43	348.72	348.88	350.07	351.34	351.48
352.76	353.07	353.68	355.42	355.67	355.13	353.9	351.67	349.8	349.99	351.29	352.52	352.91
353.66	354.7	355.39	356.2	357.16	356.23	354.82	352.91	350.96	351.18	352.83	354.21	354.19
354.72	355.75	357.16	358.6	359.33	358.24	356.17	354.02	352.15	352.21	353.75	354.99	355.59
355.98	356.72	357.81	359.15	359.66	359.25	357.02	355	353.01	353.31	354.16	355.4	356.37
356.7	357.16	358.38	359.46	360.28	359.6	357.57	355.52	353.69	353.99	355.34	356.8	357.04
358.37	358.91	359.97	361.26	361.68	360.95	359.55	357.48	355.84	355.99	357.58	359.04	358.89
359.97	361	361.64	363.45	363.79	363.26	361.9	359.46	358.05	357.76	359.56	360.7	360.88
362.05	363.25	364.02	364.72	365.41	364.97	363.65	361.48	359.45	359.6	360.76	362.33	362.64
363.18	364	364.56	366.35	366.79	365.62	364.47	362.51	360.19	360.77	362.43	364.28	363.76
365.33	366.15	367.31	368.61	369.3	368.87	367.64	365.77	363.9	364.23	365.46	366.97	366.63
368.15	368.87	369.59	371.14	371	370.35	369.27	366.93	364.63	365.13	366.67	368.01	368.31
369.14	369.46	370.52	371.66	371.82	371.7	370.12	368.12	366.62	366.73	368.29	369.53	369.48
370.28	371.5	372.12	372.87	374.02	373.3	371.62	369.55	367.96	368.09	369.68	371.24	371.02
372.43	373.09	373.52	374.86	375.55	375.41	374.02	371.49	370.7	370.25	372.08	373.78	373.1
374.68	375.63	376.11	377.65	378.35	378.13	376.62	374.5	372.99	373.01	374.35	375.7	375.64
376.79	377.37	378.41	380.52	380.63	379.57	377.79	375.86	374.07	374.24	375.86	377.47	377.38
378.37	379.69	380.41	382.1	382.28	382.13	380.66	378.71	376.42	376.88	378.32	380.04	379.67
381.38	382.03	382.64	384.62	384.95	384.06	382.29	380.47	378.67	379.06	380.14	381.74	381.84
382.45	383.68	384.23	386.26	386.39	385.87	384.39	381.78	380.73	380.81	382.33	383.69	383.55
385.07	385.72	385.85	386.71	388.45	387.64	386.1	383.95	382.91	382.73	383.96	385.02	385.34

Визуализация

t-SNE 2D Embedding of Products Data



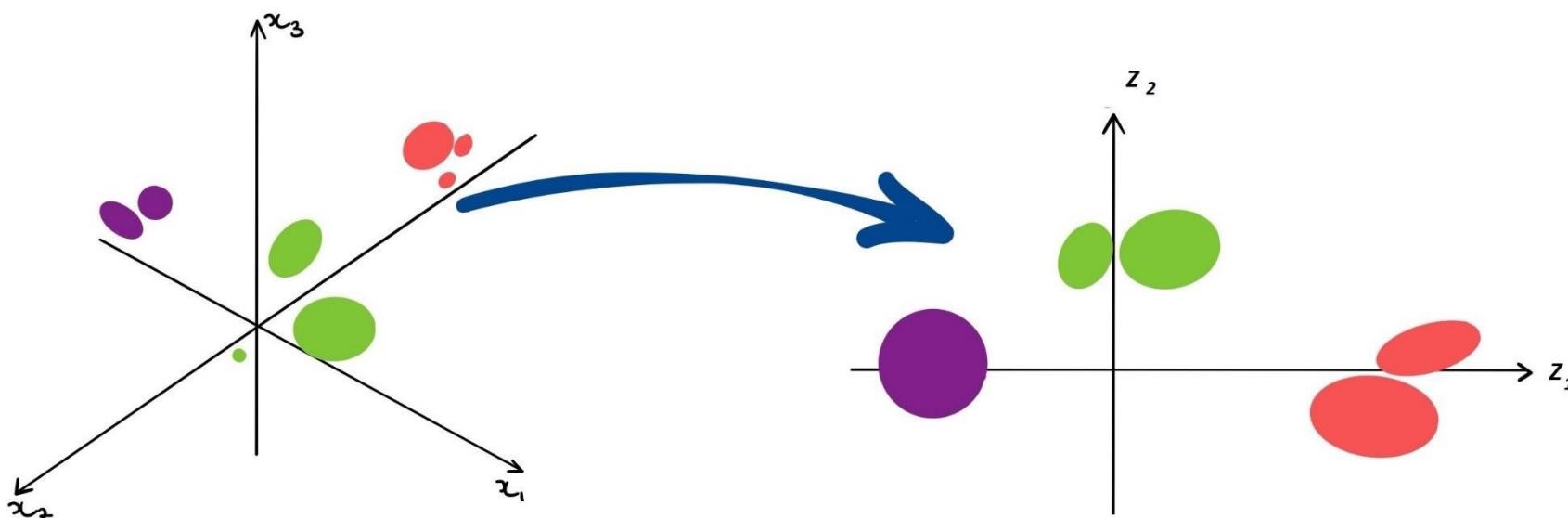
Методы понижения размерности

- Отбор признаков (feature selection)
 - Выбрать d самых важных признаков
- **Извлечение признаков (feature extraction)**
 - Найти d новых признаков, выражающихся через исходные

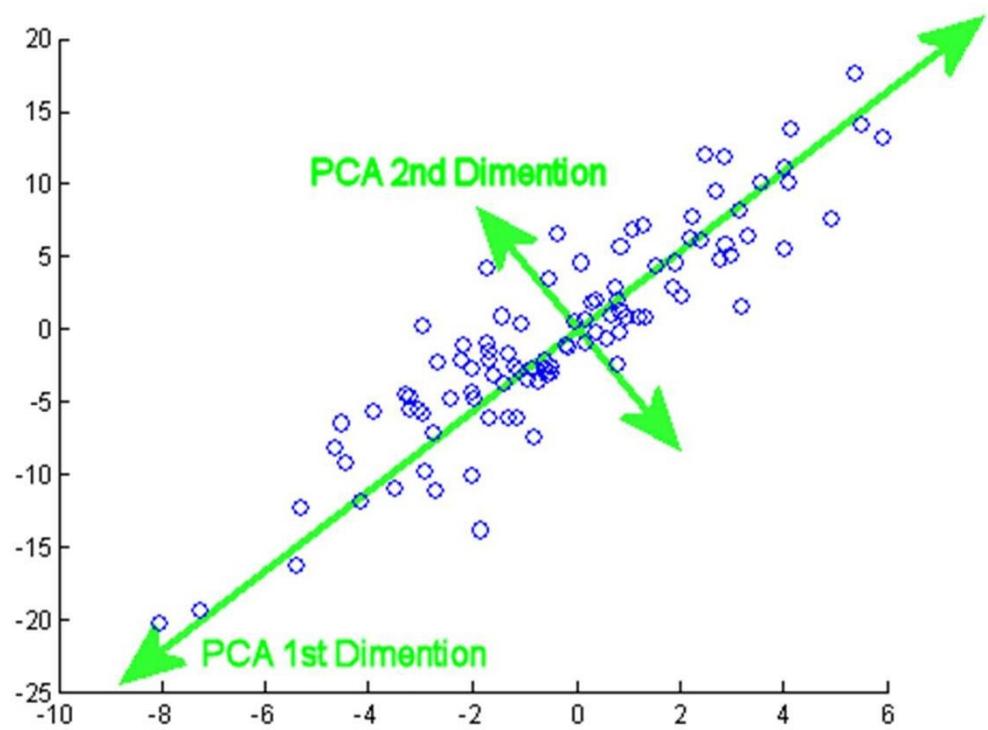
Метод главных компонент

Метод главных компонент

- Principal component analysis (PCA)
- Проецирует данные в пространство меньшей размерности
- Относится к методам извлечения признаков



Извлечение признаков



Извлечение признаков

- Порождение новых признаков
- Их должно быть меньше
- Они должны содержать как можно больше информации из исходных признаков
- Линейные методы
- Каждый новый признак — линейная комбинация исходных

Извлечение признаков

- Исходные признаки: x_{ik}, D штук
- Новые признаки: z_{ij}, d штук
- Линейный подход:

$$z_{ij} = \sum_{k=1}^D w_{jk} x_{ik}$$

Новые признаки

Вклад исходного k -го
признака в новый j -й

Исходные признаки

Метод главных компонент

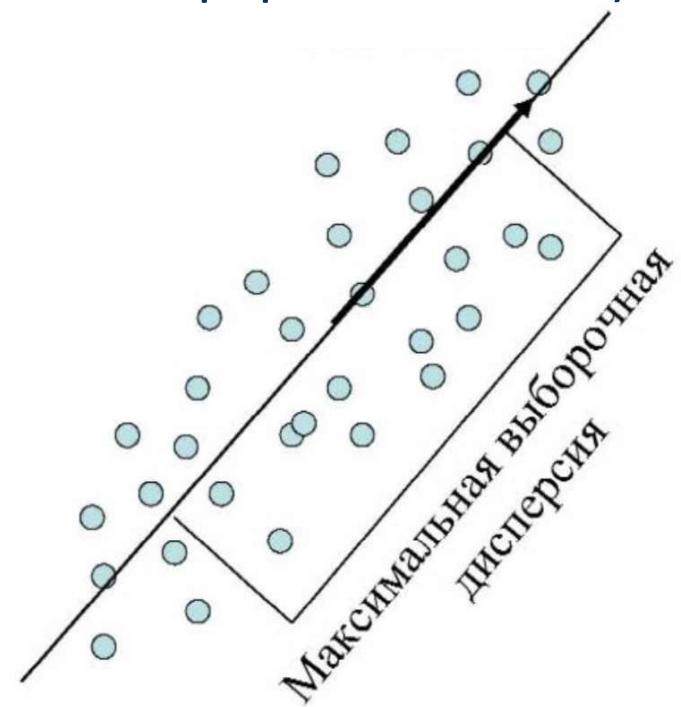
- Матричная запись:

$$Z = XW^T$$

- j -й столбец W — коэффициенты при исходных признаках для вычисления нового j -го признака

Principal Component Analysis 1

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



ПРИМЕР: FACES DATASET



FACES DATASET(ГЛАВНЫЕ КОМПОНЕНТЫ)



ВОССТАНОВЛЕННОЕ ИЗОБРАЖЕНИЕ

#efaces=1, res=57.804

#efaces=2, res=57.611

#efaces=5, res=54.054

#efaces=10, res=52.01

#efaces=20, res=45.897



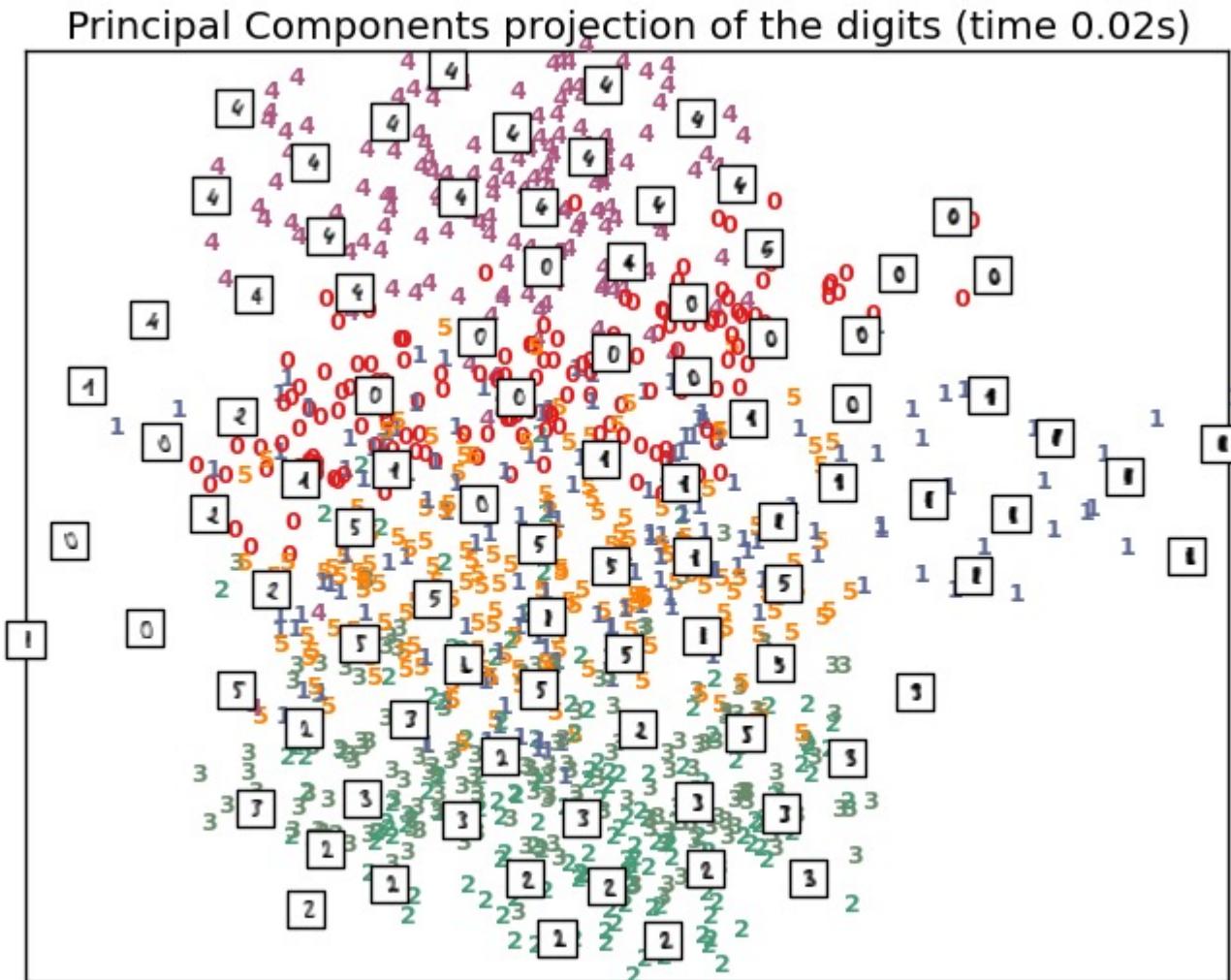
#efaces=40, res=35.868 #efaces=60, res=29.624 #efaces=80, res=24.103 #efaces=100, res=20.317 #efaces=150, res=16.154



#efaces=200, res=13.257 #efaces=300, res=9.581 #efaces=400, res=6.908 #efaces=1000, res=0.924 #efaces=1071, res=0.653

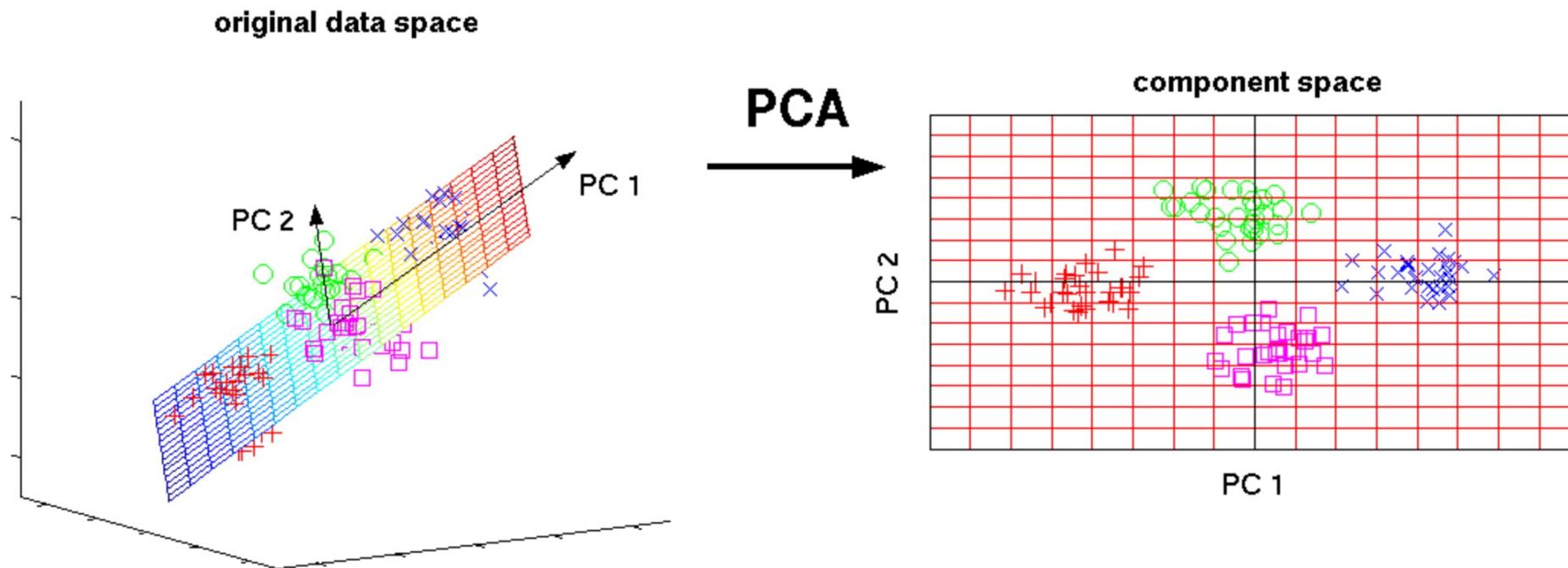


Рукописные цифры: проекция на главные компоненты



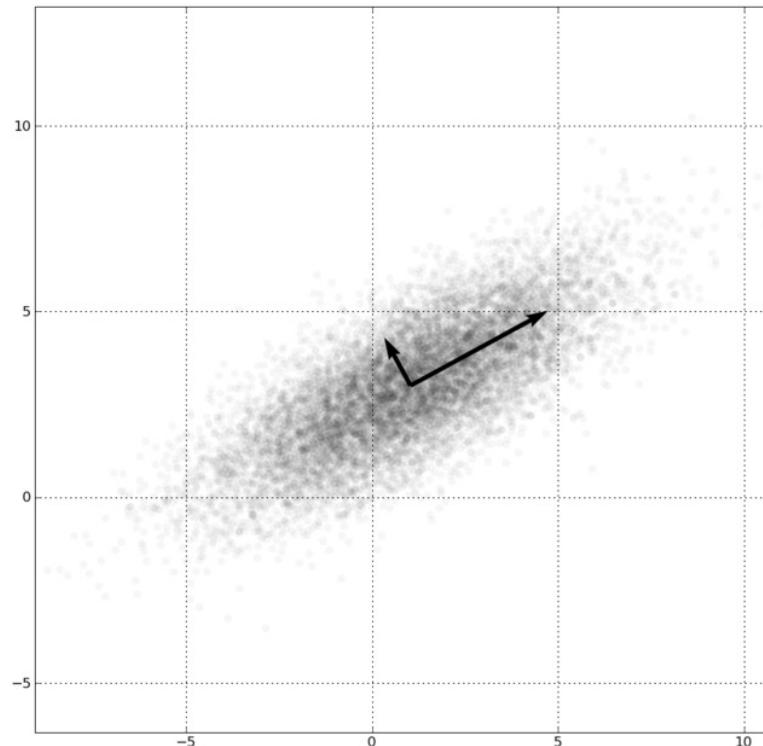
PCA (интерпретация 2)

Идея 2: давайте строить проекцию выборки на линейное подпространство меньшей размерности. А выбирать его так, чтобы квадраты отклонений точек от проекций были минимальны.

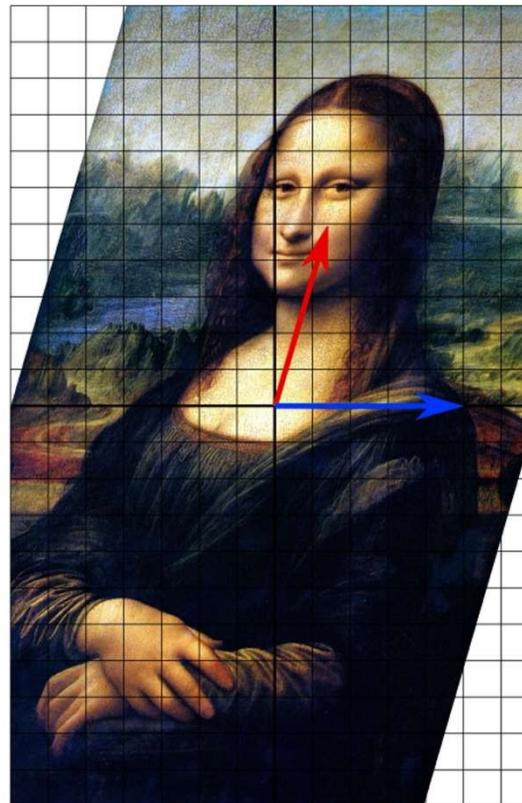
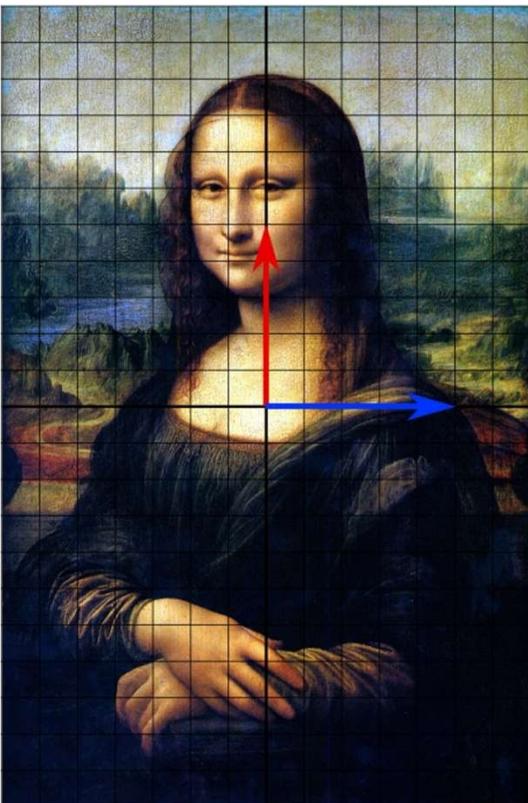


PCA (интерпретация 3)

Идея 3. Переход в базис, в котором матрица ковариаций диагональна



Собственные векторы



Собственный вектор — понятие в линейной алгебре, определяемое для произвольного линейного оператора как ненулевой вектор, применение к которому оператора даёт коллинеарный вектор — тот же вектор, умноженный на некоторое скалярное значение (которое может быть равно 0). Скаляр, на который умножается собственный вектор под действием оператора, называется собственным числом (или собственным значением) линейного оператора, соответствующим данному собственному вектору.

Собственные векторы

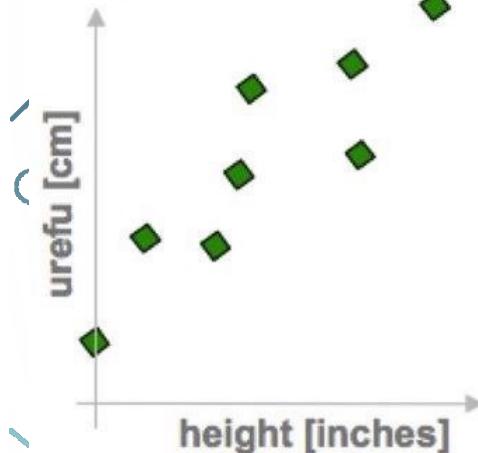
- A — матрица размера $n \times n$
- Пусть $Ax = \lambda x$
- Тогда x — собственный вектор, λ — собственное значение
- x — вектор, который не меняет направление под действием матрицы

Решение

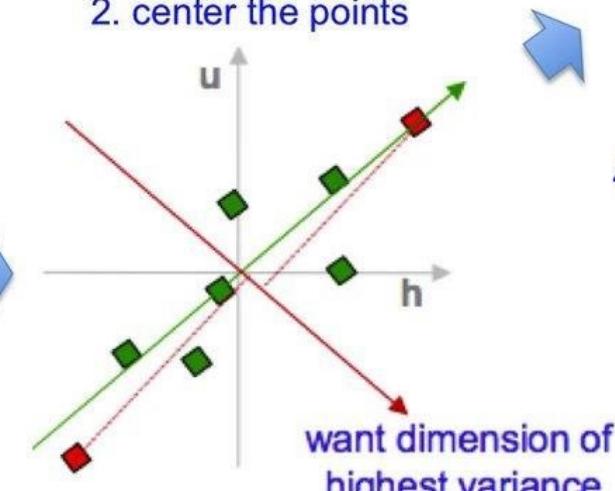
- Столбцы W — собственные векторы матрицы $X^T X$, соответствующие наибольшим собственным значениям $\lambda_1, \lambda_2, \dots, \lambda_d$
- $\frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^D \lambda_i}$ — доля дисперсии, сохранённой при понижении размерности

PCA in a nutshell

1. correlated hi-d data
("urefu" means "height" in Swahili)



2. center the points



3. compute covariance matrix

$$\begin{matrix} h & u \\ \begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \end{matrix} \rightarrow \text{cov}(h, u) = \frac{1}{n} \sum_{i=1}^n h_i u_i$$

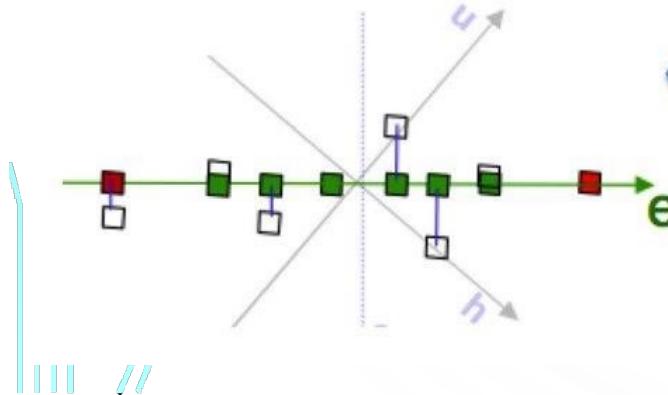
4. eigenvectors + eigenvalues

$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} e_h \\ e_w \end{pmatrix} = \lambda_e \begin{pmatrix} e_h \\ e_w \end{pmatrix}$$

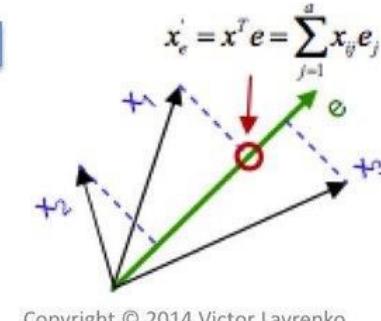
$$\begin{pmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} f_h \\ f_u \end{pmatrix} = \lambda_f \begin{pmatrix} f_h \\ f_u \end{pmatrix}$$

`eig(cov(data))`

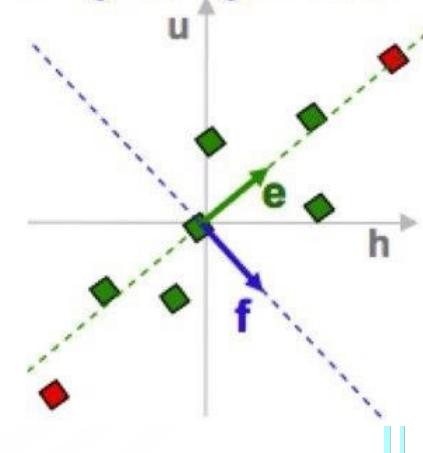
7. uncorrelated low-d data



6. project data points to those eigenvectors



5. pick m<d eigenvectors w. highest eigenvalues



Copyright © 2014 Victor Lavrenko

PCA (интерпретация 4)

Приблизим исходную матрицу признаков произведением двух матриц:

$$X \approx U \cdot V^T$$

$l \times n \quad l \times k \quad k \times n$

$$\| X - U \cdot V^T \| \rightarrow \min$$

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

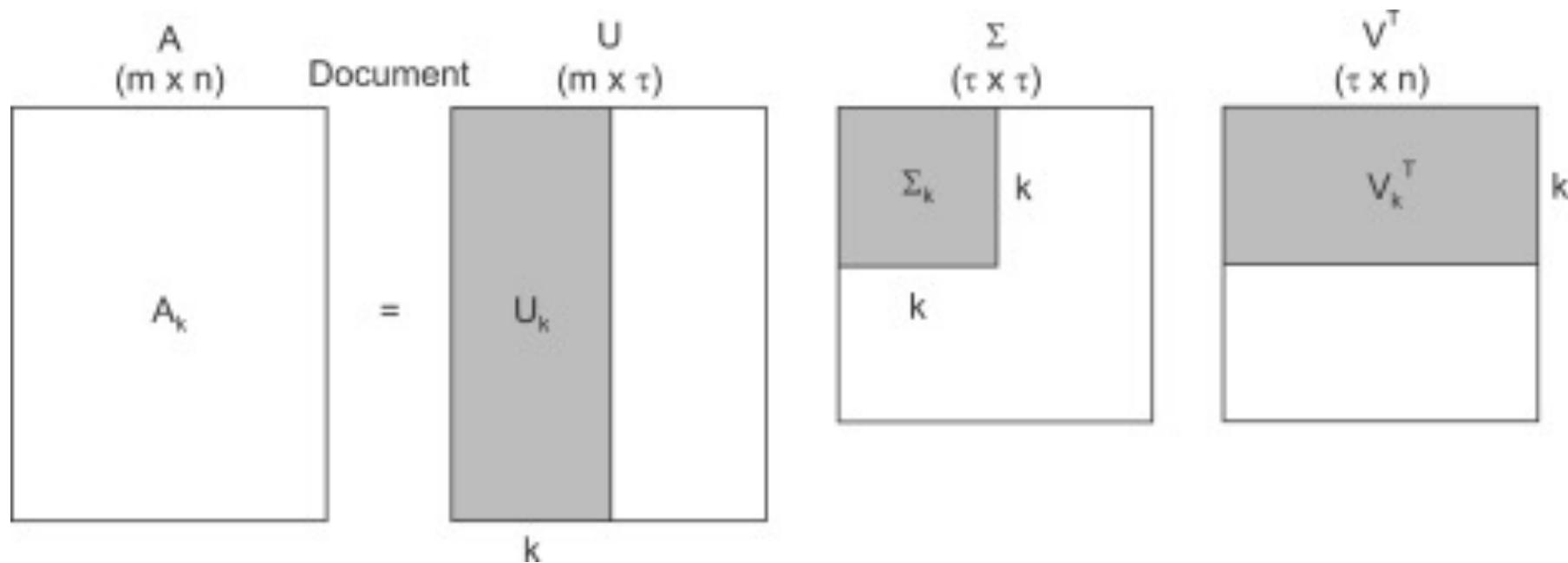
SVD

SVD = Singular Vector Decomposition (сингулярное разложение матриц)

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

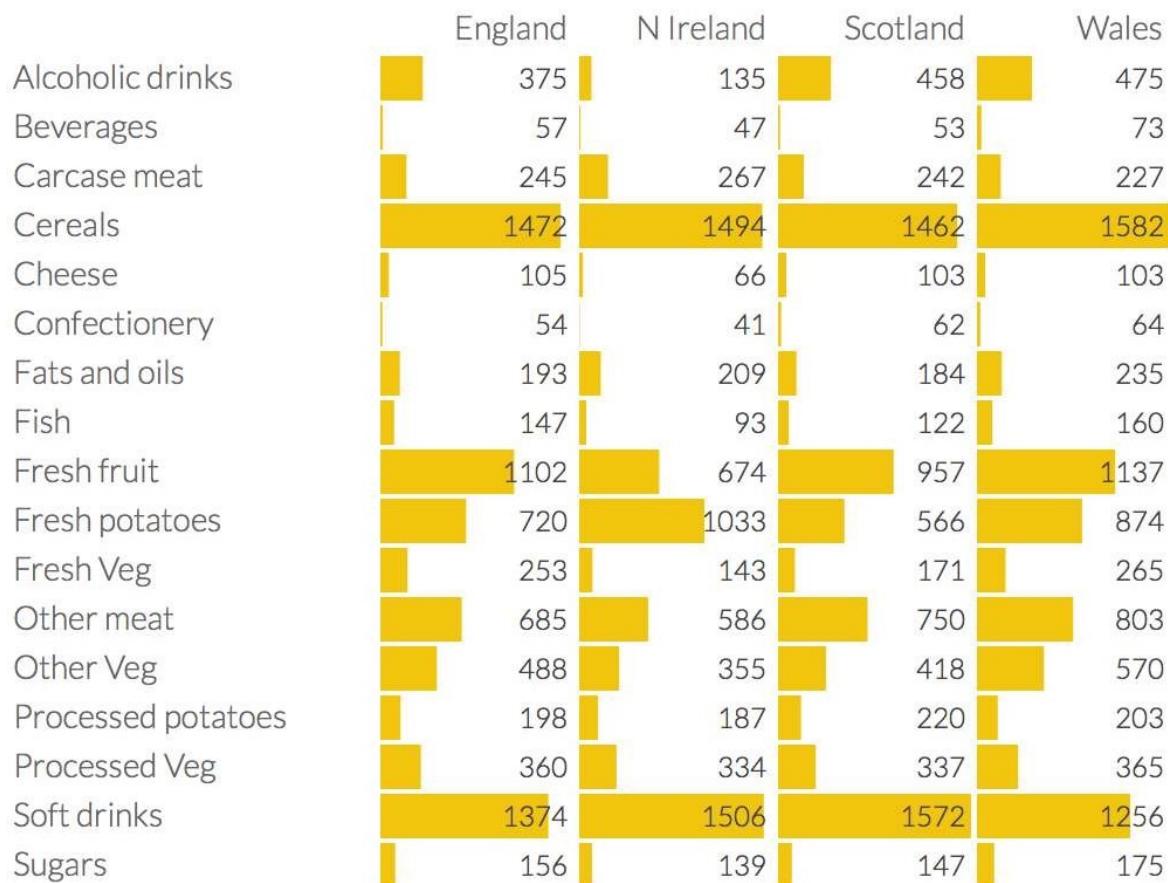
Применяется для снижения размерности пространства признаков.

SVD

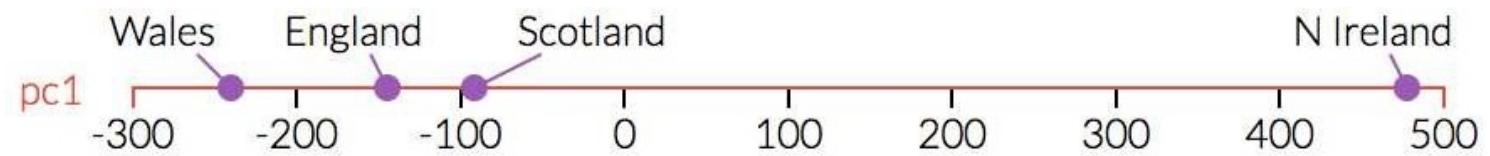


Рацион в Великобритании

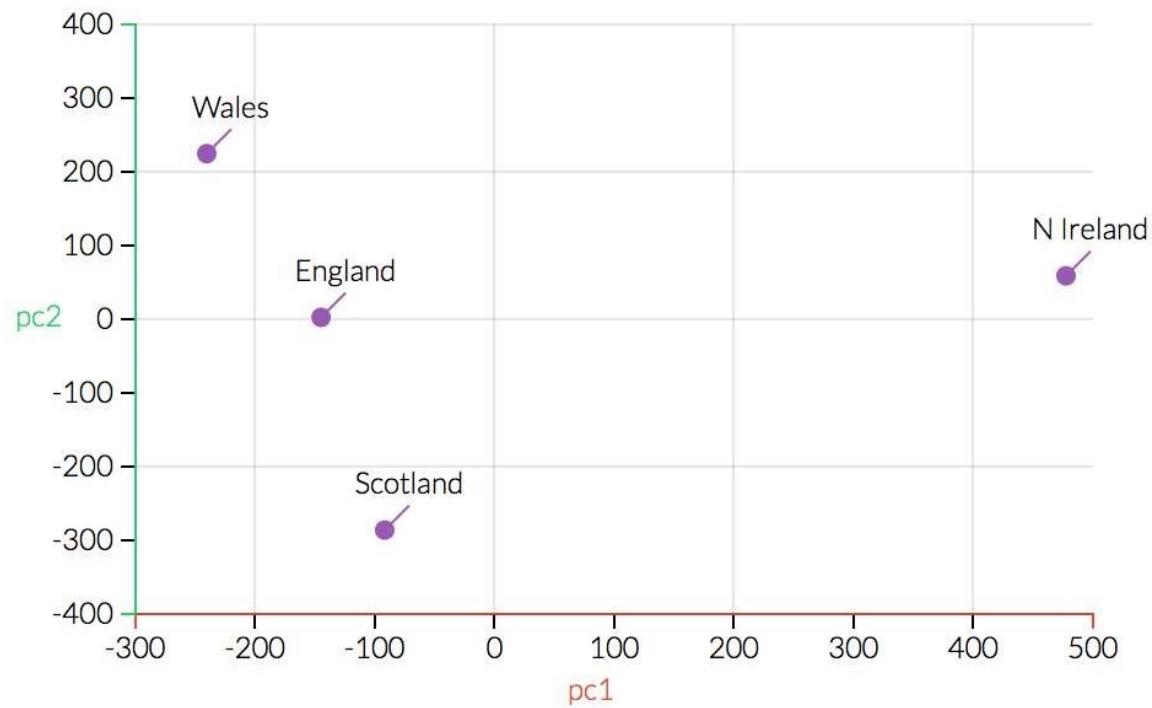
- Данные — среднее потребление продуктов в неделю в каждой провинции
- Не очень удобно смотреть на них



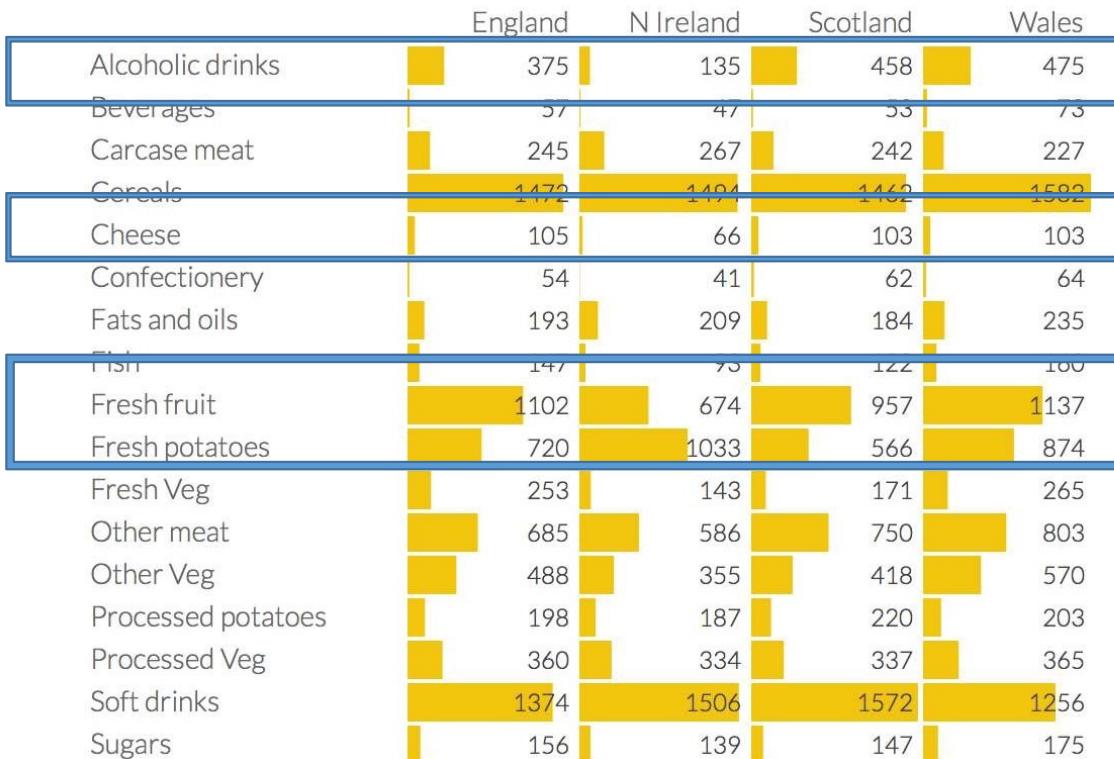
Рацион в Великобритании



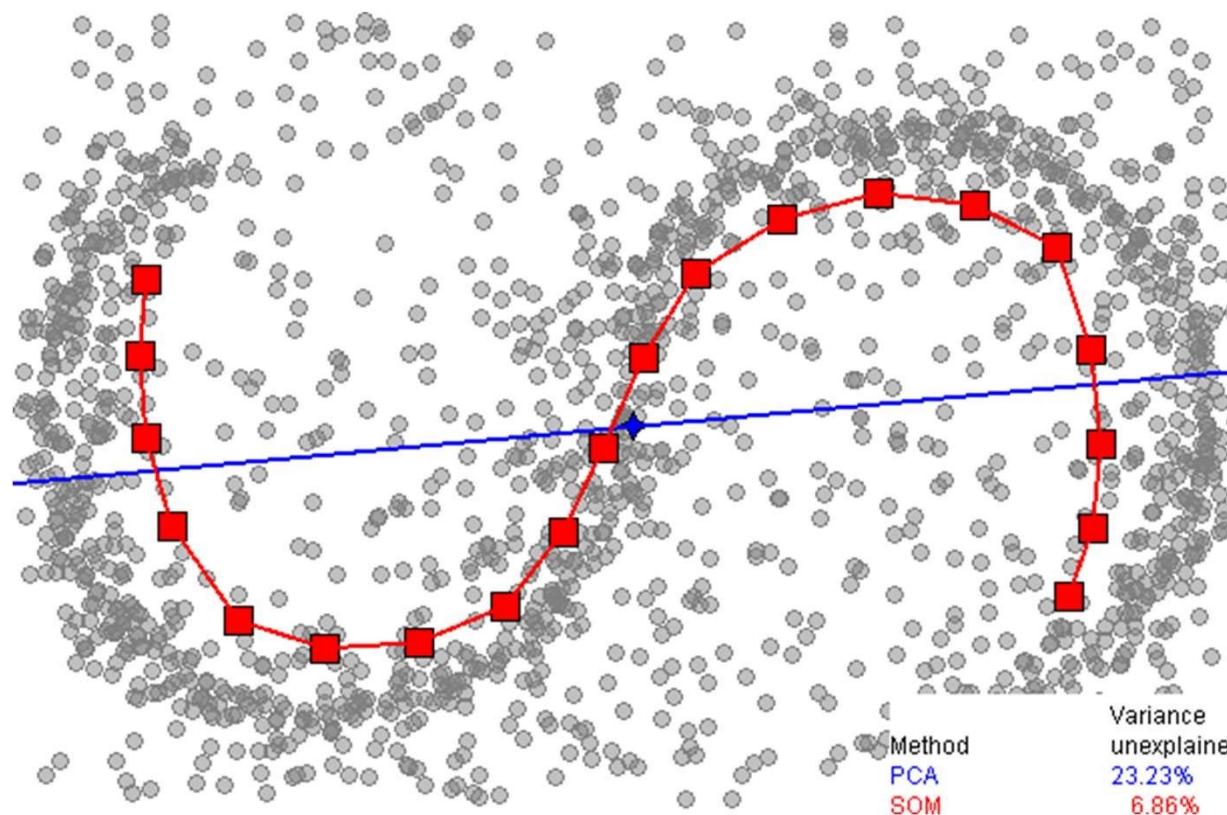
Рацион в Великобритании



Рацион в Великобритании



Ограничения

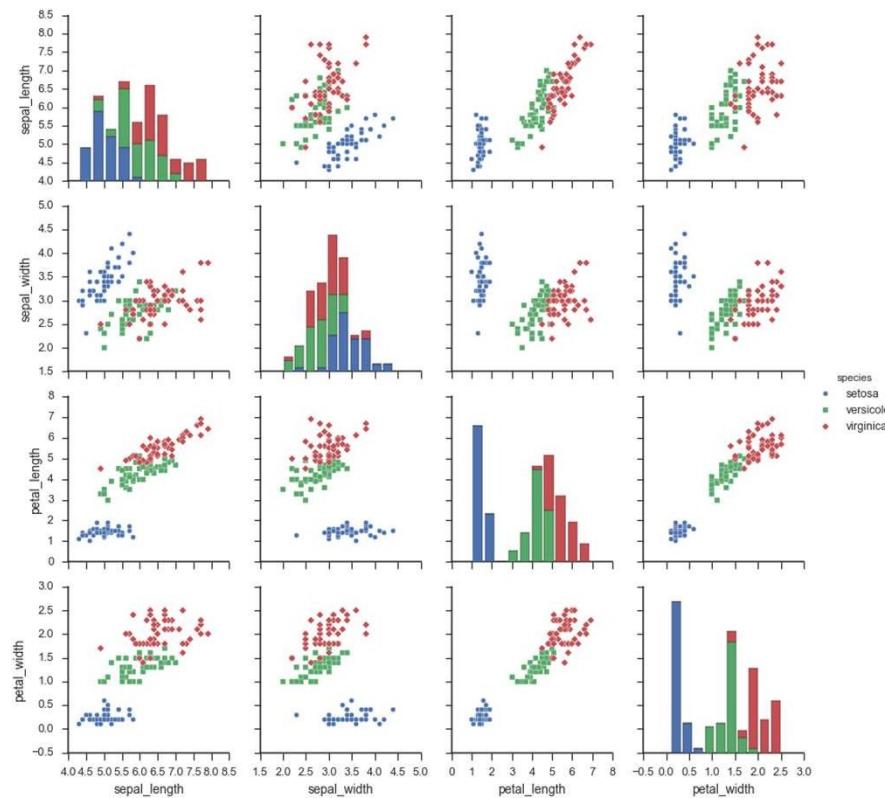


Визуализация данных

Визуализация данных

-99.99	-99.99	315.7	317.45	317.5	317.26	315.86	314.93	313.2	312.44	313.33	314.67	-99.99
315.62	316.38	316.71	317.72	318.29	318.16	316.54	314.8	313.84	313.26	314.8	315.58	315.98
316.43	316.97	317.58	319.02	320.03	319.59	318.18	315.91	314.16	313.84	315	316.19	316.91
316.93	317.7	318.54	319.48	320.58	319.77	318.57	316.79	314.8	315.38	316.1	317.01	317.64
317.94	318.56	319.68	320.63	321.01	320.55	319.58	317.4	316.25	315.42	316.69	317.7	318.45
318.74	319.08	319.86	321.39	322.24	321.47	319.74	317.77	316.21	315.99	317.12	318.31	318.99
319.57	-99.99	-99.99	-99.99	322.24	321.89	320.44	318.7	316.7	316.79	317.79	318.71	-99.99
319.44	320.44	320.89	322.13	322.16	321.87	321.39	318.8	317.81	317.3	318.87	319.42	320.04
320.62	321.59	322.39	323.87	324.01	323.75	322.39	320.37	318.64	318.1	319.79	321.08	321.38
322.06	322.5	323.04	324.42	325	324.09	322.55	320.92	319.31	319.31	320.72	321.96	322.16
322.57	323.15	323.89	325.02	325.57	325.36	324.14	322.03	320.41	320.25	321.31	322.84	323.05
324	324.42	325.64	326.66	327.34	326.76	325.88	323.67	322.39	321.78	322.85	324.12	324.63
325.03	325.99	326.87	328.14	328.07	327.66	326.35	324.69	323.1	323.16	323.98	325.13	325.68
326.17	326.68	327.18	327.78	328.92	328.57	327.34	325.46	323.36	323.57	324.8	326.01	326.32
326.77	327.63	327.75	329.72	330.07	329.09	328.05	326.32	324.93	325.06	326.5	327.55	327.45
328.55	329.56	330.3	331.5	332.48	332.07	330.87	329.31	327.51	327.18	328.16	328.64	329.68
329.35	330.71	331.48	332.65	333.09	332.25	331.18	329.4	327.43	327.37	328.46	329.57	330.25
330.4	331.41	332.04	333.31	333.96	333.6	331.91	330.06	328.56	328.34	329.49	330.76	331.15
331.75	332.56	333.5	334.58	334.87	334.34	333.05	330.94	329.3	328.94	330.31	331.68	332.15
332.93	333.42	334.7	336.07	336.74	336.27	334.93	332.75	331.59	331.16	332.4	333.85	333.9
334.97	335.39	336.64	337.76	338.01	337.89	336.54	334.68	332.76	332.55	333.92	334.95	335.51
336.23	336.76	337.96	338.89	339.47	339.29	337.73	336.09	333.91	333.86	335.29	336.73	336.85
338.01	338.36	340.08	340.77	341.46	341.17	339.56	337.6	335.88	336.02	337.1	338.21	338.69
339.23	340.47	341.38	342.51	342.91	342.25	340.49	338.43	336.69	336.86	338.36	339.61	339.93
340.75	341.61	342.7	343.57	344.13	343.35	342.06	339.81	337.98	337.86	339.26	340.49	341.13
341.37	342.52	343.1	344.94	345.75	345.32	343.99	342.39	339.86	339.99	341.15	342.99	342.78
343.7	344.5	345.28	347.08	347.43	346.79	345.4	343.28	341.07	341.35	342.98	344.22	344.42
344.97	346	347.43	348.35	348.93	348.25	346.56	344.68	343.09	342.8	344.24	345.55	345.9
346.3	346.96	347.86	349.55	350.21	349.54	347.94	345.9	344.85	344.17	345.66	346.9	347.15
348.02	348.47	349.42	350.99	351.84	351.25	349.52	348.1	346.45	346.36	347.81	348.96	348.93
350.43	351.73	352.22	353.59	354.22	353.79	352.38	350.43	348.72	348.88	350.07	351.34	351.48
352.76	353.07	353.68	355.42	355.67	355.13	353.9	351.67	349.8	349.99	351.29	352.52	352.91
353.66	354.7	355.39	356.2	357.16	356.23	354.82	352.91	350.96	351.18	352.83	354.21	354.19
354.72	355.75	357.16	358.6	359.33	358.24	356.17	354.02	352.15	352.21	353.75	354.99	355.59
355.98	356.72	357.81	359.15	359.66	359.25	357.02	355	353.01	353.31	354.16	355.4	356.37
356.7	357.16	358.38	359.46	360.28	359.6	357.57	355.52	353.69	353.99	355.34	356.8	357.04
358.37	358.91	359.97	361.26	361.68	360.95	359.55	357.48	355.84	355.99	357.58	359.04	358.89
359.97	361	361.64	363.45	363.79	363.26	361.9	359.46	358.05	357.76	359.56	360.7	360.88
362.05	363.25	364.02	364.72	365.41	364.97	363.65	361.48	359.45	359.6	360.76	362.33	362.64
363.18	364	364.56	366.35	366.79	365.62	364.47	362.51	360.19	360.77	362.43	364.28	363.76
365.33	366.15	367.31	368.61	369.3	368.87	367.64	365.77	363.9	364.23	365.46	366.97	366.63
368.15	368.87	369.59	371.14	371	370.35	369.27	366.93	364.63	365.13	366.67	368.01	368.31
369.14	369.46	370.52	371.66	371.82	371.7	370.12	368.12	366.62	366.73	368.29	369.53	369.48
370.28	371.5	372.12	372.87	374.02	373.3	371.62	369.55	367.96	368.09	369.68	371.24	371.02
372.43	373.09	373.52	374.86	375.55	375.41	374.02	371.49	370.7	370.25	372.08	373.78	373.1
374.68	375.63	376.11	377.65	378.35	378.13	376.62	374.5	372.99	373.01	374.35	375.7	375.64
376.79	377.37	378.41	380.52	380.63	379.57	377.79	375.86	374.07	374.24	375.86	377.47	377.38
378.37	379.69	380.41	382.1	382.28	382.13	380.66	378.71	376.42	376.88	378.32	380.04	379.67
381.38	382.03	382.64	384.62	384.95	384.06	382.29	380.47	378.67	379.06	380.14	381.74	381.84
382.45	383.68	384.23	386.26	386.39	385.87	384.39	381.78	380.73	380.81	382.33	383.69	383.55
385.07	385.72	385.85	386.71	388.45	387.64	386.1	383.95	382.91	382.73	383.96	385.02	385.34

Визуализация данных



Визуализация данных

- Частный случай нелинейного понижения размерности
- $d = 2$ или $d = 3$
- Нужно сохранить структуру данных и зависимости

MNIST

A selection from the 64-dimensional digits dataset

0	1	2	3	4	5	0	1	3	4	5	0	5
5	5	0	4	1	3	5	1	0	0	2	2	0
4	4	1	5	0	5	2	2	0	0	1	3	2
3	1	4	0	5	3	1	5	4	4	2	2	5
2	3	4	5	0	1	2	3	4	5	0	5	5
0	4	4	1	3	5	1	0	0	2	2	1	0
1	5	0	5	2	2	0	0	1	3	2	1	3
0	5	3	1	5	4	4	2	2	5	5	4	0
5	0	1	2	3	4	5	0	4	2	3	4	5
3	5	4	0	0	2	2	2	0	1	2	3	3
5	2	2	0	0	4	3	2	4	3	1	4	0
3	8	5	4	4	2	2	2	5	5	4	4	1
0	1	2	3	4	5	0	1	2	3	4	5	0
5	1	0	0	1	2	2	0	1	2	3	3	4
2	2	0	0	1	3	2	4	3	1	3	4	5
1	2	0	0	1	3	2	4	3	1	3	4	5
2	3	4	5	0	1	2	3	4	5	0	5	5
0	0	2	2	2	0	1	2	3	3	3	4	4
0	0	1	3	1	4	3	1	3	1	4	3	1
4	4	2	2	2	5	5	4	4	0	0	1	2
2	3	4	5	0	1	2	3	4	5	0	5	5
0	0	2	2	2	0	1	2	3	3	3	4	4
0	0	1	3	1	4	3	1	3	1	4	0	5
4	4	2	2	2	5	5	4	4	0	0	1	2

t-SNE

- t-Stochastic Neighbor Embedding
- Метод визуализации
- Ищет такие точки на плоскости, которые лучше всего сохраняют расстояния из исходного пространства

t-SNE (t-distributed Stochastic Neighbor Embedding) работает по следующему принципу:

1. Построение вероятностного распределения в высокоразмерном пространстве:

Для каждой точки в высокоразмерном пространстве (например, в пространстве с большим количеством признаков) t-SNE вычисляет вероятности того, что другие точки являются её соседями. Это делается с использованием гауссова распределения.

Для каждой точки i вычисляется вероятность $p_{j|i}$, что точка j является соседом точки i . Это вероятность того, что точка j будет выбрана как сосед точки i при условии, что точка i выбрана. Эта вероятность зависит от расстояния между точками и параметра "перплексности", который контролирует количество соседей.

2. Построение вероятностного распределения в низкоразмерном пространстве:

t-SNE затем создает аналогичное распределение вероятностей для точек в низкоразмерном пространстве (обычно 2D или 3D). Здесь используется t-распределение Стьюдента, которое имеет более тяжелые хвосты по сравнению с гауссовым распределением. Это помогает лучше справляться с разреженными данными и позволяет более эффективно визуализировать кластеры.

3. Минимизация различий между распределениями:

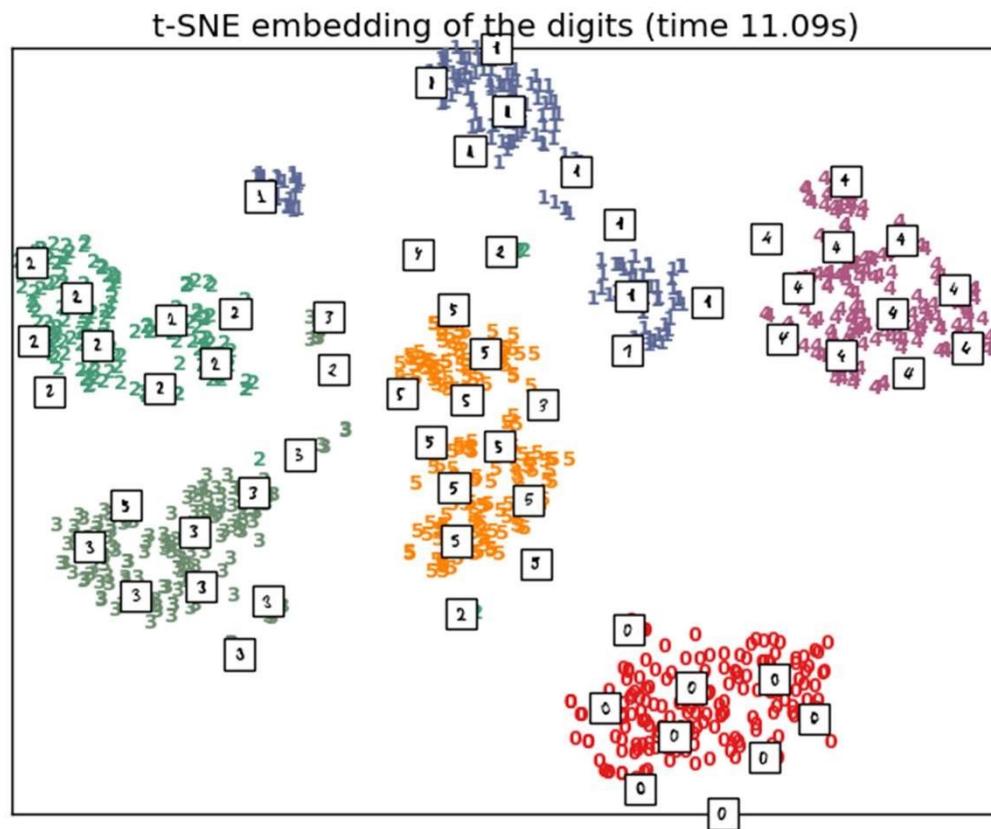
t-SNE использует метод градиентного спуска для минимизации различий между вероятностными распределениями в высокоразмерном и низкоразмерном пространствах. Это достигается путем минимизации функции потерь, которая измеряет, насколько хорошо распределение вероятностей в низкоразмерном пространстве соответствует распределению в высокоразмерном пространстве.

Функция потерь обычно представляет собой Кульбака-Лейблера (Kullback-Leibler divergence), которая измеряет различие между двумя распределениями.

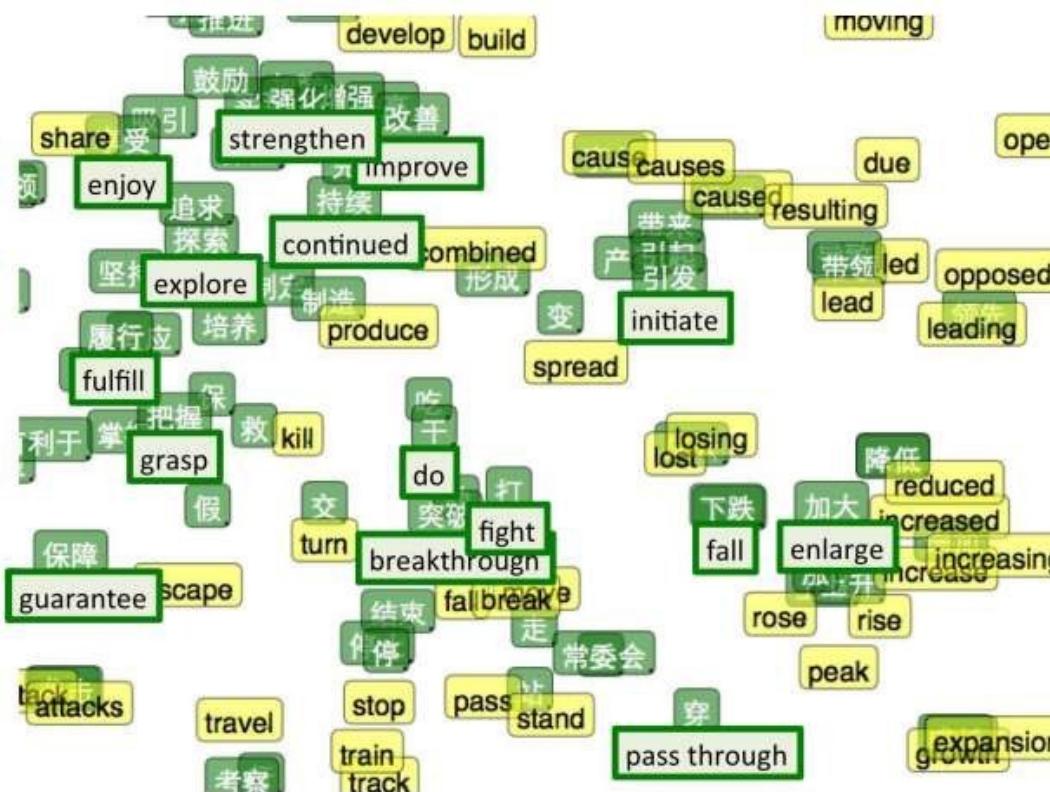
4. Итеративный процесс:

Процесс минимизации выполняется итеративно, и на каждой итерации обновляются координаты точек в низкоразмерном пространстве, чтобы лучше соответствовать структуре данных в высокоразмерном пространстве.

MNIST



Визуализация слов с помощью t-SNE



Резюме

- Методы понижения размерности позволяют убрать неинформативные признаки и ускорить работу над моделями
- Классы методов: отбор признаков и извлечение признаков
- Отбор признаков: фильтрация и использование моделей
- Извлечение признаков: PCA
- Визуализация данных: t-SNE