

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования «Ульяновский государственный университет»

Факультет математики, информационных и авиационных технологий

«Методы классификации без учителя»

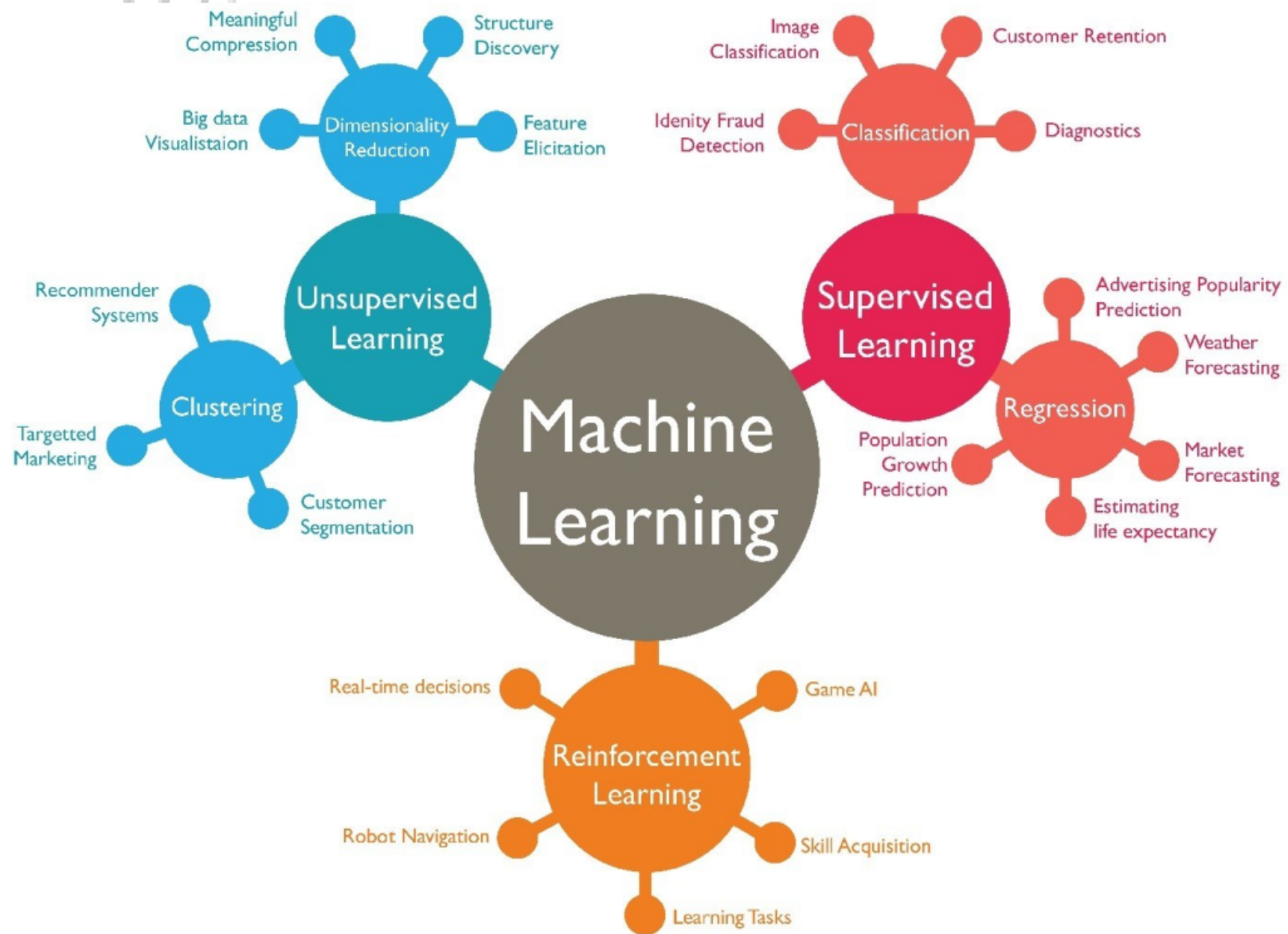
Павлов Павел Юрьевич

Кандидат технических наук

Ульяновск, 2023

Agenda

1. Обучение без учителя
2. Методы классификации без учителя.
3. Метод главных компонент
4. Метод сингулярного разложения
5. TSNE
6. Кластерный анализ
7. Метод иерархической агломерации
8. Метод k-средних
9. Нейронные сети Кохонена
10. DBSCAN
11. Мягкая кластеризация



Сокращение размерности

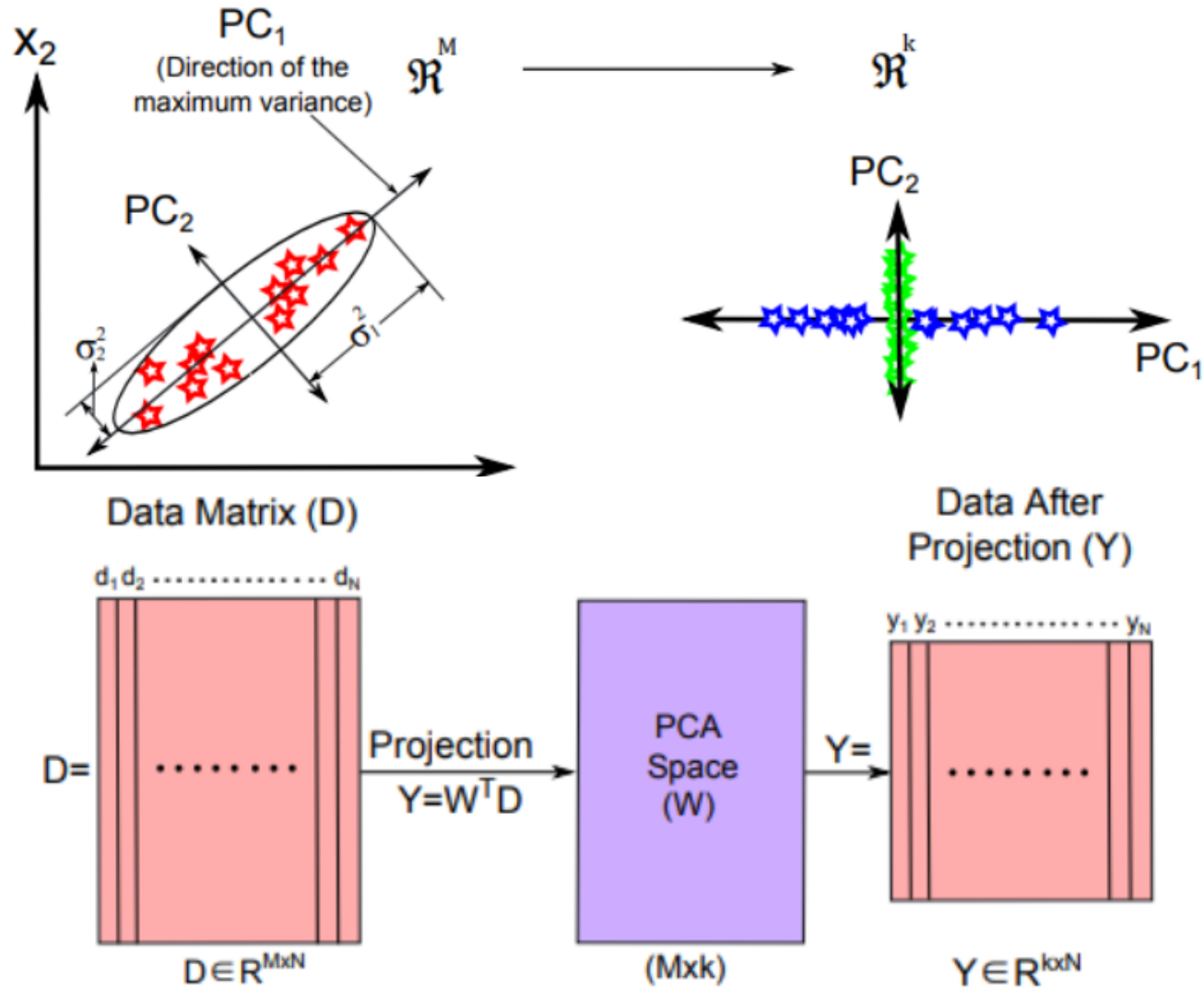
- PCA (principal component analysis) — метод главных компонент
- SVD (Singular Value Decomposition) — метод сингулярного разложения
- TSNE (t-distributed stochastic neighbor embedding) — стохастическое вложение соседей с t-распределением

Метод главных компонент (МГК)

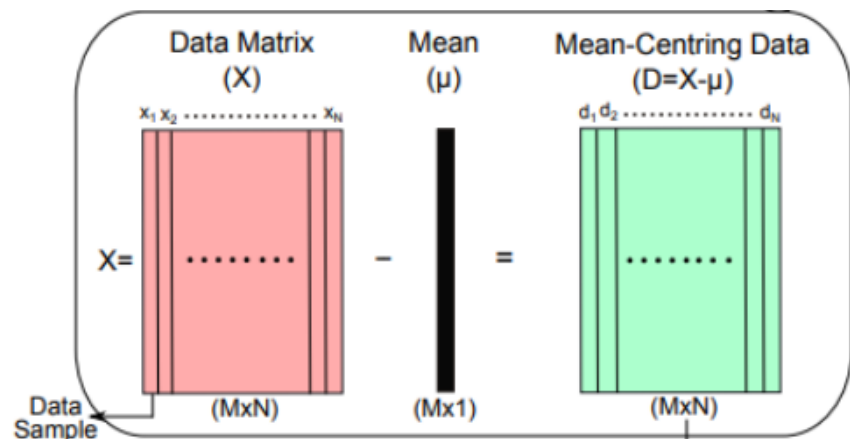
Метод главных компонент (МГК) применяется для снижения размерности пространства наблюдаемых векторов, не приводя к существенной потере информативности. Предпосылкой МГК является нормальный закон распределения многомерных векторов. В МГК линейные комбинации случайных величин определяются характеристическими векторами ковариационной матрицы. Главные компоненты представляют собой ортогональную систему координат, в которой дисперсии компонент характеризуют их статистические свойства. МГК не относят к ФА, хотя он имеет схожий алгоритм и решает схожие аналитические задачи. Его главное отличие заключается в том, что обработке подлежит не редуцированная, а обычная матрица парных корреляций, ковариаций, на главной диагонали которой расположены единицы.

Пусть дан исходный набор векторов X линейного пространства L^k . Применение метода главных компонент позволяет перейти к базису пространства L_m ($m \leq k$), такому что: первая компонента (первый вектор базиса) соответствует направлению, вдоль которого дисперсия векторов исходного набора максимальна. Направление второй компоненты (второго вектора базиса) выбрано таким образом, чтобы дисперсия исходных векторов вдоль него была максимальной при условии ортогональности первому вектору базиса. Аналогично определяются остальные векторы базиса. В результате, направления векторов базиса выбраны так, чтобы максимизировать дисперсию исходного набора вдоль первых компонент, называемых главными компонентами (или главными осями). Получается, что основная изменчивость векторов исходного набора векторов представлена несколькими первыми компонентами, и появляется возможность, отбросив менее существенные компоненты, перейти к пространству меньшей размерности.

Постановка задачи метода главных компонент



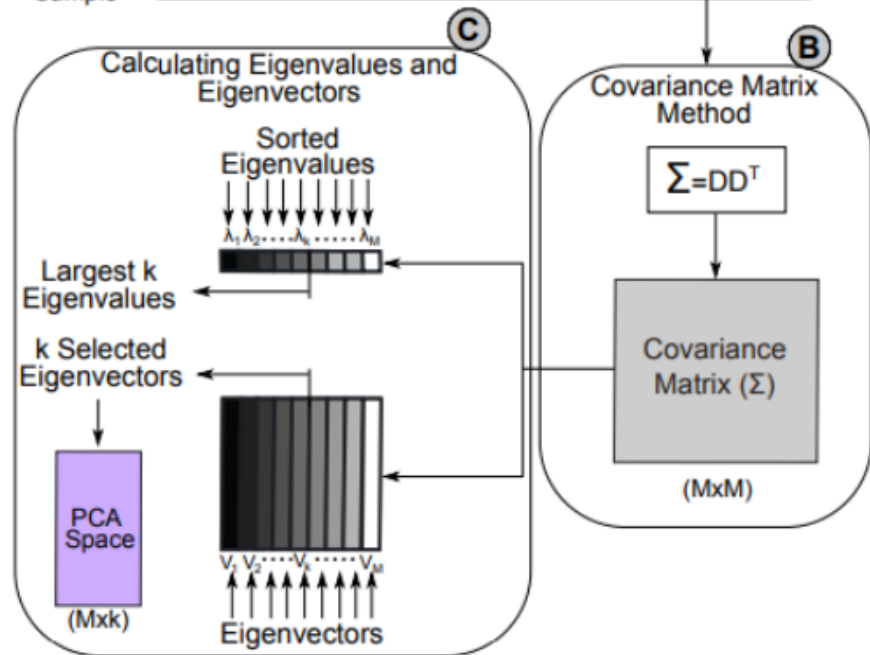
Работа метода главных компонент через матрицу ковариации



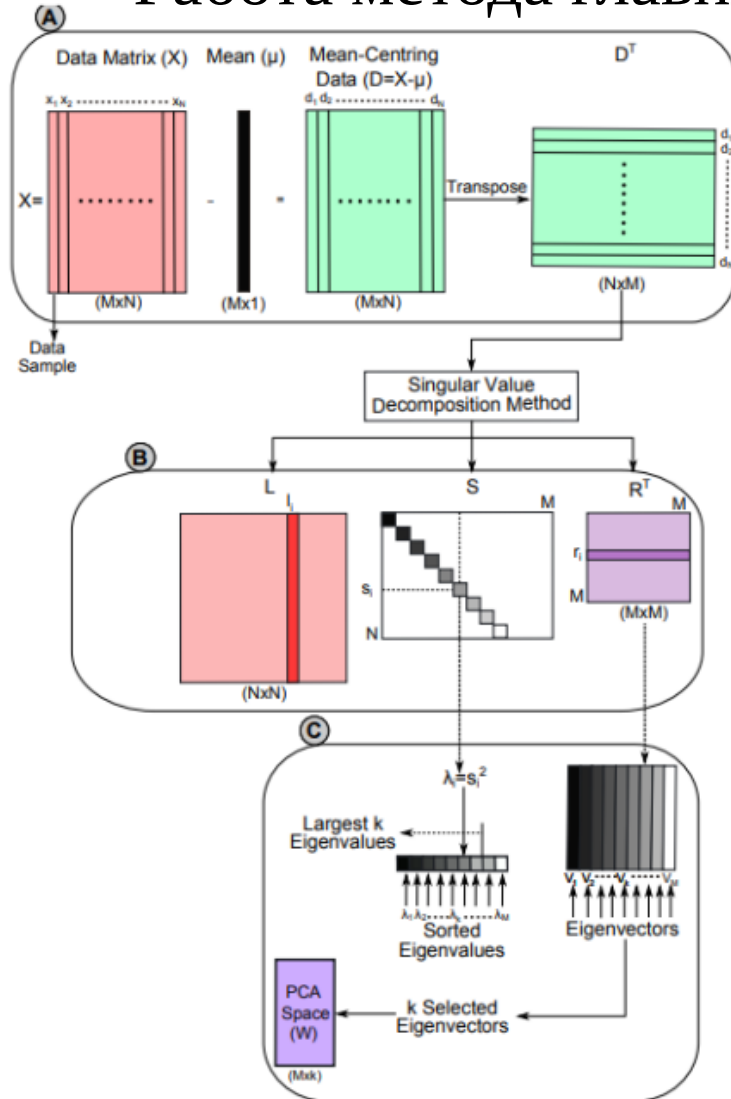
$$\sigma^2(x) = \text{Var}(x) = E((x - \mu)^2) = E\{x^2\} - (E\{x\})^2$$

$$\begin{aligned} \Sigma_{ij} &= E\{x_i x_j\} - E\{x_i\}E\{x_j\} = \\ &= E[(x_i - \mu_i)(x_j - \mu_j)] \end{aligned}$$

$$V\Sigma = \lambda V$$

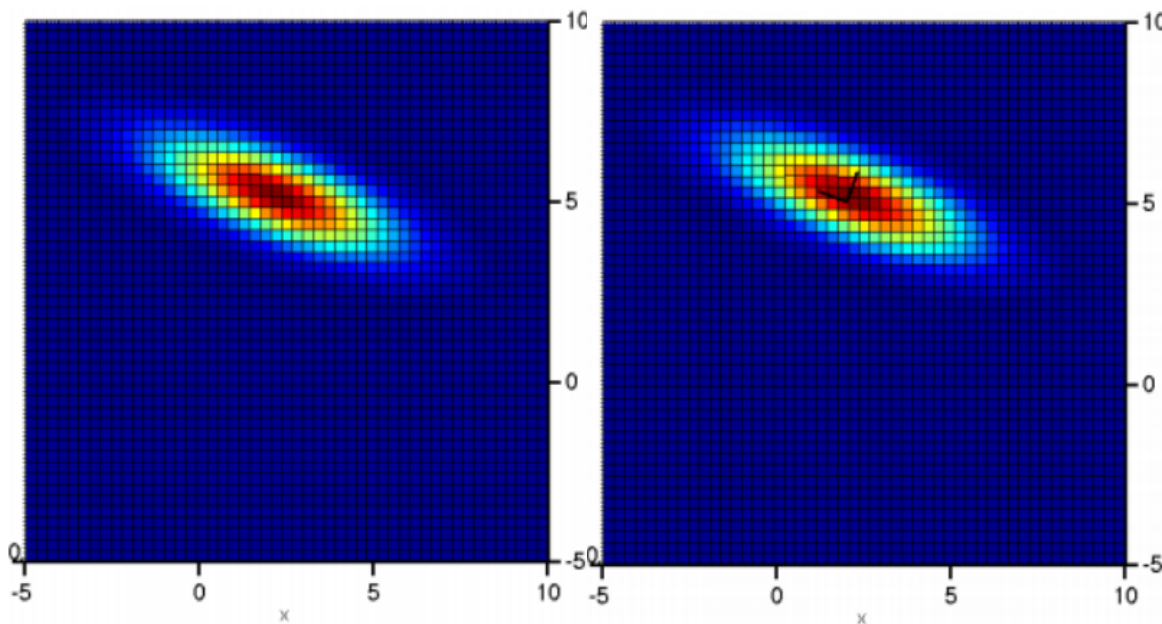


Работа метода главных компонент через SVD



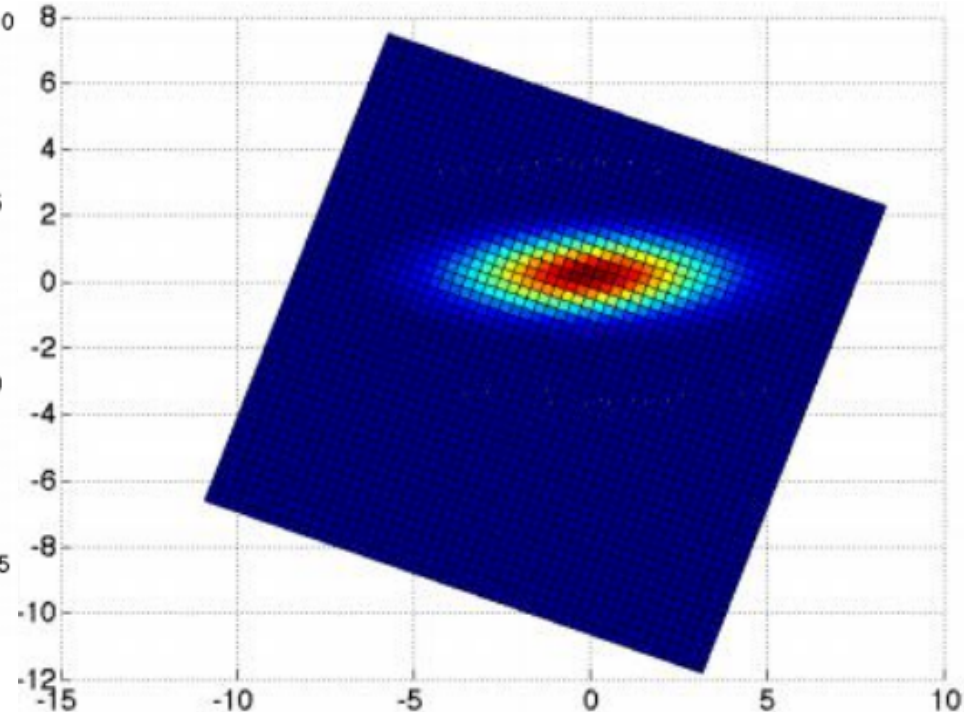
$$X = LSR^T = \begin{bmatrix} l_1 & \dots & l_p \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & s_q \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -r_1^T \\ -r_2^T \\ \vdots \\ -r_q^T \end{bmatrix}$$

Суть работы метода главных компонент



Исходное облако точек
(гауссово
распределение)

Исходное облако
точек с
собственными
векторами



Проекция точек на две
главные компоненты

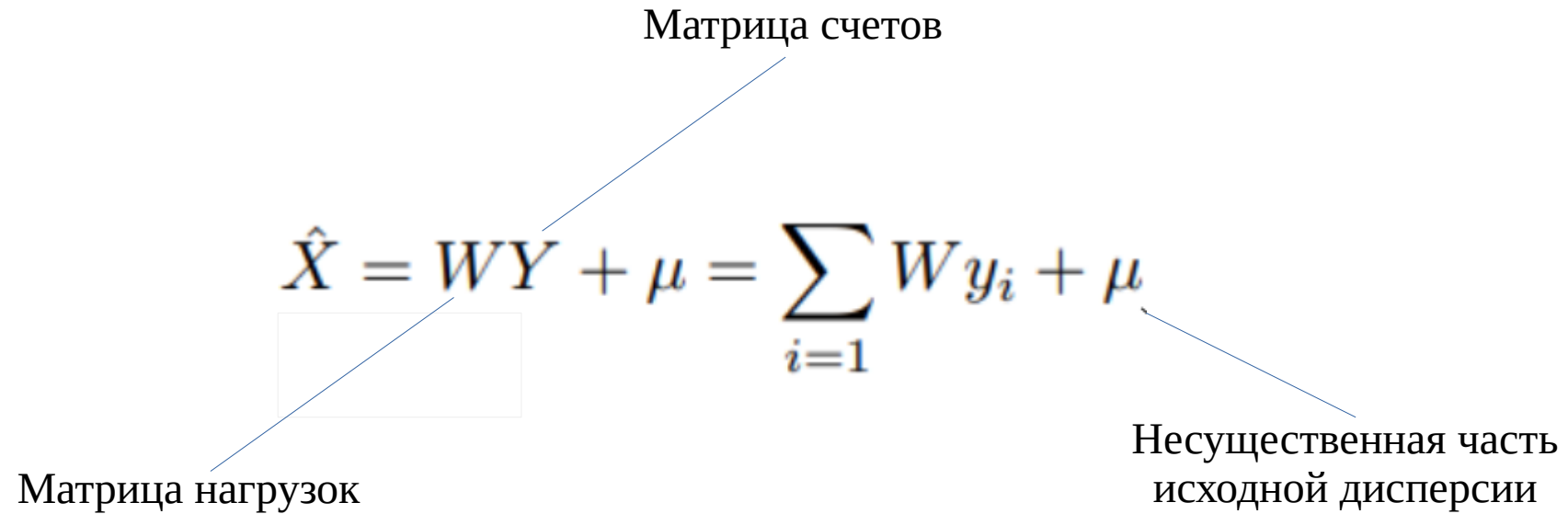
Результаты работы метода главных компонент

Матрица счетов

$$\hat{X} = WY + \mu = \sum_{i=1} W y_i + \mu$$

Матрица нагрузок

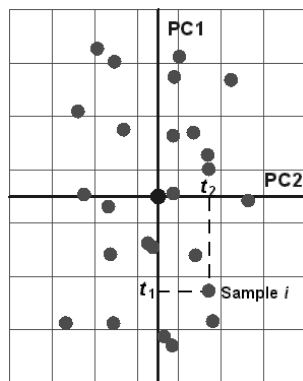
Несущественная часть
исходной дисперсии



Метод главных компонент. Матрица счетов

Матрица счетов \mathbf{T} дает нам проекции исходных образцов (J – мерных векторов $\mathbf{x}_1, \dots, \mathbf{x}_J$) на подпространство главных компонент (A -мерное). Строки $\mathbf{t}_1, \dots, \mathbf{t}_I$ матрицы \mathbf{T} – это координаты образцов в новой системе координат. Столбцы $\mathbf{t}_1, \dots, \mathbf{t}_A$ матрицы \mathbf{T} – ортогональны и представляют проекции всех образцов на одну новую координатную ось.

При исследовании данных методом РСА, особое внимание уделяется графикам счетов. Они несут в себе информацию, полезную для понимания того, как устроены данные. На графике счетов каждый образец изображается в координатах $(\mathbf{t}_i, \mathbf{t}_j)$, чаще всего – $(\mathbf{t}_1, \mathbf{t}_2)$, обозначаемых РС1 и РС2. Близость двух точек означает их схожесть, т. е. положительную корреляцию. Точки, расположенные под прямым углом, являются некоррелированными, а расположенные диаметрально противоположно – имеют отрицательную корреляцию.

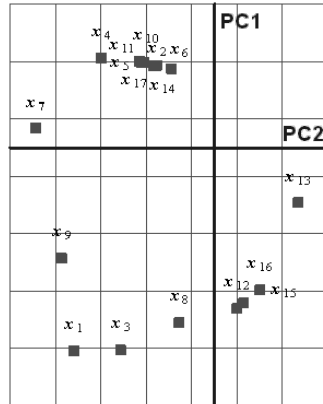


$$\mathbf{T} = \begin{matrix} & \begin{matrix} t_{11} & t_{12} & \dots & t_{1a} & \dots & t_{1A} \\ t_{21} & t_{22} & \dots & t_{2a} & \dots & t_{2A} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ t_{i1} & t_{i2} & \dots & t_{ia} & \dots & t_{iA} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ t_{I1} & t_{I2} & \dots & t_{Ia} & \dots & t_{IA} \end{matrix} \\ \begin{matrix} \vdots \\ \vdots \\ \vdots \end{matrix} & \end{matrix}$$

Метод главных компонент. Матрица нагрузок

Матрица нагрузок \mathbf{P} – это матрица перехода из исходного пространства переменных x_1, \dots, x_J (J -мерного) в пространство главных компонент (A -мерное). Каждая строка матрицы \mathbf{P} состоит из коэффициентов, связывающих переменные \mathbf{t} и \mathbf{x} . Например, a -я строка – это проекция всех переменных x_1, \dots, x_J на a -ю ось главных компонент. Каждый столбец \mathbf{P} – это проекция соответствующей переменной x_j на новую систему координат.

График нагрузок применяется для исследования роли переменных. На этом графике каждая переменная x_j отображается точкой в координатах (p_{1j}, p_{2j}) , например (p_{11}, p_{21}) . Анализируя его аналогично графику счетов, можно понять, какие переменные связаны, а какие независимы. Совместное исследование парных графиков счетов и нагрузок, также может дать много полезной информации о данных.



$$\mathbf{P} = \begin{array}{cc|cc} p_{11} & p_{12} & p_{1j} & p_{1J} \\ p_{21} & p_{22} & p_{2j} & p_{2J} \\ \dots & \dots & \dots & \dots \\ p_{a1} & p_{a2} & p_{aj} & p_{aJ} \\ \dots & \dots & \dots & \dots \\ p_{A1} & p_{A2} & p_{Aj} & p_{AJ} \end{array}$$

Метод главных компонент. Вращение факторов

Вращение — это способ превращения факторов, полученных на предыдущем этапе, в более осмысленные. Вращение делится на:

- графическое (проведение осей, не применяется при более чем двухмерном анализе),
- аналитическое (выбирается некий критерий вращения, различают ортогональное и косоугольное) и
- матрично-приближенное (вращение состоит в приближении к некой заданной целевой матрице).

Результатом вращения является вторичная структура факторов. Первичная факторная структура (состоящая из первичных нагрузок (полученных на предыдущем этапе) - это, фактически, проекции точек на ортогональные оси координат. Очевидно, что если проекции будут нулевыми, то структура будет проще. А проекции будут нулевыми, если точка лежит на какой-то оси. Таким образом, можно считать вращение переходом от одной системы координат к другой при известных координатах в одной системе (первичные факторы) и итеративно подбираемых координатах в другой системе (вторичные факторы). При получении вторичной структуры стремятся перейти к такой системе координат, чтобы провести через точки (объекты) как можно больше осей, чтобы как можно больше проекции (и соответственно нагрузок) были нулевыми. При этом могут сниматься ограничения ортогональности и убывания значимости от первого к последнему факторам, характерные для первичной структуры.

Метод главных компонент. Методы вращения матрицы нагрузок

$$V = A\Lambda$$

Вращение исходной матрицы главных компонент A

$$K = \sum_{i=1}^m \sum_{p=1}^g v_{ip}^4$$

Критерий вращения квартимакс

$$K = n \sum_{p=1}^g \sum_{i=1}^m \left(\frac{v_{ip}}{h_i} \right)^2 - \sum_{p=1}^g \left(\sum_{i=1}^m \frac{v_{ip}^2}{h_i^2} \right)^2$$

Критерий вращения варимакс

Метод главных компонент. Особенности метода главных компонент

- допущение о том, что размерность данных может быть эффективно понижена путем линейного преобразования;
- допущение о том, что больше всего информации несут те направления, в которых дисперсия входных данных максимальна.
- направления, максимизирующие дисперсию, далеко не всегда максимизируют информативность

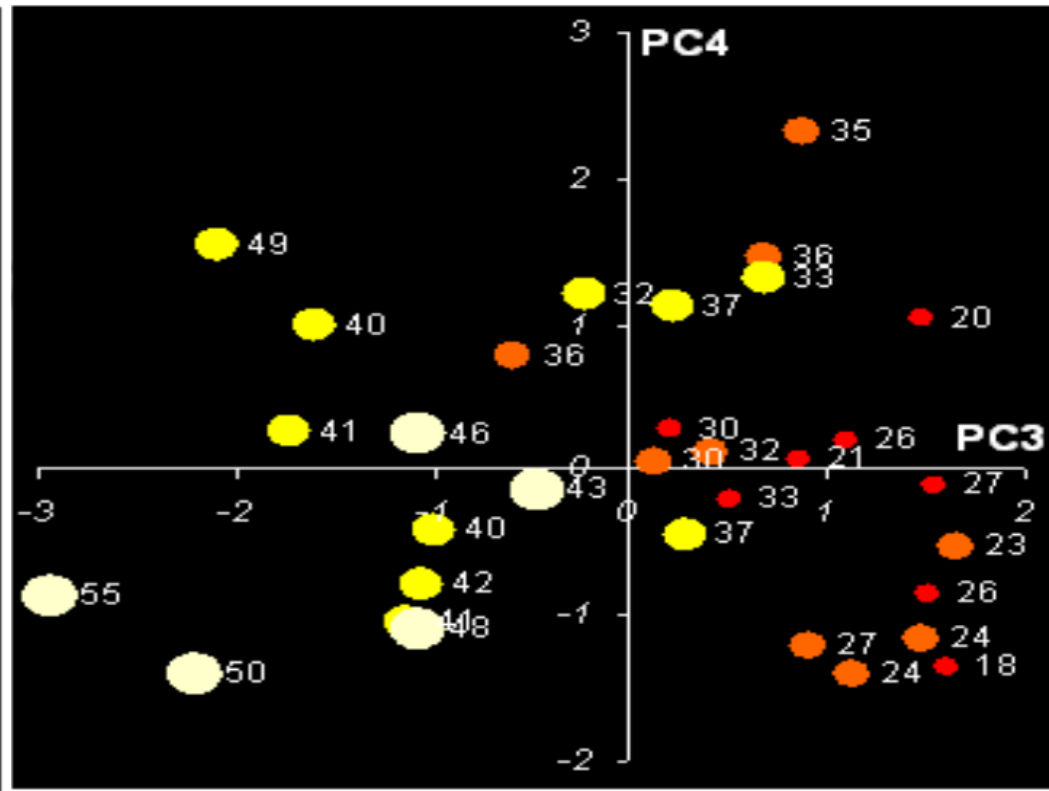
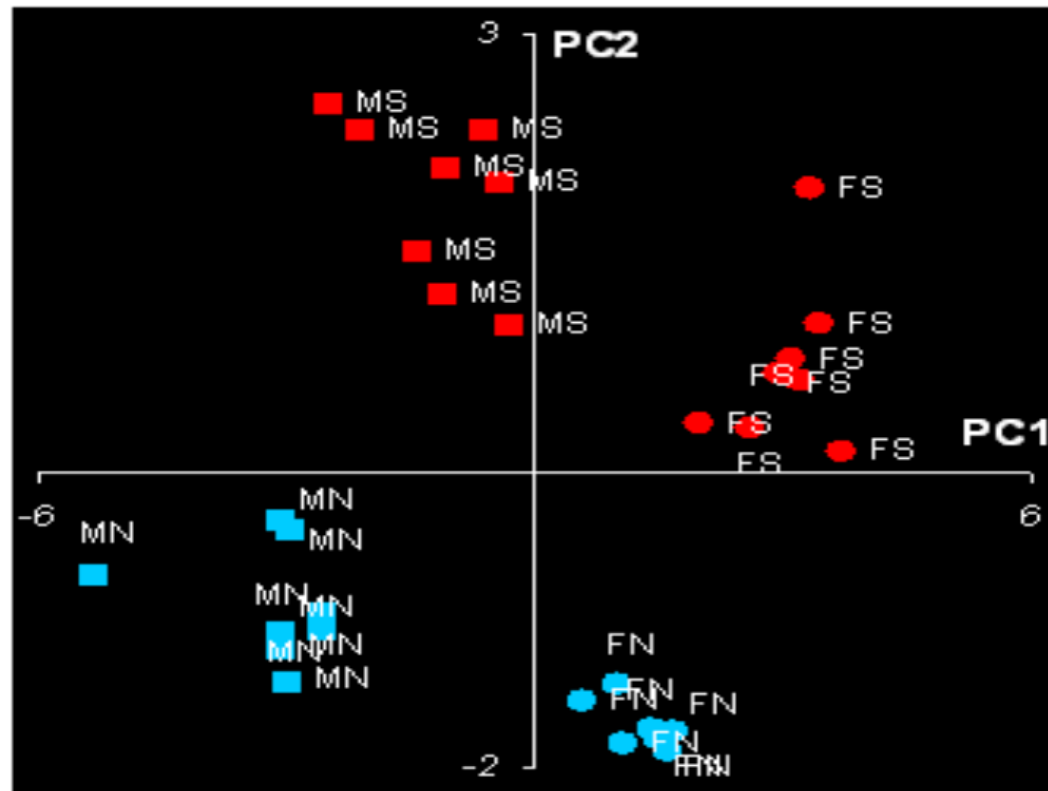
Метод главных компонент. Пример. Матрица счетов

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		PCA												
2		T Scores												
3		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	
4	1	=ScoresPCA(Xraw,12,3)				1.060	-0.017	0.563	0.085	0.280	0.078	0.109	0.133	
5	2	-3.114	-0.293	0.671	1.310	0.435	0.119	0.109	-0.456	-0.088	-0.034	0.051	-0.009	
6	3	-2.997	-0.360	0.212	1.117	0.204	-0.017	0.120	-0.469	-0.149	-0.018	0.184	-0.208	
7	4	-2.591	-0.928	0.863	2.321	-0.095	-0.076	-0.270	0.306	0.240	-0.192	-0.168	0.005	
8	5	-2.588	-1.037	0.686	1.461	-0.347	-0.098	-0.447	0.345	0.083	0.021	-0.043	-0.005	
9	6	-3.027	-1.400	0.284	-0.440	-0.633	-0.538	0.252	-0.343	0.042	-0.132	-0.247	0.020	
10	7	-3.095	-1.069	-0.472	-0.144	-0.304	-0.059	-0.147	-0.025	-0.217	0.150	-0.077	-0.088	
11	8	-3.113	-1.181	-1.068	0.242	-0.232	0.141	-0.211	-0.008	-0.038	0.203	-0.060	0.087	
12	9	-2.130	2.356	1.503	-0.858	0.284	0.070	0.020	0.246	-0.171	0.095	-0.123	0.060	
13	10	-2.510	2.525	1.542	-0.105	0.653	-0.087	0.222	0.228	-0.305	0.078	0.024	0.035	
14	11	-0.463	1.992	1.090	0.206	-0.617	0.084	0.138	-0.186	-0.202	-0.421	0.069	0.174	
15	12	-1.098	2.094	0.496	-0.198	-0.376	0.127	-0.126	0.299	-0.071	-0.122	0.112	-0.115	
16	13	-1.438	1.527	-1.059	-0.788	-0.734	-0.124	0.015	-0.060	0.264	0.185	0.156	-0.057	
17	14	-1.137	1.241	-2.200	-1.404	-0.953	0.200	0.033	0.063	-0.075	-0.002	-0.058	-0.009	
18	15	-0.334	1.034	-2.931	-0.866	0.420	-0.654	-0.781	0.037	0.016	-0.283	0.060	-0.004	
19	16	-0.652	2.360	0.125	0.055	-0.334	0.674	-0.398	-0.283	0.359	0.192	-0.007	-0.007	
20	17	1.084	-1.845	0.409	0.123	-1.323	0.872	0.704	0.328	0.032	-0.029	0.078	-0.026	
21	18	0.981	-1.434	1.645	-0.526	0.714	-0.035	-0.096	0.127	-0.005	0.003	-0.124	-0.053	
22	19	0.567	-1.551	1.474	-1.154	0.677	-0.234	-0.047	0.031	0.068	0.213	-0.058	-0.052	
23	20	1.663	-1.762	1.122	-1.394	0.018	-0.081	0.085	-0.218	0.224	-0.284	0.028	0.004	
24	21	1.486	-1.813	0.904	-1.208	0.070	-0.147	-0.003	-0.023	0.043	-0.160	0.109	-0.034	
25	22	2.464	-2.040	-0.222	1.202	-0.140	0.268	-0.491	-0.185	-0.091	0.166	0.191	0.233	
26	23	1.396	-1.736	-1.130	-1.041	0.367	0.487	-0.170	0.133	-0.220	-0.007	0.123	-0.074	
27	24	1.622	-1.894	-0.992	-0.419	0.287	0.463	-0.204	0.141	-0.225	-0.048	-0.033	0.015	
28	25	2.005	0.344	-2.085	1.549	0.603	-0.377	0.509	0.471	0.105	-0.061	0.070	0.020	
29	26	3.335	1.956	0.851	0.063	0.884	0.897	-0.141	-0.048	0.252	-0.128	-0.057	-0.046	
30	27	3.711	0.147	0.195	0.279	-0.774	-0.624	-0.091	0.120	0.002	0.081	-0.114	-0.074	
31	28	3.207	0.630	1.602	-1.358	-0.420	-0.480	-0.057	-0.047	-0.001	0.168	-0.026	0.112	
32	29	3.423	1.021	1.482	1.036	0.016	-0.395	-0.084	-0.024	-0.035	0.124	0.177	-0.027	
33	30	2.615	0.320	-0.600	0.784	-0.038	-0.789	0.455	-0.109	0.045	0.093	0.108	-0.007	
34	31	2.958	0.688	-1.728	0.267	0.268	0.181	0.320	-0.246	-0.082	0.016	-0.249	-0.012	
35	32	3.102	0.788	-1.603	0.988	0.359	0.249	0.220	-0.232	-0.078	0.055	-0.203	0.009	
36														

Метод главных компонент. Пример. Матрица нагрузок

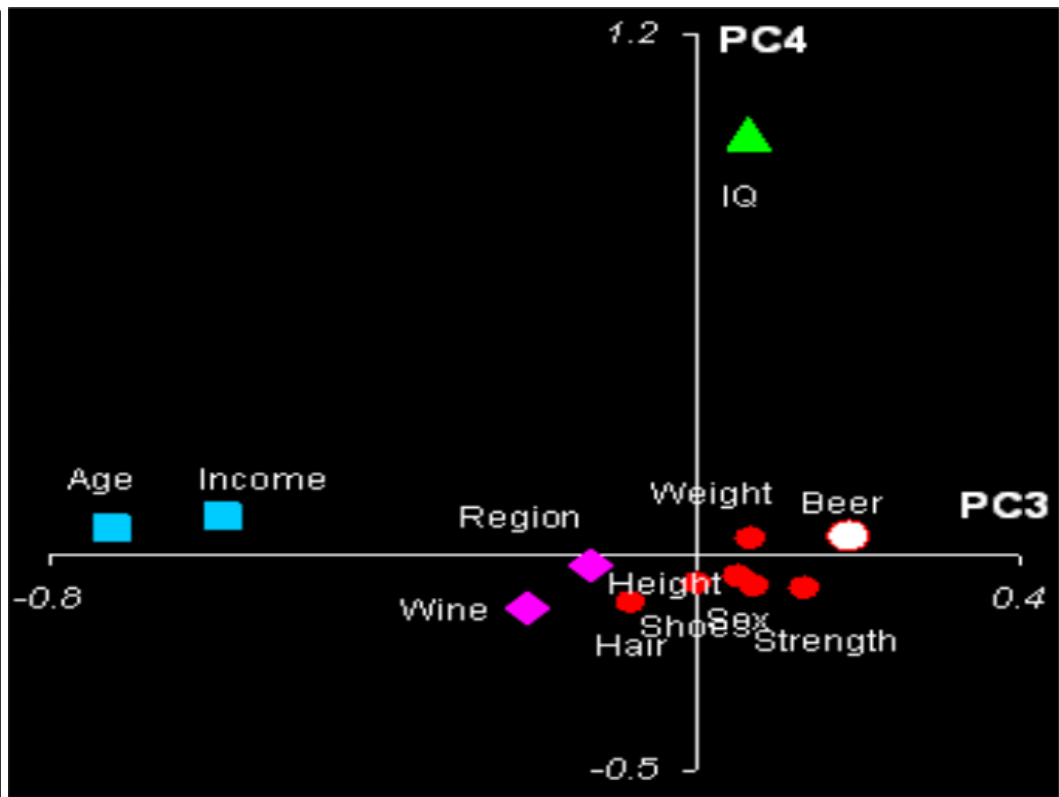
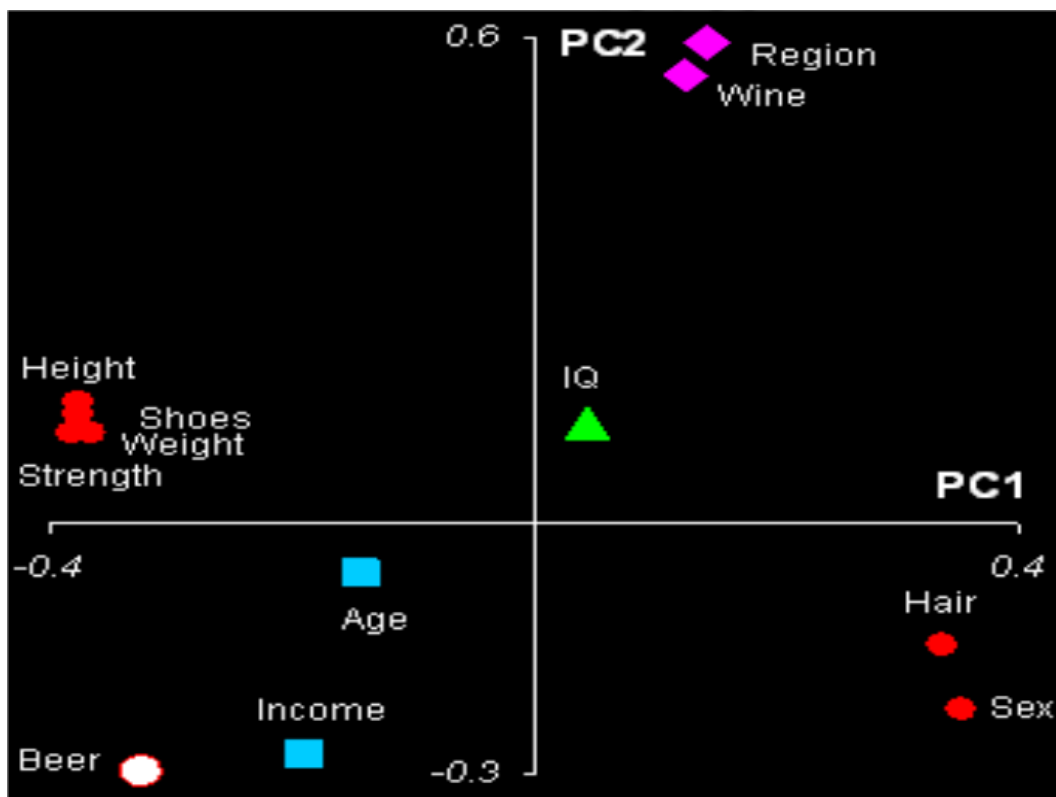
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
2			P Loadings												
3			PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	
4		Height	=LoadingsPCA(Xraw,12,3)				0.186	-0.124	0.268	0.118	0.729	-0.307	0.255	-0.065	
5		Weight	-0.381	0.111	0.068	0.033	0.100	-0.192	-0.224	-0.219	0.190	0.572	-0.415	-0.395	
6		Hair	0.338	-0.150	-0.079	-0.114	0.660	-0.489	-0.368	-0.081	0.041	-0.140	0.007	0.078	
7		Shoes	-0.378	0.151	0.001	-0.066	0.152	-0.031	-0.234	0.171	-0.280	0.376	0.685	0.183	
8		Age	-0.143	-0.061	-0.720	0.055	-0.029	-0.165	0.043	0.435	-0.174	-0.134	-0.036	-0.430	
9		Income	-0.190	-0.287	-0.586	0.085	0.063	0.137	0.129	-0.434	0.180	0.167	-0.038	0.492	
10		Beer	-0.325	-0.308	0.188	0.040	0.231	0.239	-0.170	0.567	-0.015	-0.049	-0.420	0.350	
11		Wine	0.124	0.554	-0.212	-0.125	0.415	0.638	-0.120	-0.040	0.024	-0.054	-0.095	-0.093	
12		Sex	0.352	-0.232	0.052	-0.051	0.313	0.098	0.580	0.254	0.078	0.529	0.084	-0.124	
13		Strength	-0.365	0.112	0.135	-0.081	0.336	-0.160	0.512	-0.258	-0.530	-0.232	-0.165	-0.001	
14		Region	0.144	0.595	-0.130	-0.022	-0.151	-0.402	0.161	0.265	0.050	0.180	-0.255	0.476	
15		IQ	0.044	0.123	0.062	0.969	0.180	-0.010	0.024	0.001	-0.006	-0.033	0.076	-0.010	
16															

Объекты выборки в пространстве новых компонент



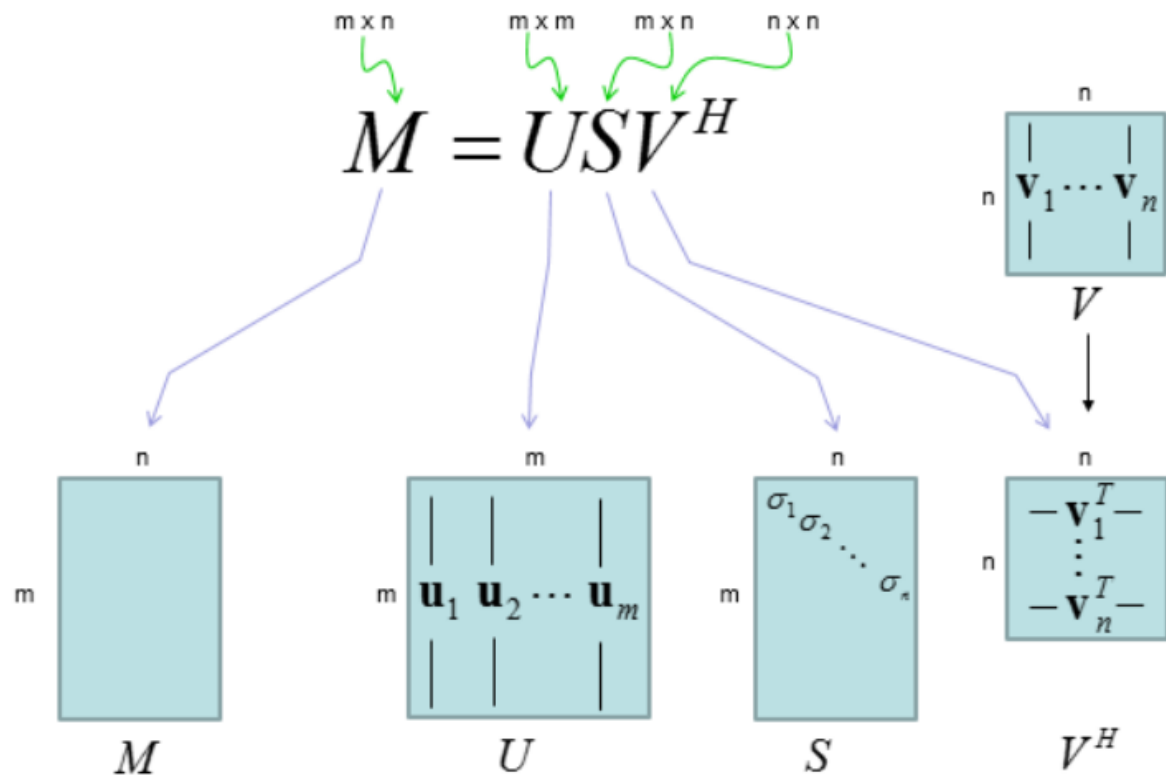
Женщины (F) обозначены кружками ● и ●, а мужчины (M) – квадратами ■ и ■. Север (N) представлен голубым ■, а юг (S) – красным ●.

Исходные переменные в пространстве новых компонент



Метод сингулярного разложения (Singular Value Decomposition)

Сингулярное разложение — определённого типа разложение прямоугольной матрицы. Имеющее широкое применение, в силу своей наглядной геометрической интерпретации, при решении многих прикладных задач.

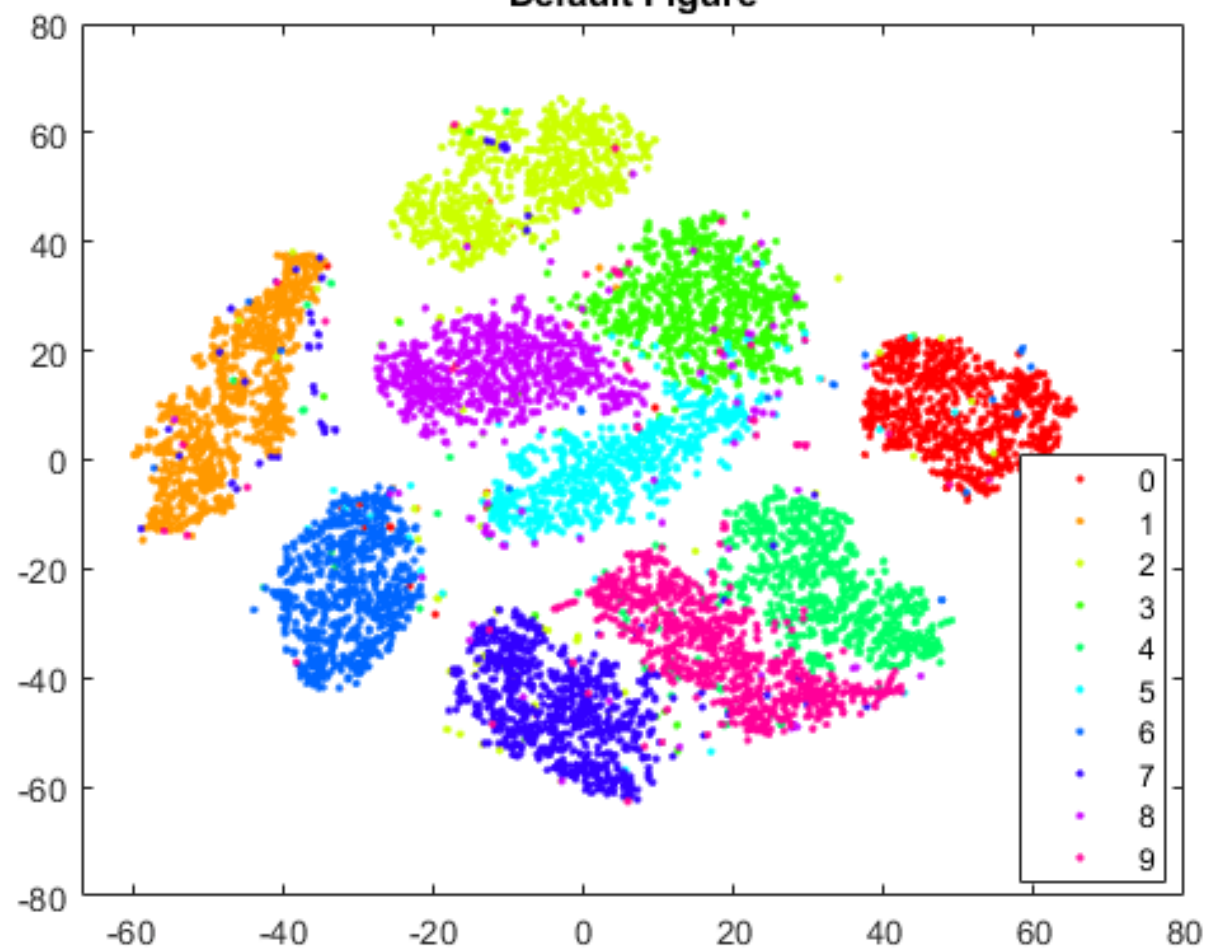


TSNE. Возможное применение

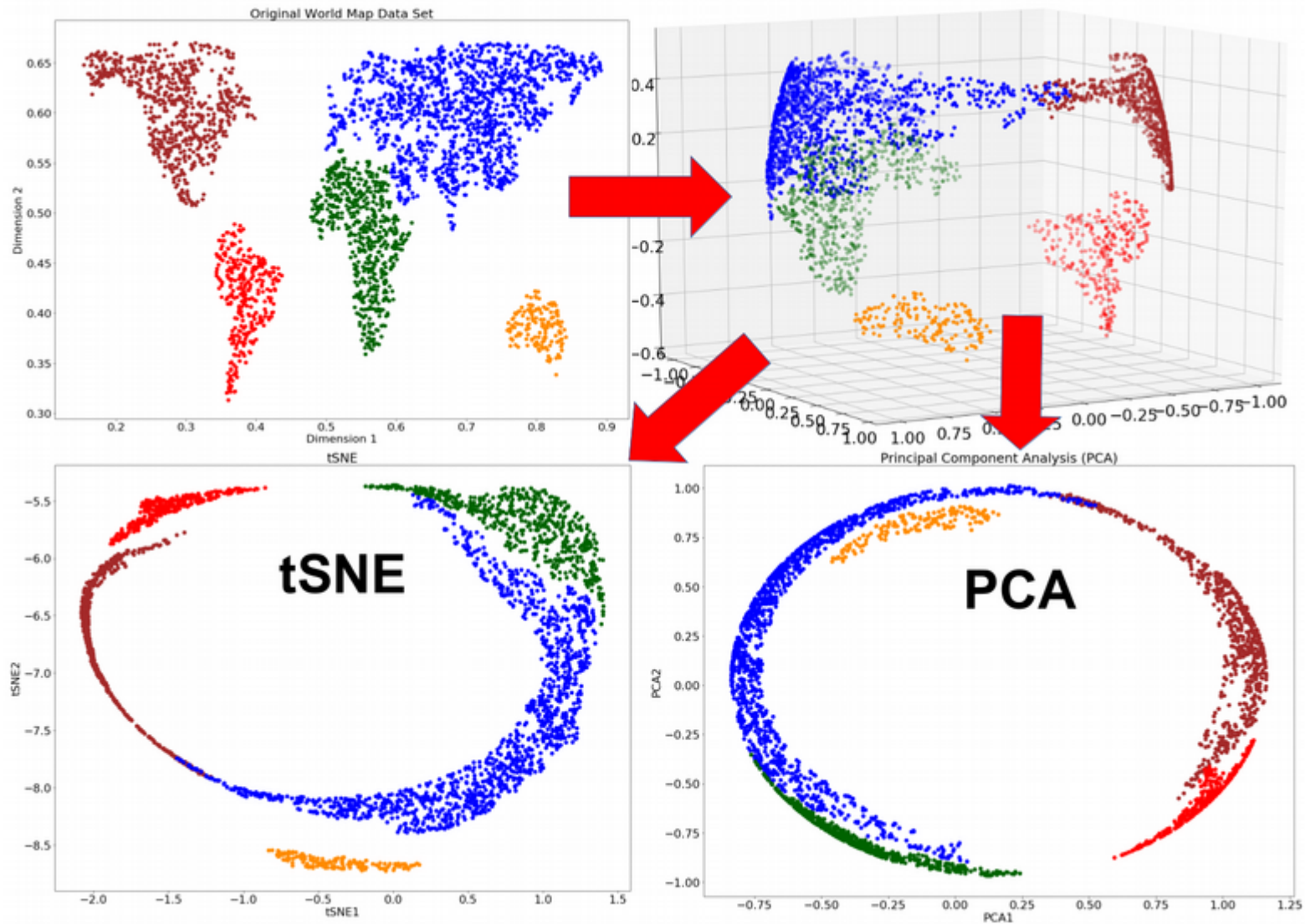
Стохастическое вложение соседей с t-распределением (англ. t-distributed Stochastic Neighbor Embedding, t-SNE) — это алгоритм машинного обучения для визуализации, является техникой нелинейного снижения размерности, хорошо подходящей для вложения данных высокой размерности для визуализации в пространство низкой размерности (двух- или трехмерное). В частности, метод моделирует каждый объект высокой размерности двух- или трёхмерной точкой таким образом, что похожие объекты моделируются близко расположенными точками, а непохожие точки моделируются с большой вероятностью точками, далеко друг от друга отстоящими.

TSNE

Default Figure



TSNE



TSNE. Возможное применение

- Снижение размерности исходной задачи
- Поиск латентных факторов
- Избавление от мультиколлинеарности в данных
- Поиск топиков (тем) в текстовых документах
- Улучшение интерпретируемости моделей классификации и кластеризации

Кластерный анализ

Кластерный анализ (кластер в переводе с лат. означает скопление или гроздь) является совокупностью методов позволяющих исследователю производить т.н. классификацию «без учителя», т. е. разбивать множества объектов на группы, сходные по свойствам, не имея исходных представлений о структуре таких групп. Разбиение множества происходит, таким образом, чтобы элементы, входящие в одну группу были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга.

Кластерный анализ

Кластеризация применяется для решения таких задач как:

- Изучение данных
- Облегчение анализа
- Сжатие данных
- Прогнозирование
- Обнаружение аномалий

Кластерный анализ в теории

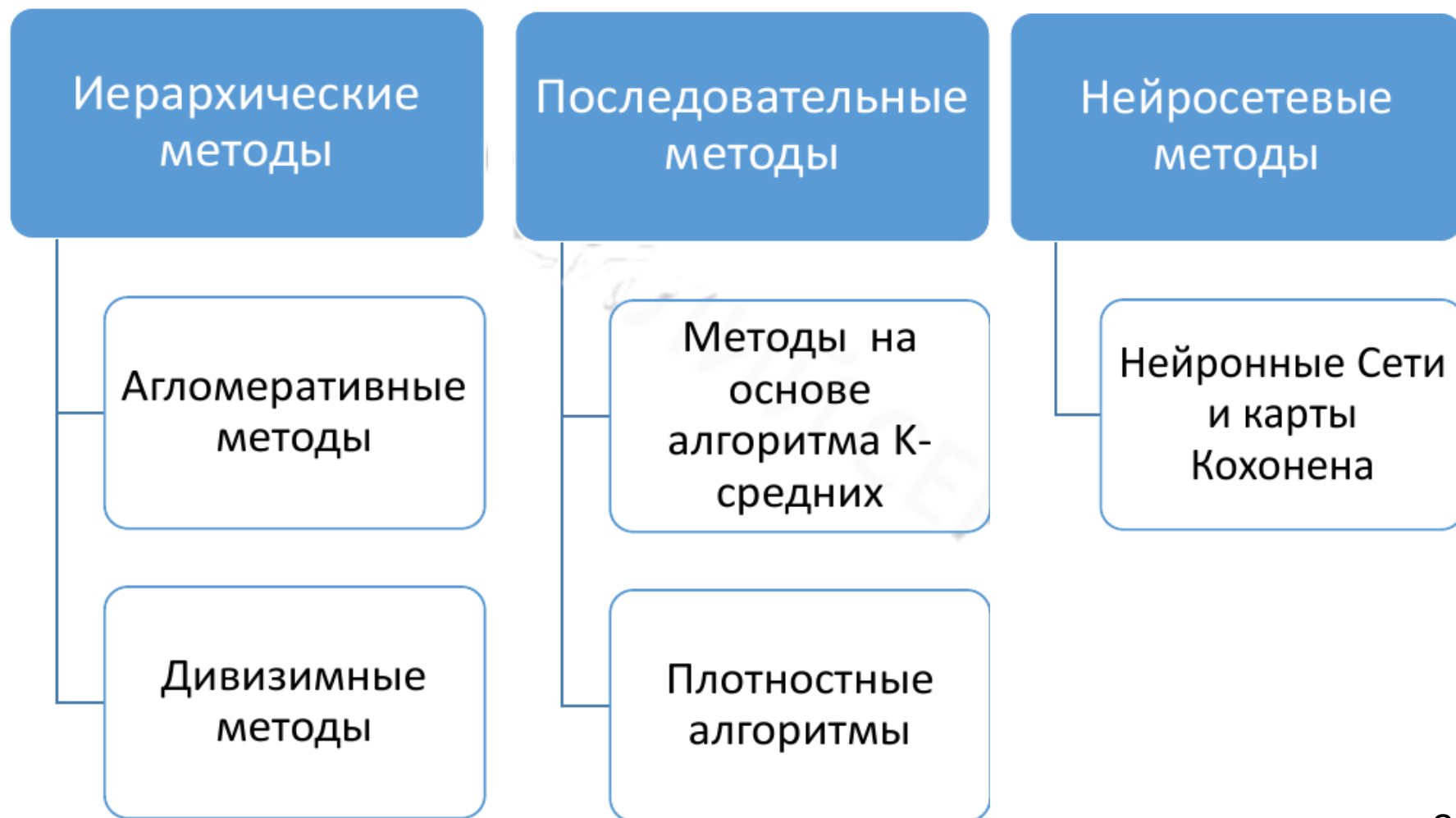


Кластерный анализ на практике

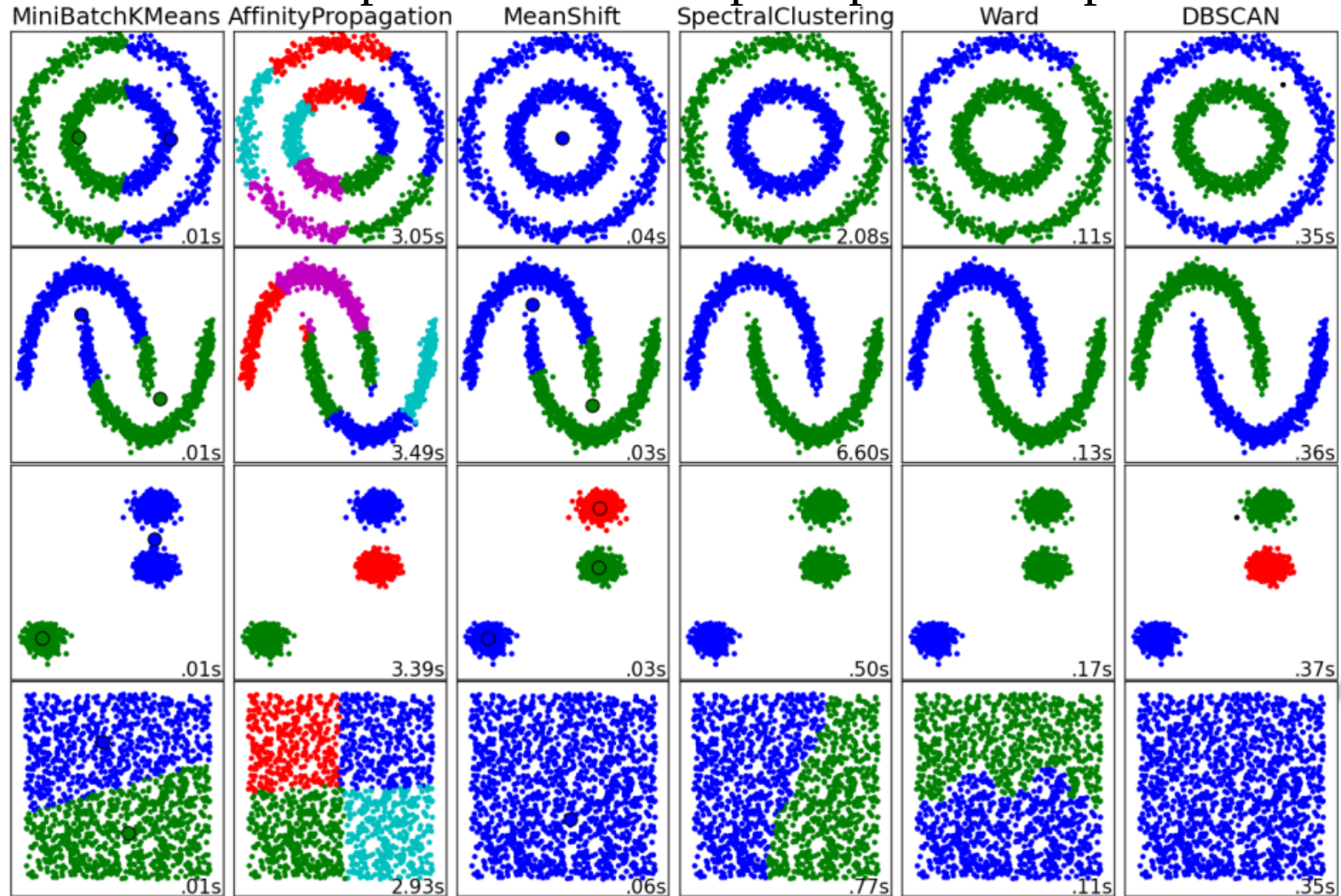


Как очертить границу кластеров? Сколько их следует выделить?

Кластерный анализ. Классификация методов кластерного анализа



Кластерный анализ. Примеры кластеров



Кластерный анализ. Меры расстояния и сходства

$$\text{расстояние}(x,y) = \{ \sum_i (x_i - y_i)^2 \}^{1/2}$$

Евклидово расстояние

$$\text{расстояние}(x,y) = \sum_i (x_i - y_i)^2$$

Квадрат евклидова расстояния

$$\text{расстояние}(x,y) = \sum_i |x_i - y_i|$$

Расстояние городских кварталов (манхэттенское расстояние)

$$\text{расстояние}(x,y) = \text{Максимум}|x_i - y_i|$$

Расстояние Чебышева

$$\text{расстояние}(x,y) = (\sum_i |x_i - y_i|^p)^{1/p}$$

Степенное расстояние

$$\text{расстояние}(x,y) = (\text{Количество } x_i \neq y_i) / i$$

Процент несогласия

Метод иерархической агломерации

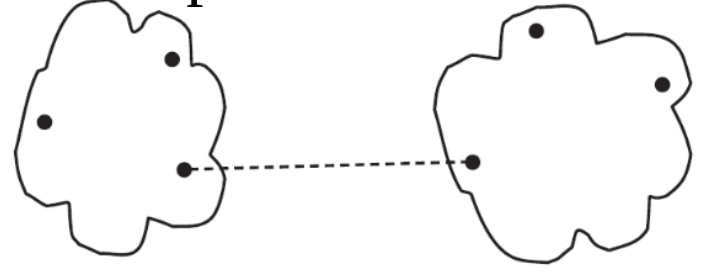
- Метод иерархической агломерации заключается в последовательном объединении N исходных объектов до момента, пока все они не будут объединены в один кластер объёма N .
- С учётом того, что на каждом шаге подвергаются слиянию только два кластера, процедура содержит $N-1$ шагов объединения.

Методы определения расстояния между кластерами

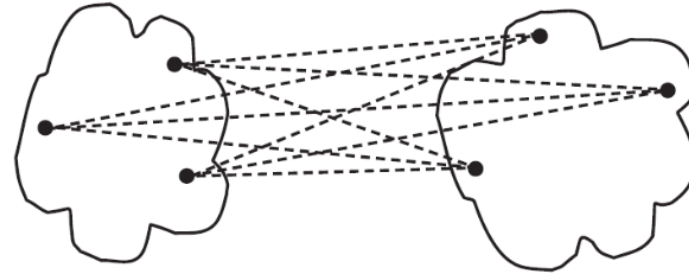
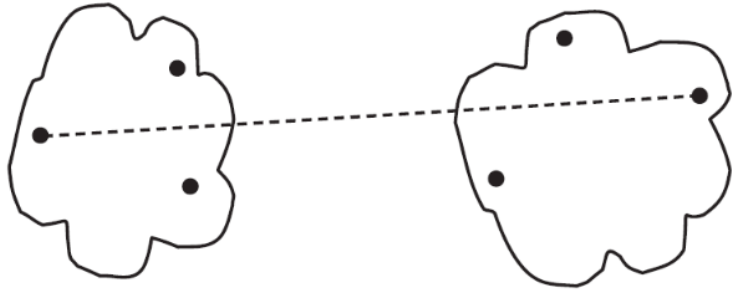
- Метод ближайшего соседа (простого связывания)
- Метод самого дальнего соседа (полного связывания)
- Метод средней связи
- Метод центроидов
- Метод Варда

Способы объединения кластеров

Метод «ближайшего соседа» - объединяются кластеры с наименьшим расстоянием между элементами

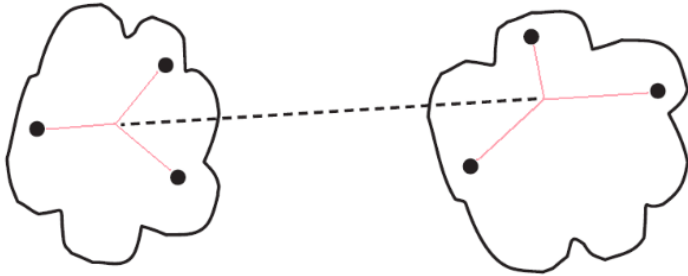


Метод «дальнего соседа» - объединяются кластеры с наибольшим расстоянием между элементами



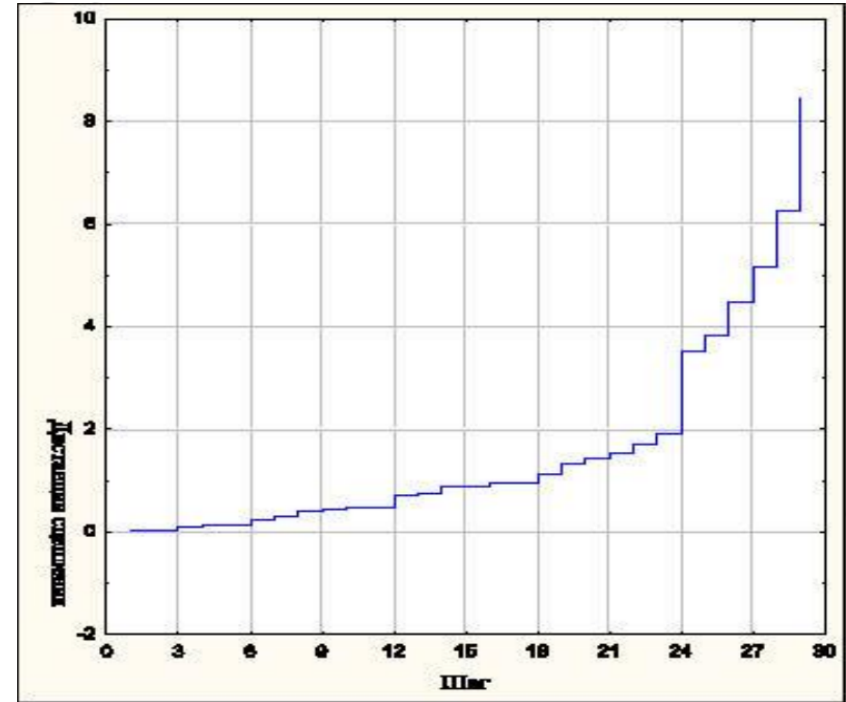
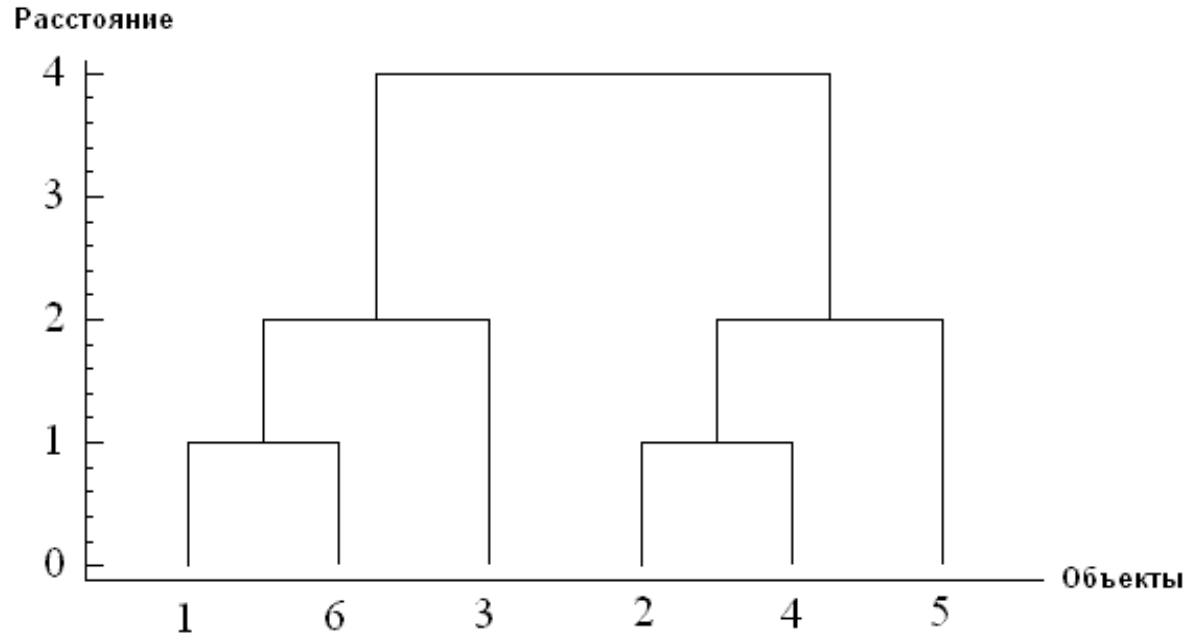
Групповое среднее расстояние

Расстояние между центрами



Метод Уорда (Варда, Ward) – объединяются кластеры, дающие наименьшую суммарную дисперсию

Метод иерархической агломерации



Метод k-средних

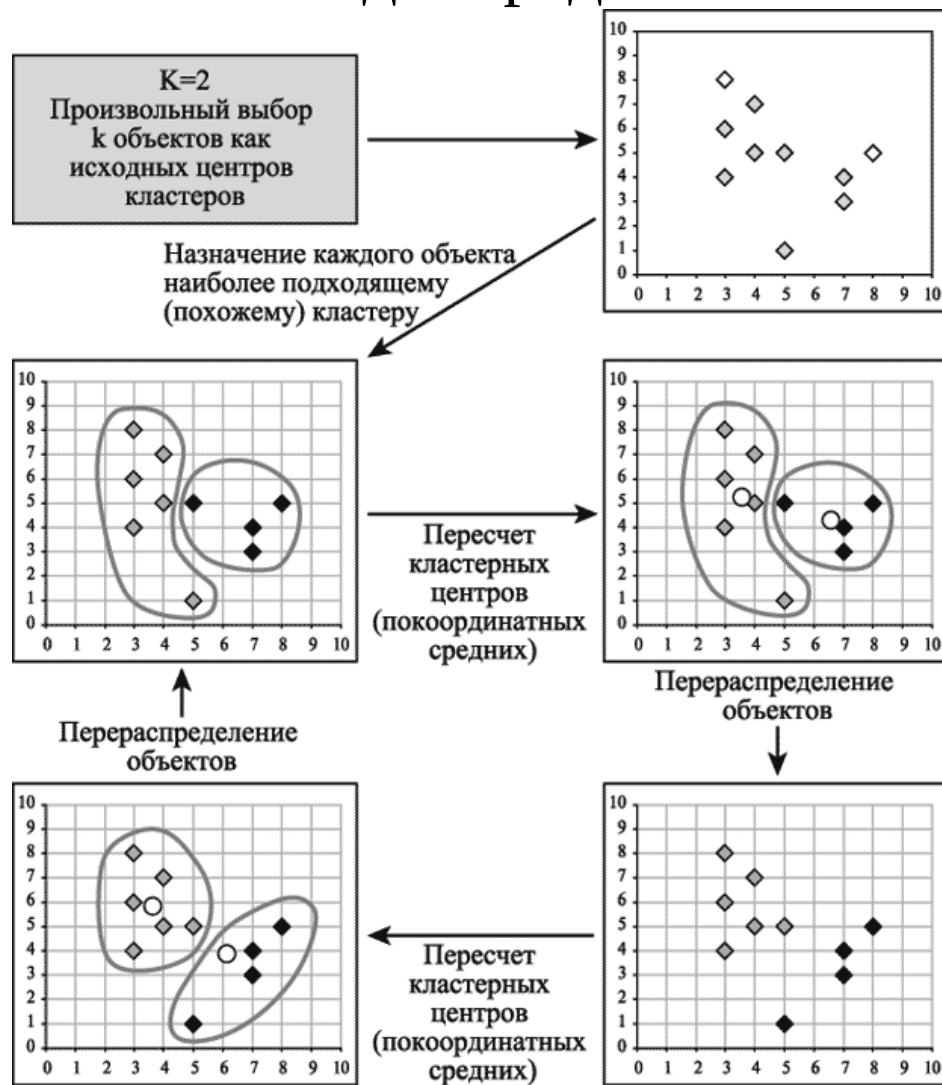
Метод k-средних — наиболее популярный метод кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров

Алгоритм представляет собой версию ЕМ-алгоритма, применяемого также для разделения смеси гауссиан. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k .

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение V уменьшается, поэтому заикливание невозможно.

Метод к-средних



Нейронные сети Кохонена

Сети, называемые картами Кохонена, - это одна из разновидностей нейронных сетей, использующих неконтролируемое обучение. Идея сети Кохонена принадлежит финскому ученому Тойво Кохонену (1982 год). Основным принцип работы сетей - введение в правило обучения нейрона информации относительно его расположения.

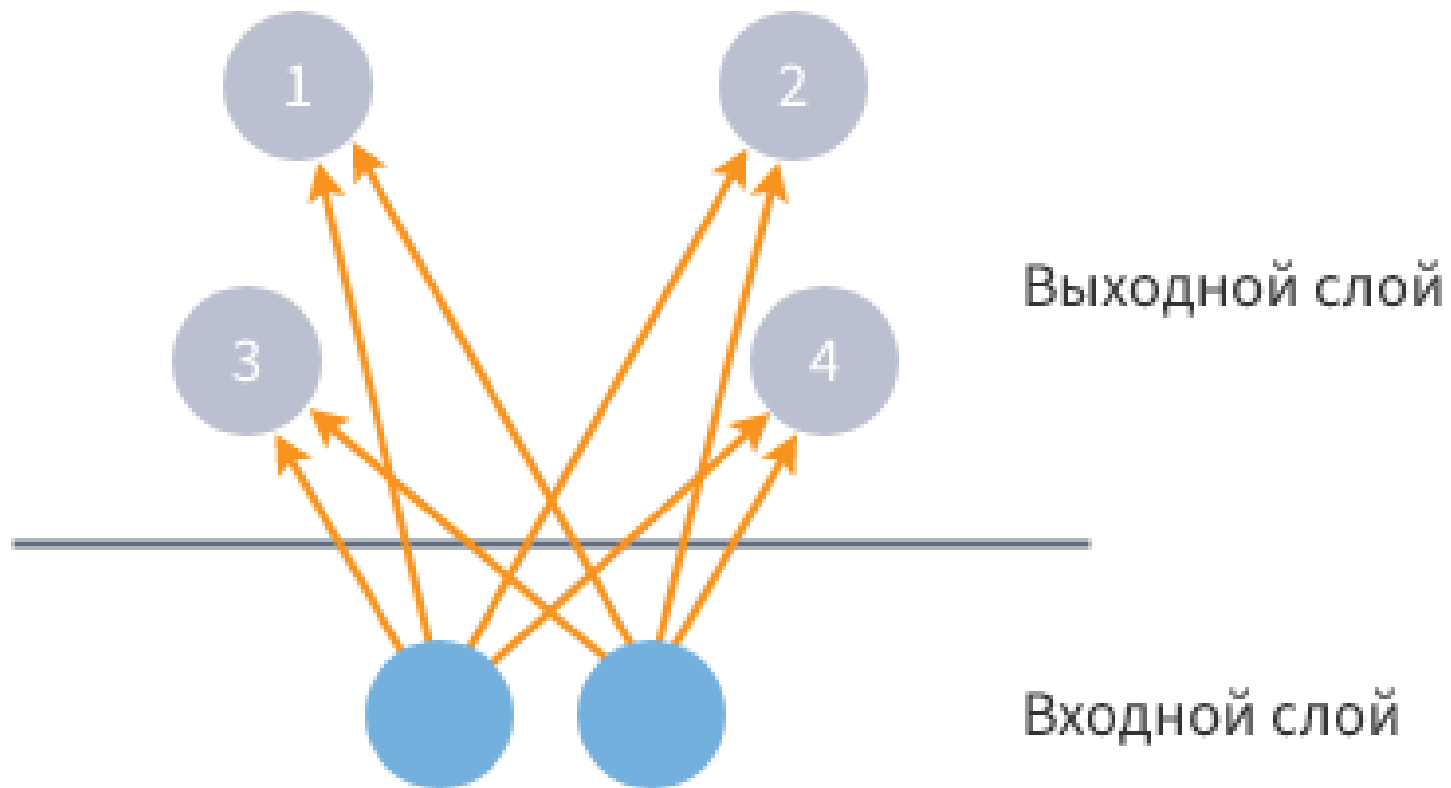
Самоорганизующиеся карты могут использоваться для решения таких задач, как моделирование, прогнозирование, поиск закономерностей в больших массивах данных, выявление наборов независимых признаков и сжатие информации. Наиболее распространенное применение сетей Кохонена - решение задачи классификации без учителя, т.е. кластеризации.

Нейронные сети Кохонена

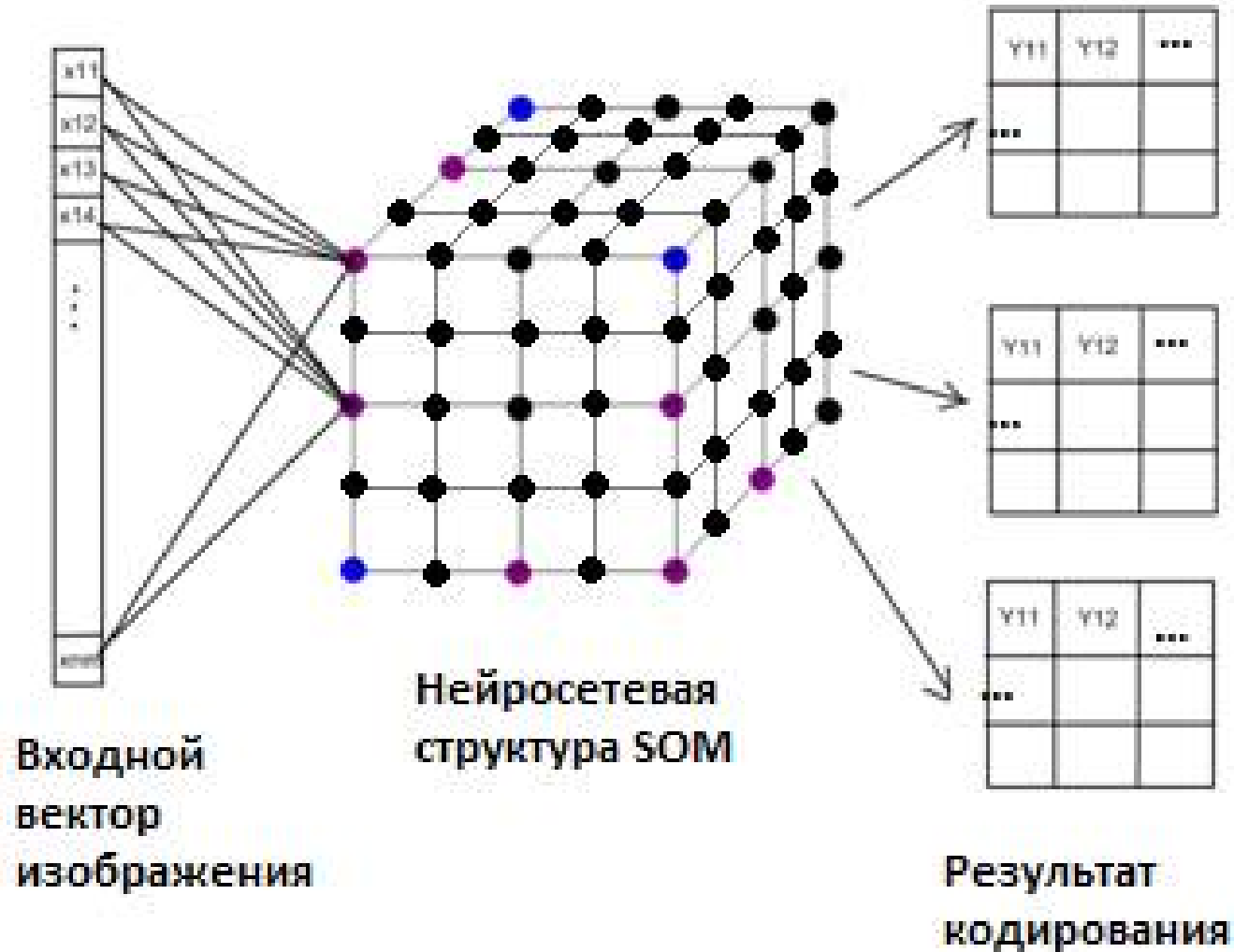
Сеть Кохонена обучается методом последовательных приближений. В процессе обучения таких сетей на входы подаются данные, но сеть при этом подстраивается не под эталонное значение выхода, а под закономерности во входных данных. Начинается обучение с выбранного случайным образом выходного расположения центров.

В процессе последовательной подачи на вход сети обучающих примеров определяется наиболее схожий нейрон. Этот нейрон объявляется победителем и является центром при подстройке весов у соседних нейронов. Такое правило обучения предполагает "соревновательное" обучение с учетом расстояния нейронов от "нейрона-победителя".

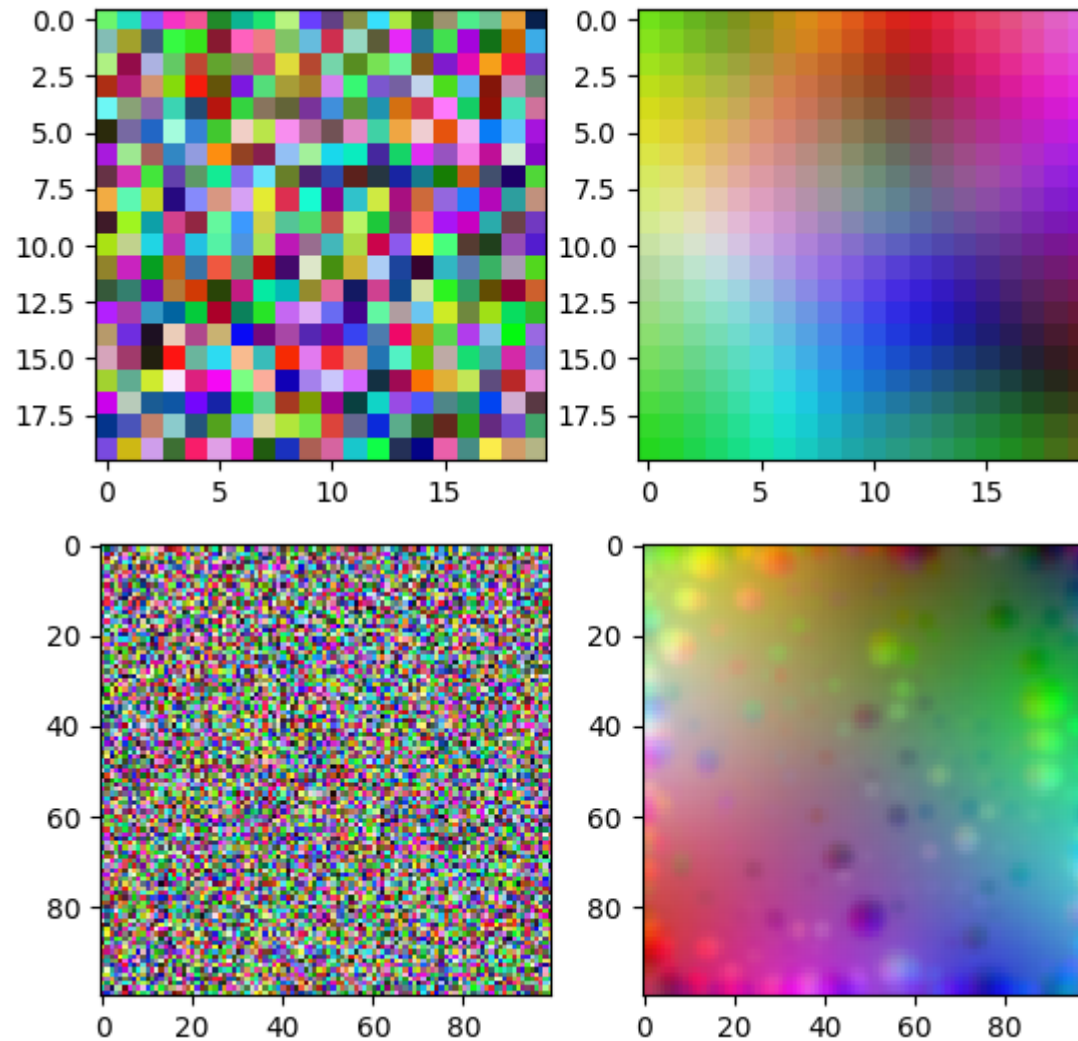
Компоненты сети Кохонена



Компоненты сети Кохонена



Результаты работы



DBSCAN – Density-Based Spatial Clustering

Основанная на плотности пространственная кластеризация для приложений с шумами - это алгоритм кластеризации данных, который предложили Маритин Эстер, Ганс-Петер Кригель, Ёрг Сандер и Сяовэй Су в 1996.

Это алгоритм кластеризации, основанной на плотности — если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены (точки со многими близкими соседями, помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко)).

Схема работы метода

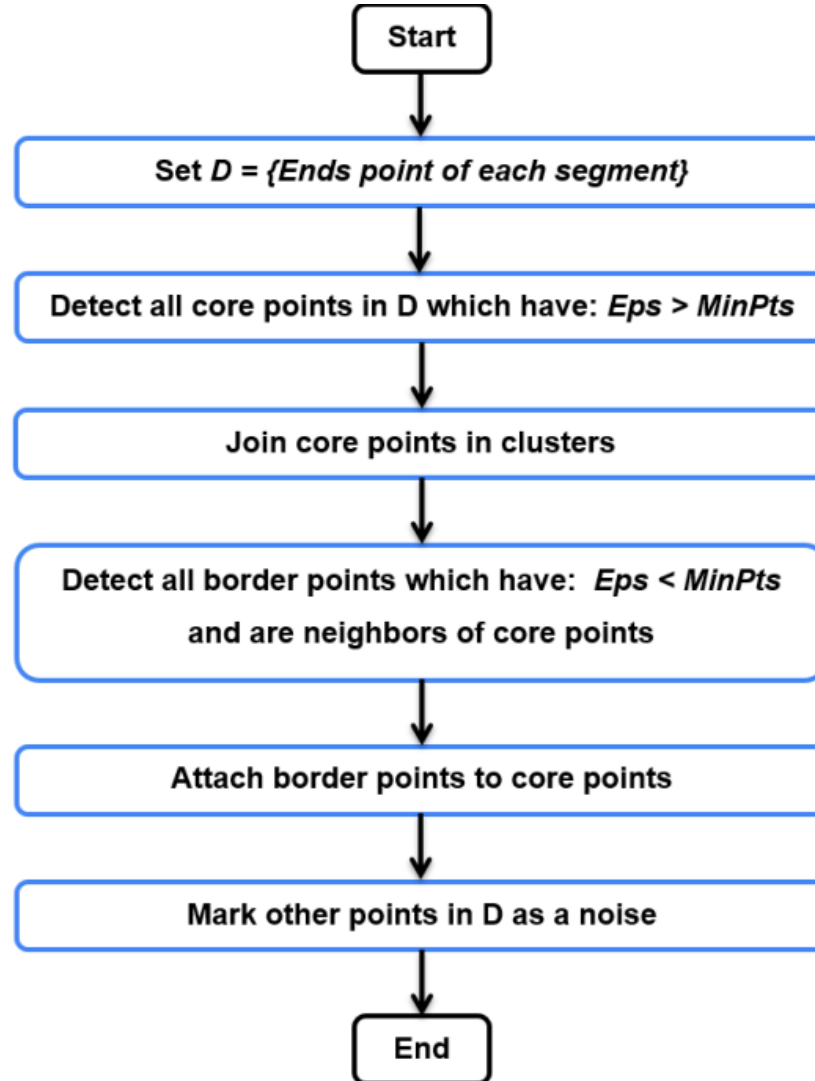
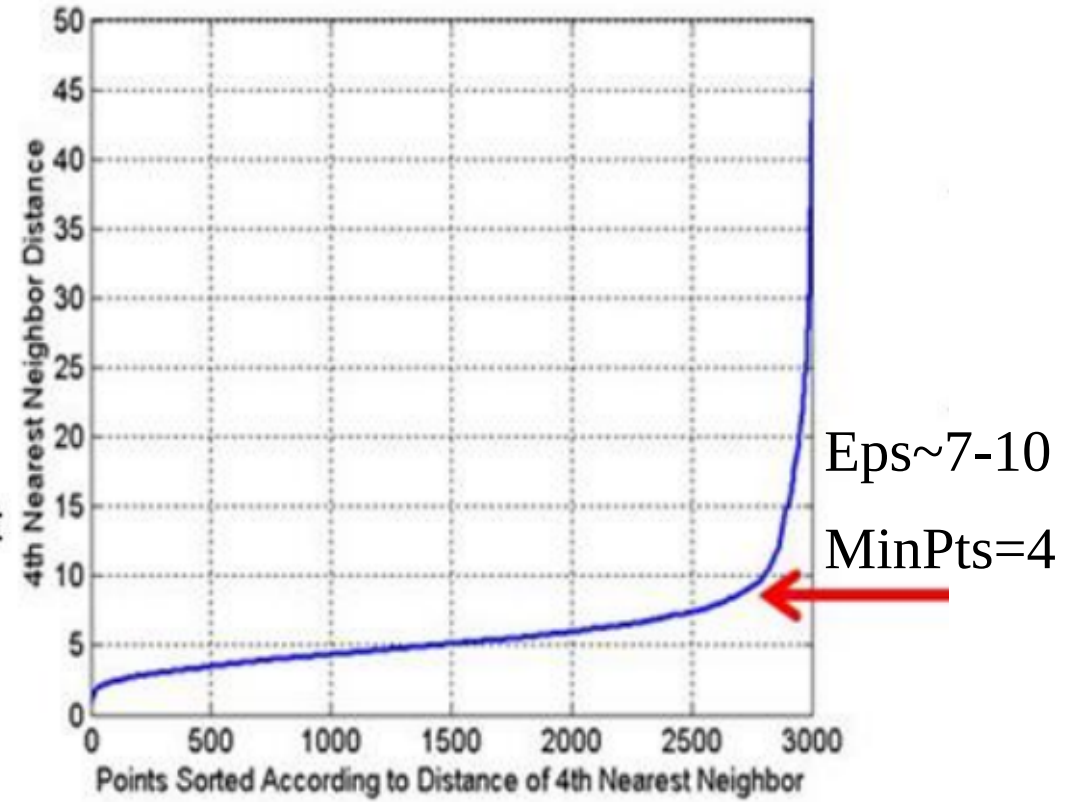
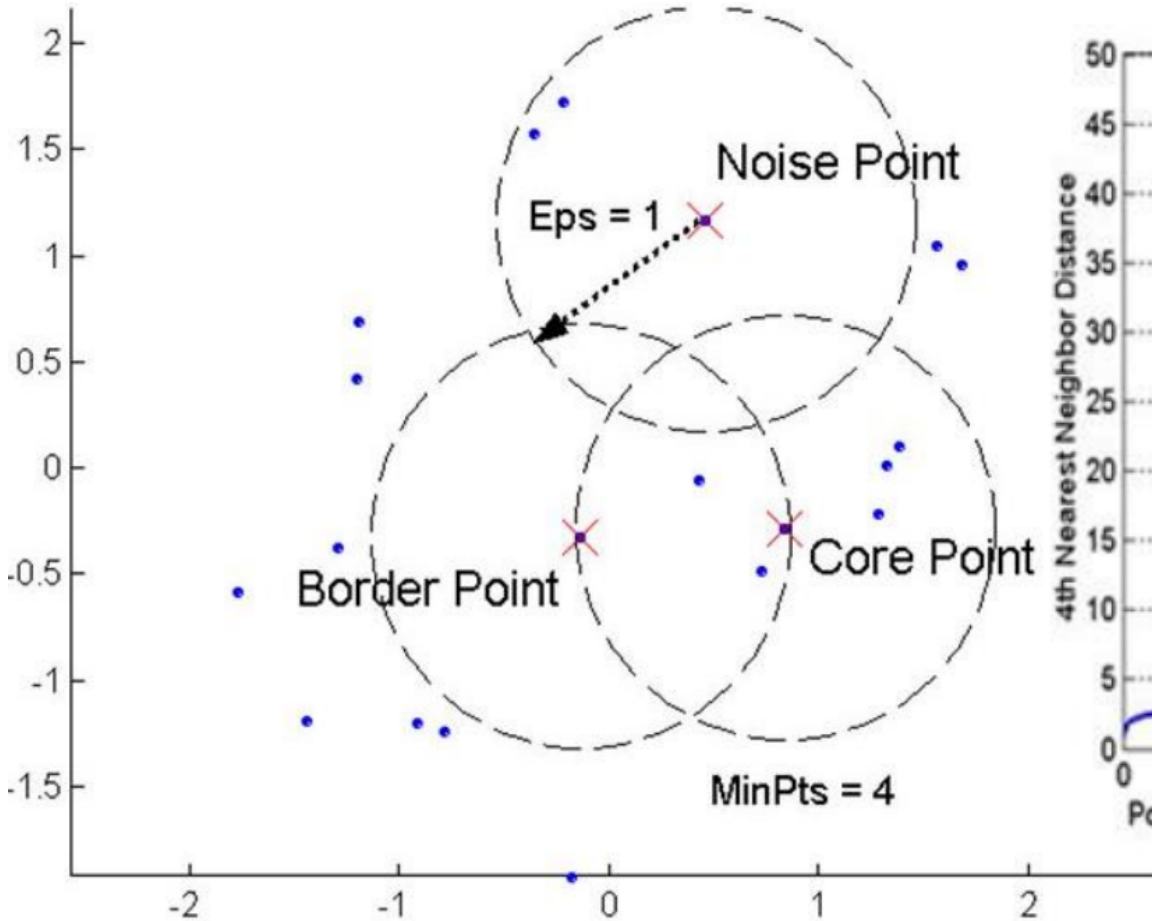


Схема работы метода



DBSCAN – преимущества

- DBSCAN не требует спецификации числа кластеров в данных априори в отличие от метода k-средних.
- DBSCAN может найти кластеры произвольной формы.
- DBSCAN имеет понятие шума и устойчив к выбросам.
- DBSCAN требует лишь двух параметров и большей частью нечувствителен к порядку точек в базе данных.
- DBSCAN имеет скорость работы $O(n \log n)$

DBSCAN – недостатки

- DBSCAN не полностью однозначен — краевые точки, которые могут быть достигнуты из более чем одного кластера, могут принадлежать любому из этих кластеров, что зависит от порядка просмотра точек.
- Качество DBSCAN сильно зависит от измерения расстояния
- DBSCAN не может хорошо кластеризовать наборы данных с большой разницей в плотности, поскольку не удастся выбрать приемлемую для всех кластеров комбинацию ϵ

Мягкая кластеризация

Мягкая кластеризация (англ. Fuzzy/soft clustering) — тип кластеризации, при котором каждая точка может принадлежать одному или нескольким кластерам. Мягкая кластеризация также называется нечёткой кластеризацией и используется при решении задач обработки естественного языка, в том числе в лексической семантике.

Критерии качества кластеризации

$$F_0 = \frac{\sum_{i < j} [y_i = y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i = y_j]} \rightarrow \min$$

Среднее внутрикластерное расстояние

$$F_1 = \frac{\sum_{i < j} [y_i \neq y_j] \rho(x_i, x_j)}{\sum_{i < j} [y_i \neq y_j]} \rightarrow \max$$

Среднее межкластерное расстояние

$$F_0/F_1 \rightarrow \min$$

Отношение пары функционалов

Критерии качества кластеризации

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{i: y_i = y} \rho^2(x_i, \mu_y) \rightarrow \min$$

Сумма средних внутрикластерных расстояний

$K_y = \{x_i \in X^\ell \mid y_i = y\}$ - кластер y μ_y - центр масс кластера y

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \rightarrow \max$$

Сумма межкластерных расстояний

μ - центр масс всей выборки.

Спасибо за внимание!