# 1 Методы кластерного анализа

# 2 Иерархические алгоритмы

# 3 Расстояния между кластерами

*Возможные расстояния*:

1. **Расстояние "ближайшего соседа"** (одиночная связь):

$$\rho_{\min}(K_i, K_j) = \min_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно расстоянию между самыми ближними объектами кластеров.

2. **Расстояние "дальнего соседа"** (полная связь):

$$\rho_{\max}(K_i, K_j) = \max_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно расстоянию между самыми дальними объектами кластеров.

3. **Невзвешенное попарное среднее**:

$$\rho_{\text{mean}}(K_i, K_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно среднему между всеми парами расстояний.

4. **Взвешенное попарное среднее**:

$$\rho_{\text{mean2}}(K_i, K_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \rho(k_1 \cdot x_i, k_2 \cdot x_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \frac{k_1 x_i + k_2 x_j}{k_1 + k_2}.$$

   Равно среднему между всеми парами расстояний с учетом весов $k_1, k_2$, равных ёмкости кластеров.

5. **Незвешенный центроидный метод**: расстояние между кластерами равно расстоянию между их центрами тяжести, где центр тяжести есть среднее арифметическое всех объектов в кластере:

$$x_c = \frac{\sum_{i=1}^{k} 1 \cdot x_i}{k} = \text{mean}_{x_i \in K_i}(x_i)$$

6. **Взвешенный центроидный метод**. Как я понял, это расстояние между центрами, но с учетом весов, как в пункте 4. Но вообще начиная с пункта 4 нигде нет формул, так что не факт, что они у меня правильные.

7. **Метод Варда**. Целевой функцией является внутригрупповая сумма квадратов:

$$\text{SS} = \sum_i \sum_{x_j \in K_i} \rho(x_j, x_i),$$

   где сумма берётся по всем кластерам, $x_i$ – среднее по кластеру $K_i$. Объединение в кластеры на каждой итерации происходит так, чтобы увеличение этой функции было минимальным.

# 4 Процедуры эталонного типа

Наряду с иерархическими методами кластеризации существуют итеративные ($k$-средних), суть которых заключается в следующем (возможны модификации):

1. Среди объектов $x_i$ некоторым образом выбираются $k$ штук – центры будущих кластеров.
2. Объекты, не отнесённые к какому-либо кластеру, приписываются тому кластеру, до которого будет наименьшее расстрояние.
3. Центры кластеров пересчитываются.
4. Для всех точек пересматривается их принадлежность к кластеру.
5. Пункты 3-4 повторяются, пока точки могут менять принадлежность.

На этом сайте хорошо показано, как работает метод $k$-средних: `https://www.naftaliharris.com/blog/visualizing-k-means-clustering/`

# 5 Методика дискриминантного анализа

# 6 Что характеризует Лямбда Уилкса?

# 7 Что показывают квадраты расстояний Махаланобиса?

# 8 Какое максимальное число канонических дискриминантных функций допустимо в дискриминантном анализе?

# 9 Какую информацию дают стандартизованные и структурные коэффициенты дискриминантной функции?

# 10 Опишите процедуру отбора переменных с помощью стандартизованных и структурных коэффициентов

# 11 Какова интерпретация канонического коэффициента корреляции?

Общее число дискриминантных функций не превышает числа дискриминантных переменных и, по крайней мере, на единицу меньше числа групп. Степень разделения выборочных групп зависит от величины собственных чисел: чем больше собственное число, тем сильнее разделение. Наибольшей разделительной способностью обладает первая дискриминантная функция, соответствующая наибольшему собственному числу $\lambda_i$, вторая обеспечивает максимальное различение после первой и т. д. Различительную способность $i$-й функции оценивают по относительной величине в процентах собственного числа $\lambda_i$ от суммы всех $\lambda$.

**Коэффициент канонической корреляции**. Другой характеристикой, позволяющей оценить полезность дискриминантной функции является коэффициент канонической корреляции $r_i$. Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1. Будем считать, что группы составляют одно множество, а

другое множество образуют дискриминантные переменные. Коэффициент канонической корреляции для $i$-й дискриминантной функции определяется формулой:

$$r_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}.$$

**Чем больше величина $r_i$, тем лучше разделительная способность дискриминантной функции.**

# 12 В каком случае учет априорных вероятностей может сильно изменить результаты классификации?

Априорные вероятности оказывают наибольшее влияние при перекрытии групп и, следовательно, многие объекты с большой вероятностью могут принадлежать ко многим группам. Если группы сильно различаются, то учет априорных вероятностей практически не влияет на результат классификации, поскольку между классами будет находиться очень мало объектов

# 13 Методика факторного анализа

# 14 Суть задачи вращения общих факторов

Задача вращения общих факторов решается с целью улучшения их *интерпретируемости*. Производится попытка достижения простой структуры, в которой каждая переменная характеризуется преобладающим влиянием какого–то одного фактора. Факторные нагрузки могут быть изображены в виде диаграммы рассеяния, на которой каждая переменная представлена точкой. Можно повернуть оси в любом направлении без изменения относительного положения точек. При этом действительные координаты точек, то есть факторные нагрузки, изменяются.

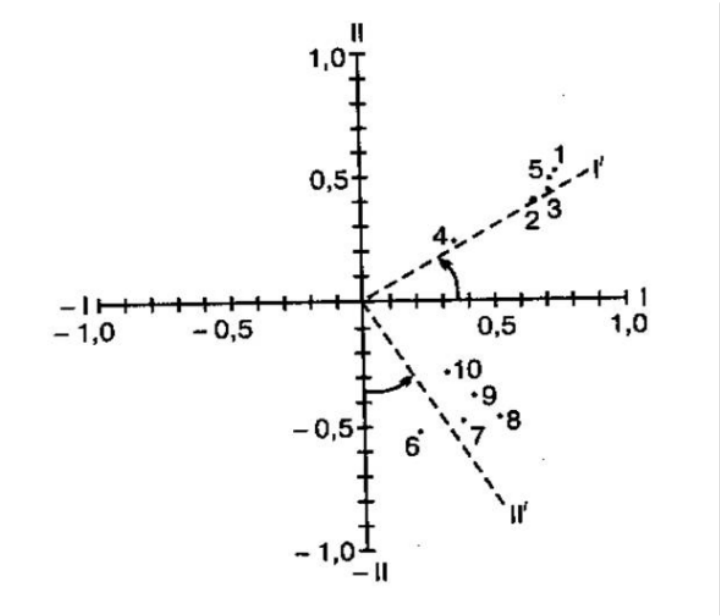


Рис. 1: Зависимость сигнала от шума для данных.

Существуют различные методы вращения факторов. Целью этих методов является получение понятной (интерпретируемой) матрицы нагрузок, то есть факторов, которые ясно отмечены высокими нагрузками для некоторых переменных и низкими – для других.

*Примеры методов вращения*:

- **Варимакс**. Ортогональный метод вращения, минимизирующий число переменных с высокими нагрузками на каждый фактор.
- **Метод косоугольного (неортогонального) вращения**. Самое косоугольное решение соответствует дельте, равной 0 (по умолчанию). По мере того, как дельта отклоняется в отрицательную сторону, факторы становятся более ортогональными.
- **Квартимакс**. Метод вращения, который минимизирует число факторов, необходимых для объяснения каждой переменной.
- **Эквимакс**. Метод вращения, объединяющий методы варимакс, упрощающий факторы, и квартимакс, упрощающий переменные. Минимизируется число переменных с большими факторными нагрузками и число факторов, требуемых для объяснения переменной.
- **Промакс-вращение**. Косоугольное вращение в предположении, что факторы могут коррелировать между собой. Оно производится быстрее, чем вращение типа косоугольного вращения, поэтому оно полезно для больших наборов данных.

**Примечание**. КОСОУГОЛЬНОЕ ВРАЩЕНИЕ — такая трансформация факторного пространства, которая предполагает возможность проведения факторных осей под углами друг к другу, отличающимися от 90 градусов (от ортогональности)

# 15 Критерий Кайзера

**Критерий Кайзера или критерий собственных чисел**: отбираются только факторы с собственными значениями равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером (Kaiser, 1960), и является, вероятно, наиболее широко используемым.

# 16 Критерий Каменистой осыпи

# 17 Метод главных компонент