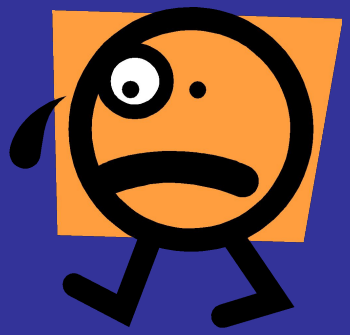
# «АНАЛИЗ И ИНТЕРПРЕТАЦИЯ ДАННЫХ»

**1.5. Матрица объект – объект и признак – признак, расстояние и близость**

**1.6. Измерение признаков**

**1.7. Основные типы шкал**



## 1.5. Матрица объект – объект и признак – признак, расстояние и близость

- Пусть имеется матрица данных *X(N* x n). Если рассматривать строки данной матрицы как *N* векторов xi, в пространстве *n* признаков, то естественно рассмотреть расстояние между двумя некоторыми векторами. Расстояния между всевозможными парами векторов дают матрицу R(N x *N)* расстояний типа объект - объект.

- Часто рассматривается величина, обратная в некотором смысле расстоянию - близость. На практике часто используют функции близости вида

$$\mu(x_1, x_2) = \exp[-\alpha d^2(x_1, x_2)] \quad \text{или} \quad \mu(x_1, x_2) = \frac{1}{1+\alpha \cdot d(x_1, x_2)},$$

где альфа- определяет крутизну функции близости. Очевидно, что матрица близостей также является симметричной с единичной главной диагональю, так как m(x1, x1)= 1.

- Если рассмотреть признаки как *n* векторов в N-мерном пространстве объектов, то получим другое преобразование матрицы данных в матрицу R(n x *n)* типа признак - признак. Элементом rij такой матрицы является значение расстояния или близости между признаками *Xi* и *Xj.* Наиболее распространено представление в виде матрицы близостей между признаками, где под близостью понимается, например, корреляция соответствующих признаков.
- Легко заметить, что содержательные задачи на матрице данных *X(N* x n) интерпретируются на квадратных матрицах R(N x *N)* и R(n x *n)* как выделение блочно - диагональной структуры путем одновременной перегруппировки строк и столбцов. Тогда в каждом диагональном блоке группируются элементы, близкие в соответствующем пространстве и далекие от элементов других блоков. Такая задача группировки известна как задача диагонализации матрицы связей .Задача о диагонализации матрицы связей является наиболее общей для матриц связей произвольной природы.

## 1.6. Измерение признаков

Данные получают в результате измерения некоторых свойств объектов. Для того, чтобы провести измерение, должны присутствовать собственно объекты с интересующими нас физическими свойствами и измерительное устройство.

- объекты обладают обычно самыми разными свойствами. В результате измерения фиксируются только некоторые свойства объекта и не учитываются многие другие. Следовательно, в матрице данных содержится заведомо неполная информация об объектах исследования.

*Например* , объекты могут оказаться эквивалентными по весу или длине, если значения таких характеристик присутствуют в матрице данных как значения соответствующих признаков. Те же объекты могут оказаться совершенно различными по цвету или форме. Но это различие никак не отразится на результатах обработки, если эти свойства не были представлены в матрице данных в виде значений соответствующих признаков.

- Под измерительным устройством может пониматься не только некоторый прибор, но и человек, например, респондент, отвечающий на вопросы некоторой анкеты.

  Важно, чтобы измерительное устройство было способно изменить свое состояние в ответ на изменение состояния объекта. Очевидно, что измеряющая способность устройства зависит от того, насколько структурированы свойства объектов.

- Простейшая структурированность свойств объектов позволяет судить о совпадении или различии состояний. Для представления такой довольно грубой структуры не обязательно использовать числа, так как словами можно легко обозначить факт простого совпадения состояний или их различия. Таким образом, язык можно использовать для выражения классификационных понятий, совокупность которых образует шкалу наименований или номинальную шкалу.

- Признаки, значения которых измеряются в шкалах наименований или порядка, называются качественными.
- Признаки, значения которых измеряются в числовых, то есть количественных шкалах, называются количественными.

# 1.7. Основные типы шкал

- Тип шкалы определяется типом преобразований, с помощью которых одна числовая система переводится в другую числовую систему .

- К числу преобразований, характеризующих основные типы шкал, относятся: тождественное, подобия, сдвига, линейное, монотонное и взаимнооднозначное.

  Чем меньше множество числовых систем, в которые гомоморфно (разные по форме) отображается данная эмпирическая система, тем мощнее шкала, в которой она измеряется, по набору допустимых операций над ее числовыми значениями.

- Наименее мощным типом шкалы является номинальная шкала.

  Измерение признака в номинальной шкале состоит в разбиении объектов на классы эквивалентности, где объектам одного класса соответствует одно число. В номинальной шкале значения числовой системы Uz определены с точностью до взаимно - однозначного преобразования j(x), где x- исходное числовое значение. Это означает, что *k* различным значениям xijÎ{1,…*k*} компоненты i признака *Xj* можно поставить в соответствие *k* произвольных различных значений j( xij )Î{j(1), j(2),… j(k)}.

- Более мощной является порядковая шкала. Числовые системы, в которые гомоморфно отображается эмпирическая система с отношением линейного порядка, должны сохранять порядок на множестве объектов, соответствующий их ранжированию. *В* порядковой шкале значения числовой системы определены с точностью до монотонных преобразований.

- Следующая шкала уже относится к количественному типу - шкала интервалов. В такой шкале значения числовой системы измеряются с точностью до линейного преобразования вида j(x)=ax+b, a> 0. В шкале интервалов сохраняется отношение разности численных значений. Действительно, пусть объектам a1, a2, a3, a4 в некоторой числовой системе соответствуют значения f(a1)= x11, f(a2)= x21, f(a3)= x31, f(a4)= x41 ,то есть измерен признак X1=(x11, x21, x31, x41)T . Пусть в другой числовой системе измерен признак F(X1) == (j(x11),j(x21),j(x31),j(x41))T . Тогда получим : $$\frac{\varphi(x_{11})-\varphi(x_{21})}{\varphi(x_{31})-\varphi(x_{41})} = \frac{\alpha\cdot x_{11}-\alpha\cdot x_{21}+\beta-\beta}{\alpha\cdot x_{31}-\alpha\cdot x_{41}+\beta-\beta} = \frac{x_{11}-x_{21}}{x_{31}-x_{41}}$$
- *Примером* измерения в шкале интервалов является значение температуры по шкалам Цельсия, Кельвина, Фаренгейта.

- Следующая шкала - шкала отношений. В такой шкале значения числовой системы измеряются с точностью до преобразования подобия вида j(x) = ax, a > 0. В такой шкале сохраняются отношения численных значений. Действительно, пусть объектам a1 и a2 соответствуют значения f(a1) =x11 и f(a2) =x21 в одной числовой системе и значения j( f(a1)) и j( f(a2)) в другой числовой системе, то есть значениям признака *X1*= (x11,x21) T соответствуют значения признака Ф(*X1*) = (j(x11), j(x21))T . Тогда получим $$\frac{\varphi(x_{11})}{\varphi(x_{21})} = \frac{\alpha\cdot x_{11}}{\alpha\cdot x_{21}} = \frac{x_{11}}{x_{21}}$$

Измерениями в шкале отношений являются измерения веса, длины и прочих именованных величин, характеризующихся масштабом.

- Наиболее мощной является абсолютная шкала.

  В ней значения числовой системы определяются с точностью до тождественного преобразования j(x)= x. Результаты измерения в абсолютной шкале определяются однозначно, например, число стульев, количество рабочих. Любое преобразование, кроме тождественного, исказит эти измерения и приведет к неправильным данным.