Введение.
Основные структуры данных.

Рассмотрим систему координат, изображенную на рис. 1.

Возможны три ситуации (рис.2,3,4):

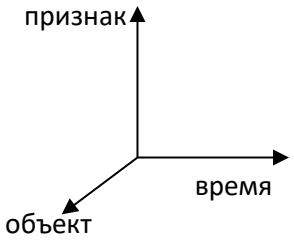


Рис.1.

Рис.2.

Рис.3.

Рис.4.

Соответственно различают три структуры данных:

1. матрица типа “объект-признак”;
2. временной ряд;
3. матрица близостей.

Рассмотрим подробнее эти ситуации:

Ситуация 1.

Пусть матрица X - матрица типа “ объект-признак ”.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ & \dots & & \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{bmatrix}$$

p - число признаков,
N - число объектов,
x_{ij} – значение j-го признака на i-ом объекте, i=[1,..,N], j=[1,..,p].

Основные задачи:

1) Сжатие информации.

Содержательная постановка задачи: найти небольшое число наиболее важных свойств исследуемого явления (например: рост, размер, полнота при изготовлении одежды).

Формальная постановка задачи: устранить дублирующие друг друга признаки или построить новые признаки (меньшее число) описывающие данные. Построение новых признаков должно производиться без потери информации. Они объективно существуют, но не измеряются непосредственно. Этими задачами занимается **факторный анализ**.

2) Задача исследования зависимости одного признака от других (этот признак называется целевым).

Содержательная постановка: описать зависимость избранного свойства исследуемых объектов от остальных свойств (например, прогноз погоды по набору народных примет).

Формальная постановка: найти функциональную зависимость, приближенно описывающую изменение целевого признака при изменении других признаков. Такими задачами занимается **регрессионный анализ**.

3) Задачи классификации.

а) **Кластерный анализ**.

Содержательная постановка: среди множества исследуемых объектов найти естественные группы с похожими свойствами. Например, создание классификации животных и растений по классам, родам и видам.

Формальная постановка: обнаружить в пространстве описания компактные скопления точек.

б) **Классификация с обучением**.

Содержательная постановка: найти правило, пользуясь которым можно определить принадлежность любого объекта к одному из заданных образов или классов.

Формальная постановка задачи: найти в пространстве описания поверхность, разделяющую группы точек, соответствующих различным образам, и описать ее как функцию исходных признаков или найти к какой группе точек (образу) относятся данные точки-объекты.

Ситуация 2.

Определение: Совокупность наблюдений некоторого признака, осуществляемых последовательно во времени через равные его промежутки называется временным рядом.

Пример 1: средний балл студента от сессии к сессии.

Пример 2: y – состояние популяции грызунов,

$y_t = 1$, если вспышка численности,

$y_t = 0$,если вспышка отсутствует.

Это бинарный временной ряд.

Пример 3: x – электрическая активность головного мозга пациента

ряд x_t–непрерывный временной ряд.

Основные задачи.

1. Извлечение из временного ряда информации о механизме, его генерирующем.
2. Прогнозирование получения новых значений временного ряда.
3. Управление системой.

Раздел прикладной статистики, занимающийся указанными выше задачами, называется **анализом временных рядов**.

Временные ряды делятся на одномерные и многомерные. Примеры 1-3 – одномерные.

Пример многомерного временного ряда:

Имеется временной ряд x. Пусть в данный момент времени:

x_1 – производительность труда

x_2 - рентабельность

x_3 – прибыль

x_4 – травматизм

x_5 - энерговооруженность

Ситуация 3.

Пример структуры данных: численность промысловой популяции горностая, соболя.

Основная задача: найти объекты со схожей динамикой.

Основой метод, с помощью которого изучается данный вопрос, называется многомерным шкалированием. С помощью этого метода исследуется матрица близости (размерность матрицы $n \times n$).

x_{ij} – экспериментальная оценка близости i-го объекта к j-му.

Шкалы измерений и типы признаков.

Определение. Признаки называются разнотипными, если они измеряются в различных шкалах.

Типы признаков:

1. бинарный
2. номинальный (классификационный). (Принадлежит ли данный предмет данному классу).
3. ранговый (порядковый). (Ранг – это место, занимаемое данным объектом в совокупности объектов).
4. количественный.

Пример: пусть x – численность популяции. Рассмотрим признак как бинарный (есть(1) или нет (0)). Тот же признак может быть рассмотрен как номинальный (0 – вспышка, 1 – депрессия, 2 - норма). Если признак рассмотрен как ранговый, то:

№ района	популяция
1	(больше
2	меньше
3	меньше
...	...
n	меньше)

Основной вопрос для:

- ранговой шкалы - какой объект из данных предпочтительнее,
- номинальной шкалы - к какому классу относится объект,
- количественной шкалы - во сколько раз один объект предпочтительнее другого,
- бинарной шкалы - есть или нет.

Анализ матриц “объект-признак”.

Пусть Z - матрица данных типа “объект- признак” размером N×p (N – число объектов, p – число признаков)

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{Np} \\ \cdots & & \\ z_{N1} & \cdots & z_{Np} \end{bmatrix}$$

Особенности матрицы:

1. признаки измеряются в различных единицах.
2. у признаков различная масштабность.

Стандартизация матрицы данных.

$\bar{z}^j$ – *среднее значение* j – *го признака* $\quad \bar{z}^j = \frac{1}{N}\sum_{i=1}^{N} z_{ij}$

$S^{j^2} = \frac{1}{N}\sum_{i=1}^{N}\left(z_{ij} - \bar{z}^j\right)^2$ – *оценка дисперсии.*

Пусть $$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ & \dots & & \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{bmatrix}, \quad x_{ij} = \frac{z_{ij} - \bar{z}^j}{s^j}, \quad s^j = \sqrt{(s^2)^j}$$

X называется стандартизованной матрицей данных

Свойства X:

1. в матрице X все признаки безразмерны.
2. $\bar{x}^j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left(z_{ij} - \bar{z}^j\right)}{s^j} = \frac{1}{s^j}\left(\frac{1}{N}\sum_{i=1}^{N} z_{ij} - \frac{1}{N}\sum_{i=1}^{N} \bar{z}^j\right) = 0$

   Среднее значение каждого признака в стандартизованной матрице данных равно нулю.
3. $(s_x^j)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}^j\right)^2 = \frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = \frac{1}{N}\sum_{i=1}^{N} \frac{\left(z_{ij} - \bar{z}^j\right)^2}{s^{j\,2}} = 1,$

   $j = 1, \dots, p$

   Оценка дисперсии каждого признака в стандартизованной матрице равна единице.

$\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$.

Отсюда следует, что длина каждого вектора-признака стандартизованной матрицы данных равна $\sqrt{N}$.

Геометрическая интерпретация матрицы данных.

1. Каждому признаку (вектору - столбцу) поставим в соответствие ось координат. Тогда матрица X – это набор из N-точек в p- мерном пространстве. Это пространство называется пространством признаков, т.е. объект - это точка в пространстве признаков.
2. Каждому объекту поставим в соответствие ось некоторой системы координат в N - мерном пространстве. Тогда каждому признаку будет соответствовать в этом пространстве некоторая точка (это тоже вектор). Матрица данных - это p точек в N- мерном пространстве объектов. В силу того, что матрица стандартизована и длина всех признаков равна $\sqrt{N}$, часто задача взаимосвязи между признаками сводится к определению углов между ними в N- мерном пространстве объектов.

Ковариация и её свойства.

Ковариация - это количественная мера связи двух случайных величин.

Пусть ξ_1 и ξ_2 - случайные величины. Тогда

$\mathrm{cov}(\xi_1,\xi_2)=M[(\xi_1-M\xi_1)(\xi_2-M\xi_2)]=$

$=M(\xi_1\xi_2-\xi_2M\xi_1-\xi_1M\xi_2+M\xi_1M\xi_2)=$

$=M(\xi_1\xi_2)-M\xi_2M\xi_1-M\xi_1M\xi_2+M\xi_1M\xi_2=M(\xi_1\xi_2)-M\xi_1M\xi_2$

Свойства ковариации:

1. $\mathrm{cov}(\xi_1,\xi_2)=M(\xi_1\xi_2) - M\xi_1M\xi_2$.
2. Если случайные величины независимы, то их ковариация равна нулю.
3. $\mathrm{cov}(\xi_1,\xi_2)=\mathrm{cov}(\xi_2,\xi_1)$
4. Если a и b - некоторые константы, то $\mathrm{cov}(a\xi_1+b\xi_2,\xi_3)=a\mathrm{cov}(\xi_1,\xi_3)+b\mathrm{cov}(\xi_2,\xi_3)$. (Доказать самостоятельно на практике)
5. $\mathrm{cov}(\xi,\xi)=D\xi$.

Дисперсия линейной комбинации случайных величин .

Теорема: Пусть для случайных величин $\xi_1, \xi_2, \ldots \xi_n$ существуют ковариации: $\mathrm{cov}(\xi_i,\xi_j)=\sigma_{ij}$, i=1,..,n, j=1,..,n. Тогда для любых постоянных $c_1, c_2, \ldots, c_n$ существует дисперсия линейной комбинации случайных величин с этими весами $D\left(\sum_{i=1}^{n} c_i\xi_i\right)$, и эта дисперсия равна $\sum_{j=1}^{n}\sum_{i=1}^{n}\left(c_ic_j\sigma_{ij}\right)$.

Доказательство:

По определению, $DX=M(X-MX)^2$, где X – случайная величина.

$$D\left(\sum_{i=1}^{n} c_i\xi_i\right)=M\left(\sum_{i=1}^{n} c_i\xi_i-M\left(\sum_{i=1}^{n} c_i\xi_i\right)\right)^2=M\left(\sum_{i=1}^{n} c_i\xi_i-\sum_{i=1}^{n}M\left(c_i\xi_i\right)\right)^2=M\left(\sum_{i=1}^{n} c_i\left(\xi_i-M\left(\xi_i\right)\right)\right)^2=$$

$$=M\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left[c_ic_j\left(\xi_i-M\xi_i\right)\left(\xi_j-M\xi_j\right)\right]\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}c_ic_jM\left(\left(\xi_i-M\xi_i\right)\left(\xi_j-M\xi_j\right)\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}c_ic_j\sigma_{ij}$$

(Использована формула $\left(\sum_{i=1}^{n} a_i\right)^2=\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_j$).

Следствие 1:

Если $c_1, c_2, \ldots, c_n$ - константы, а случайные величины $\xi_1, \xi_2, \ldots, \xi_n$ независимы, то дисперсия их линейной комбинации $D\left(\sum_{i=1}^{n} c_i \xi_i\right) = \sum_{i=1}^{n} c_i^2 D\xi_i$.

Доказательство:

По пятому свойству $\sigma_{ii} = D\xi_i$, а если i не равно j , то $\sigma_{ij} = 0$, (по второму свойству). Т.о. остаются те слагаемые, где индексы совпадают: $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \sigma_{ij} = \sum_{i=1}^{n} c_i^2 D\xi_i$.

Следствие 2:

$$D(c_1\xi_1 + c_2\xi_2) = c_1^2\sigma_{11} + 2c_1c_2\sigma_{12} + c_2^2\sigma_{22} \qquad (*)$$

Оценка сверху модуля ковариации.

В (*) сделаем замену $c_1=x$, $c_2=1$. Тогда будем иметь $D(x\xi_1+\xi_2)=x^2\sigma_{11}+\sigma_{22}+2x\sigma_{12}\geq 0$

С другой стороны, относительно x, это парабола, у которой ветви направлены вверх (так как $\sigma_{11}>0$). Значение квадратного трехчлена для любого x больше или равно 0, поэтому его дискриминант меньше или равен нуля.

$\sigma_{12}^2 - \sigma_{11}\sigma_{22} \leq 0,$ откуда следует

$$|\sigma_{12}| \leq \sqrt{\sigma_{11}\sigma_{22}} \qquad (**)$$

Ковариационная матрица.

Заданы n случайных величин $\xi_1, \xi_2, \ldots, \xi_n$ и $\mathrm{cov}(\xi_i,\xi_j)=\sigma_{ij}$. Из них сформируем матрицу.

$$\sum = \begin{bmatrix} \sigma_{11} & \ldots & \sigma_{1n} \\ \ldots & \ldots & \ldots \\ \sigma_{n1} & \ldots & \sigma_{nn} \end{bmatrix}$$

Свойства ковариационной матрицы:

1. матрица $\sum$ - симметричная.
2. на главной диагонали матрицы стоят дисперсии.

Выборочная ковариация.

Пусть существуют случайные величины X и Y. Между X и Y существует стохастическая связь. Как определить эту связь?

Возьмем набор наблюдений (x_1,y_1), (x_2,y_2), …, (x_n,y_n).

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Пусть s - оценка ковариации, тогда $s = \frac{1}{n}\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]$. Это аналог определения ковариации.

Недостатки ковариации, как меры связи:
1. размерность ковариации может быть очень большой.
2. ковариация - величина не нормированная.

Коэффициент корреляции и его свойства.

Определение: Коэффициентом корреляции (КК) двух случайных величин ξ_1, ξ_2 называется $\rho(\xi_1,\xi_2) = \frac{\mathrm{cov}(\xi_1,\xi_2)}{\sqrt{D\xi_1 D\xi_2}}$.

Свойства КК:
1. ρ - величина безразмерная.
2. $|\rho| \le 1$, в силу (**).
3. $|\rho| = 1$ тогда и только тогда, когда ξ_1 и ξ_2 связаны линейной функциональной зависимостью ($\xi_2 = a\xi_1 + b$).
4. если ξ_1, ξ_2 независимы, то $\rho = 0$ (но не наоборот).
5. Если ξ_1, ξ_2 подчиняются двумерному нормальному закону распределения, то понятие некоррелированности и независимости идентичны.
6. $|\rho|$ не изменится, если ξ_1, ξ_2 подвергнуть линейному преобразованию.
   (если $\eta_1 = a_1\xi_1 + b_1$, $\eta_2 = a_2\xi_2 + b_2$,то $|\rho(\eta_1,\eta_2)| = |\rho(\xi_1,\xi_2)|$). (Доказать самостоятельно на практике)

Выборочный коэффициент корреляции.

Пусть X и Y - случайные величины. Берем наблюдения (x_1,y_1), (x_2,y_2), …, (x_n,y_n) и оцениваем коэффициент корреляции ρ(X,Y), r(X,Y) - оценка ρ(X,Y). Тогда

$$r(X,Y) = \frac{\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}.$$

Следствие определения выборочного коэффициента корреляции:

Пусть в роли X и Y выступают произвольные столбцы стандартизованной матрицы данных (их средние равны нулю, а оценки дисперсии - единице).

$$X^1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \dots \\ x_{N1} \end{bmatrix}, \dots, X^p = \begin{bmatrix} x_{1p} \\ x_{2p} \\ \dots \\ x_{Np} \end{bmatrix}.$$

Тогда выборочный коэффициент корреляции между i- ым и j- ым столбцами принимает вид

$$r\left(X^i, X^j\right) = r_{ij} = \frac{1}{N}\sum_{K=1}^{N} x_{ki}x_{kj}$$

Замечание.

$$r\left(X^i, X^j\right) = r_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ki}x_{kj} = \frac{1}{N}\left(X^i, X^j\right) = \frac{1}{N}\left|X^i\right\|X^j\left|\cos\alpha_{ij} = \frac{1}{N}\sqrt{N}\sqrt{N}\cos\alpha_{ij} = \cos\alpha_{ij}\right.$$

α_{ij} - это угол между i -м и j-м признаками в пространстве объектов.

Таким образом, коэффициент корреляции в данном случае - это косинус угла между векторами признаков в пространстве объектов.

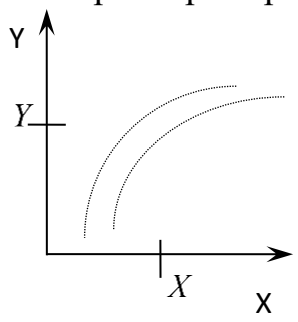
Корреляционный анализ.

Под корреляционным анализом совокупности признаков будем понимать вычисление различных мер связи между этими признаками и проверку статистических гипотез относительно этих мер.

Коэффициент корреляции как мера связи двух признаков.

Пусть имеется n наблюдений (x_1,y_1), (x_2,y_2), ..., (x_n,y_n).

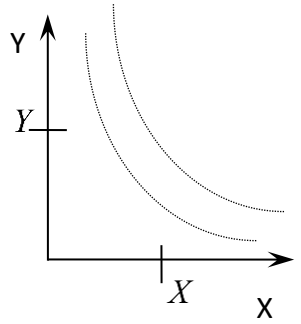
Рассмотрим пример.



1. X - производительность труда, Y - средняя заработная плата. Объекты - предприятия отрасли. Вычислим средние $\overline{X}, \overline{Y}$

Если $\left(X_i - \overline{X}\right) > 0$ и мы знаем, что зависимость возрастающая , то можно ожидать, что чаще всего $\left(Y_i - \overline{Y}\right) > 0$, т. е. если с ростом X, Y - в среднем, возрастает, то знаки сомножителей у слагаемых в числителе для формулы r(X,Y) намного чаще

будут одинаковыми, чем различными. Как следствие, r(X,Y) принимает достаточно большие положительные значения.



2. Пусть X- производительность труда, а Y- потери рабочего времени. Если очередное наблюдение таково, что $(X_i - \overline{X}) > 0$, то знак разности чаще всего «-». Если с ростом X, Y уменьшается, то знаки сомножителей слагаемых числителя в формуле для r(X,Y) чаще всего будут разные, следовательно r(X,Y) примет большое по модулю и отрицательное значение.

Принято считать, что:

- если |r|<0.3, то X и Y практически не коррелированны
- если 0.3≤|r|<0.6, то корреляция слабая
- если 0.6<|r|<0.8, то говорят, что имеется корреляция
- если 0.8<|r|, имеется сильная корреляция.

Оценка значимости коэффициента корреляции.

Пусть статистические гипотезы H_0 и H_1 состоят в следующем:

H_0: $\rho(x,y) = 0$, т.е. связи между признаками нет.

H_1: $\rho(x,y) \neq 0$,т.е. связь есть.

Действие состояние природы	*H_0 отвергаем*	*H_0 принимаем*
H_0 истинна	ошибка α	верное решение
H_1 истинна	верное решение	Ошибка β

Пусть α – вероятность ошибки первого рода, т.е. вероятность отвергнуть истинную гипотезу. Пусть β вероятность ошибки второго рода, т.е. вероятность принять неверную гипотезу. Нужно сформулировать такое правило, чтобы α и β были малыми.

Пусть $t = r\dfrac{\sqrt{n-2}}{\sqrt{1-r^2}}$.

В математической статистике показано, что статистика t при условии, что H_0 справедлива, подчиняется закону распределения Стьюдента с (n-2) степенями свободы.

Алгоритм проверки статистической гипотезы о значимости КК, (о справедливости гипотезы H_1).

КК значим, если между X и Y имеется связь.

1. даны (x_1,y_1), (x_2,y_2), …, (x_n,y_n) - экспериментальные данные. Вычислим r(X,Y).
2. пусть α =0.05 - приемлемая для нас вероятность ошибки.
3. вычисляем значение статистики $t_{расч}$.
4. по выбранному α и числу степеней свободы f=n-2 определим $t_{табл}$

$t_{табл} = F(\alpha,f)$.

Правило для вынесения решения будет таким:

если $|t_{расч}| \geq |t_{табл}|$, то решаем, что справедлива гипотеза H_1, в противном случае, если $|t_{расч}| < |t_{табл}|$, считаем, что справедлива гипотеза H_0, т.е. связи между X и Y нет, а отличие от нуля выборочного КК обусловлено случайными причинами.

Корреляционная матрица.

Пусть X(N×p) - стандартизованная матрица данных.

Тогда КК равен $r_{ij} = \frac{1}{N}\sum_{k=1}^{N} x_{ki}x_{kj}, \quad i = \overline{1,p}, \quad j = \overline{1,p}$

Матрица, состоящая из таких коэффициентов корреляции, называется корреляционной матрицей и обозначается R(p×p).

$$R(p \times p) = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ & \dots & & \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{bmatrix}$$

Матрица R получается умножением матриц X и X^T.

$R(p\times p) = \frac{1}{N} X^T(p\times N)\ X(N\times p)$

Свойства корреляционной матрицы:

1. R- симметричная матрица.
2. Элементы диагонали $r_{ii} = \cos(\alpha_{ii}) = 1$.