



Machine Learning

Лекция #1

Основы математической
статистики



Числовые атрибуты
Однофакторный анализ

Однофакторный анализ сфокусирован на одном атрибуте и его значениях в заданный момент времени.

Матрица данных D ($1 \times n$)

$$D = \begin{pmatrix} X \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

- X – числовой атрибут,
- $x_i \in R$ – случайная величина, причем x_i ($1 \leq i \leq n$)

При этом:

- Наблюдаемые данные - случайная выборка из $X \Rightarrow$ каждая переменная x_i независима и распределена аналогично X .
- В векторном представлении образец данных - n -мерный вектор $X \in R^n$.
- Плотность распределения вероятности $f(x)$ и кумулятивная функция распределения $F(x)$ для атрибута X - неизвестны.

Эмпирическая кумулятивная функция распределения

Эмпирическая кумулятивная
функция распределения (CDF)
атрибута X:

где

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x) \quad (2.1)$$

$$I(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{if } x_i > x \end{cases}$$

двоичная переменная – индикатор (показывает выполнены условия или нет).

Обратная эмпирическая кумулятивная функция распределения или квантильная функция случайной переменной X возвращает минимальное значение X , для которых значение CDF равно q :

$$F^{-1}(q) = \min\{x \mid F(x) \geq q\} \quad \text{for } q \in [0, 1] \quad (2.2)$$

Эмпирическая функция распределения вероятности

(PMF) величины X :

$$\hat{f}(x) = P(X = x) = \frac{1}{n} \sum_{i=1}^n I(x_i = x) \quad (2.3)$$

where

$$I(x_i = x) = \begin{cases} 1 & \text{if } x_i = x \\ 0 & \text{if } x_i \neq x \end{cases}$$

(вероятность значения в каждой точке x_i равна $1/n$)



Числовые атрибуты
Меры среднего значения

Математическое ожидание -
это среднее арифметическое значение величины X .

Математическое ожидание для дискретной случайной величины X :

$$\mu = E[X] = \sum_x x f(x) \quad (2.4)$$

где $f(x)$ – функция
распределения вероятности X .

Математическое ожидание для непрерывной случайной величины X :

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

где $f(x)$ – функция
распределения плотности
вероятности X .

Выборочное среднее значение -

это функция $\hat{\mu} : \{x_1, x_2, \dots, x_n\} \rightarrow \mathbb{R}$, определяющая среднее значение x_i :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.5)$$

Несмещенное среднее выборочное значение

Оценка $\hat{\theta}$ - несмещенная оценка для θ , если $E[\hat{\theta}] = \theta$ для всех значений θ . Выборочное среднее значение $\hat{\mu}$ является несмещенной оценкой для среднего значения μ , если:

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n E[x_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu \quad (2.6)$$

Где x_i распределены аналогично $X \Rightarrow E[x_i] = \mu$ для всех x_i . E – линейная функция $\Rightarrow$ для любых двух случайных величин X и Y , и натуральных чисел a и b , справедливо равенство:
 $E[aX + bY] = aE[X] + bE[Y]$.

Оценка устойчивости

статистика устойчива, если она инвариантна к выбросам значений данных.

Медиана

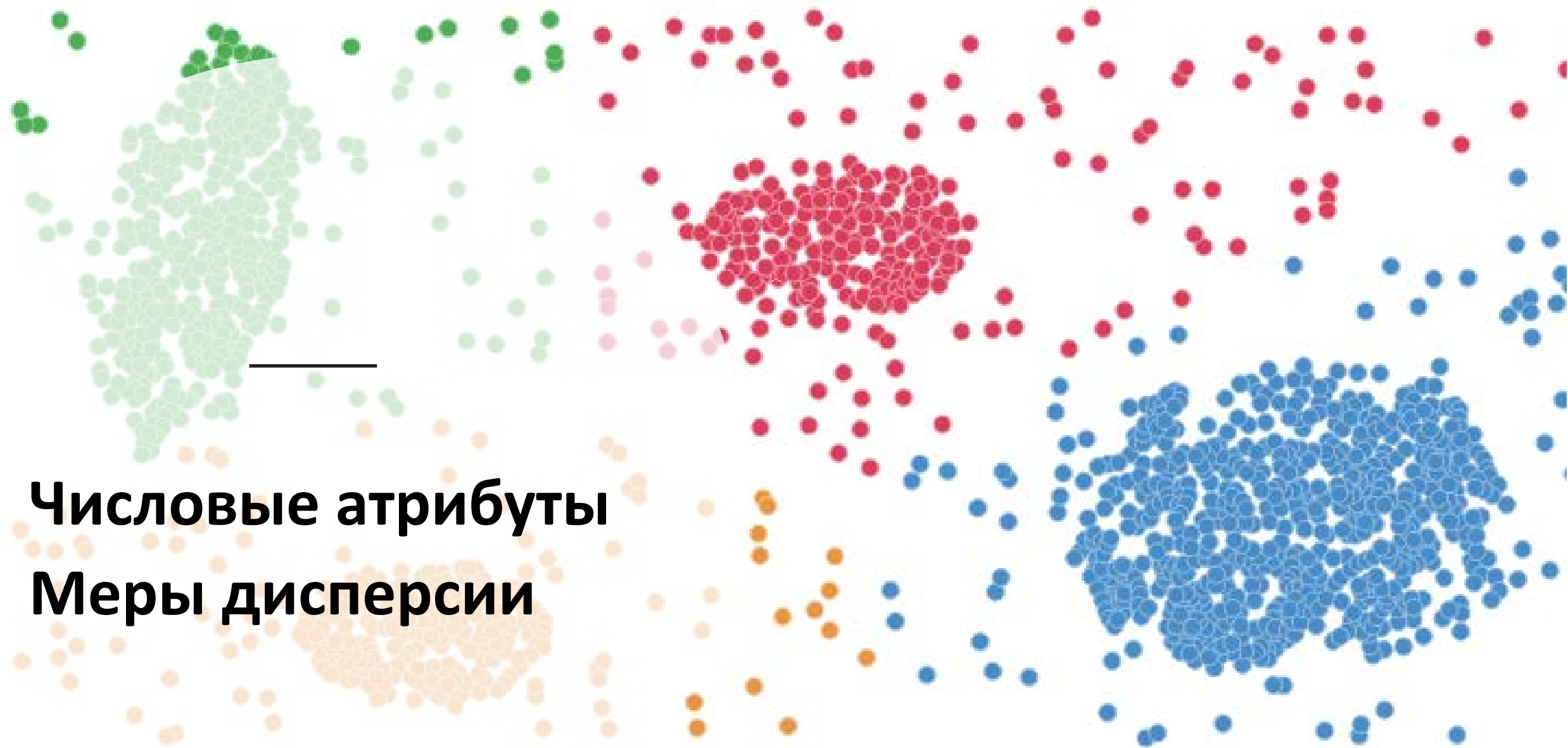
Медиана случайной величины - m , если справедливо:

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

Мода

Мода X это значение, для которого PMF достигает максимума. Для выборочной моды - это эмпирическая функция распределения [Eq. (2.3)]

$$\text{mode}(X) = \arg \max_x \hat{f}(x)$$



Числовые атрибуты
Меры дисперсии

Диапазон случайной величины X - разница между максимальным и минимальным значениями

$$r = \max\{X\} - \min\{X\}$$

Выборочный диапазон – статистика вида:

$$\hat{r} = \max_{i=1}^n \{x_i\} - \min_{i=1}^n \{x_i\}$$

Межквартильный диапазон

Квартили – это определенные значения квартиль-функции, которая делит данные на 4 равные части (0.25, 0.5, 0.75, и 1.0).

Межквартильный диапазон (IQR) :

$$IQR = q_3 - q_1 = F^{-1}(0.75) - F^{-1}(0.25)$$

Дисперсия X - мера отклонения от среднего значения (математического ожидания).

Или дисперсия - это ожидаемая величина квадрата отклонения от математического ожидания.

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases} \quad (2.8)$$

Среднеквадратическое отклонение σ определяется как квадратный корень от дисперсии σ^2 .

Выборочная дисперсия

Выборочная дисперсия:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

Выборочное среднеквадратическое отклонение:

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2}$$

Стандартизированная оценка (z-оценка) выборки x_i :

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

Дисперсия выборочного среднего значения

Так как все x_i независимы:

$$\text{var} \left(\sum_{i=1}^n x_i \right) = \sum_{i=1}^n \text{var}(x_i)$$

Так как x_i распределены аналогично X , то и дисперсии равны:

$$\text{var}(x_i) = \sigma^2 \text{ for all } i$$

Комбинируя эти два факта:

$$\text{var} \left(\sum_{i=1}^n x_i \right) = \sum_{i=1}^n \text{var}(x_i) = \sum_{i=1}^n \sigma^2 = n\sigma^2$$

При этом:

$$E \left[\sum_{i=1}^n x_i \right] = n\mu$$

Дисперсия выборочного среднего $\hat{\mu}$ может быть вычислена:

$$\begin{aligned} \text{var}(\hat{\mu}) &= E[(\hat{\mu} - \mu)^2] = E[\hat{\mu}^2] - \mu^2 = E\left[\left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2\right] - \frac{1}{n^2} E\left[\sum_{i=1}^n x_i\right]^2 \\ &= \frac{1}{n^2} \left(E\left[\left(\sum_{i=1}^n x_i\right)^2\right] - E\left[\sum_{i=1}^n x_i\right]^2 \right) = \frac{1}{n^2} \text{var}\left(\sum_{i=1}^n x_i\right) \\ &= \frac{\sigma^2}{n} \end{aligned} \quad (2.14)$$

Выборочная дисперсия является смещенной, но асимптотически она не смещенная:

$$\sum_{i=1}^n (x_i - \mu)^2 = n(\hat{\mu} - \mu)^2 + \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (2.15)$$

Используя (2.15) на первом шаге мы имеем:

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2\right] = E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2\right] - E[(\hat{\mu} - \mu)^2] \quad (2.16)$$

Так как x_i распределены аналогично X , то они имеют те же среднее значение μ и дисперсию σ^2 .
Следовательно:

$$E[(x_i - \mu)^2] = \sigma^2$$

Далее, из (2.14) выборочное среднее $\hat{\mu}$ имеет дисперсию:

$$E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$$

Подставляя это в (2.16), получаем:

$$\begin{aligned} E[\hat{\sigma}^2] &= \frac{1}{n} n\sigma^2 - \frac{\sigma^2}{n} \\ &= \left(\frac{n-1}{n}\right) \sigma^2 \end{aligned}$$

Выборочная дисперсия $\hat{\sigma}^2$ - смещенная оценка σ^2 , так как ее ожидаемое значение отличается от самой дисперсии в $(n-1)/n$ раз. Однако, она асимптотически не смещенная, так как смещение исчезает при $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} \frac{n-1}{n} = \lim_{n \rightarrow \infty} 1 - \frac{1}{n} = 1$$



Числовые атрибуты
Двухфакторный анализ

В двухфакторном анализе мы рассматриваются два числовых атрибута - x_1 и x_2 , совместно с данными D ($n \times 2$ матрица):

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Геометрически D может быть представлена:

- Вид из n точек или векторов в 2-мерном пространстве над атрибутами X_1 и X_2 , таким образом:

$$\mathbf{x}_i = (x_{i1}, x_{i2})^T \in \mathbb{R}^2$$

- Вид двух точек или векторов в n -мерном пространстве состоящем из точек, (каждый столбец это вектор в $\mathbb{R}^n$):

$$X_1 = (x_{11}, x_{21}, \dots, x_{n1})^T$$

$$X_2 = (x_{12}, x_{22}, \dots, x_{n2})^T$$

Эмпирическая совместная функция распределения вероятности

$$\hat{f}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n I(\mathbf{x}_i = \mathbf{x}) \quad (2.17)$$

$$\hat{f}(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) = \frac{1}{n} \sum_{i=1}^n I(x_{i1} = x_1, x_{i2} = x_2)$$

где $\mathbf{x} = (x_1, x_2)^T$ и I – переменная-индикатор, принимающая значение 1, только когда ее аргумент - истина:

$$I(\mathbf{x}_i = \mathbf{x}) = \begin{cases} 1 & \text{if } x_{i1} = x_1 \text{ and } x_{i2} = x_2 \\ 0 & \text{otherwise} \end{cases}$$

Математическое ожидание

Двухфакторное математическое ожидание:

$$\mu = E[\mathbf{X}] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad (2.18)$$

Вектор выборочного математического ожидания может быть получен из $\widehat{f}_{X_1}$ и $\widehat{f}_{X_2}$ - эмпирических функций распределения вероятностей величин X_1 и X_2 .

$$\hat{\mu} = \sum_{\mathbf{x}} \mathbf{x} \hat{f}(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x} \left(\frac{1}{n} \sum_{i=1}^n I(\mathbf{x}_i = \mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (2.19)$$

Дисперсия

Пусть σ_1^2 - дисперсия для X_1 и σ_2^2 для X_2 . Общая дисперсия представляется в виде:

$$\text{var}(\mathbf{D}) = \sigma_1^2 + \sigma_2^2$$

Ковариация между X_1 и X_2

показывает меру ассоциации или линейной зависимости между ними:

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

Исходя из ожидания линейности, имеем:

$$\begin{aligned}\sigma_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[X_1X_2 - X_1\mu_2 - X_2\mu_1 + \mu_1\mu_2] \\ &= E[X_1X_2] - \mu_2E[X_1] - \mu_1E[X_2] + \mu_1\mu_2 \\ &= E[X_1X_2] - \mu_1\mu_2 \\ &= E[X_1X_2] - E[X_1]E[X_2]\end{aligned}$$

Если X_1 и X_2 - независимые случайные, то их ковариация равна нулю. Если X_1 и X_2 - независимы, то:

$$E[X_1 X_2] = E[X_1] \cdot E[X_2]$$

Что в свою очередь означает:

$$\sigma_{12} = 0$$

Выборочная ковариация X_1 и X_2 :

$$\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$$

Корреляция

Корреляция между X_1 и X_2 - стандартная ковариация, нормализованная среднеквадратическим отклонением каждой переменной:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

Выборочная корреляция между атрибутами X_1 и X_2 :

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^n (x_{i2} - \hat{\mu}_2)^2}}$$

Геометрическая интерпретация выборочной ковариации и корреляции

Пусть Z_1 и Z_2 обозначают центральные векторы атрибутов в $\mathbb{R}^n$, представленные в виде:

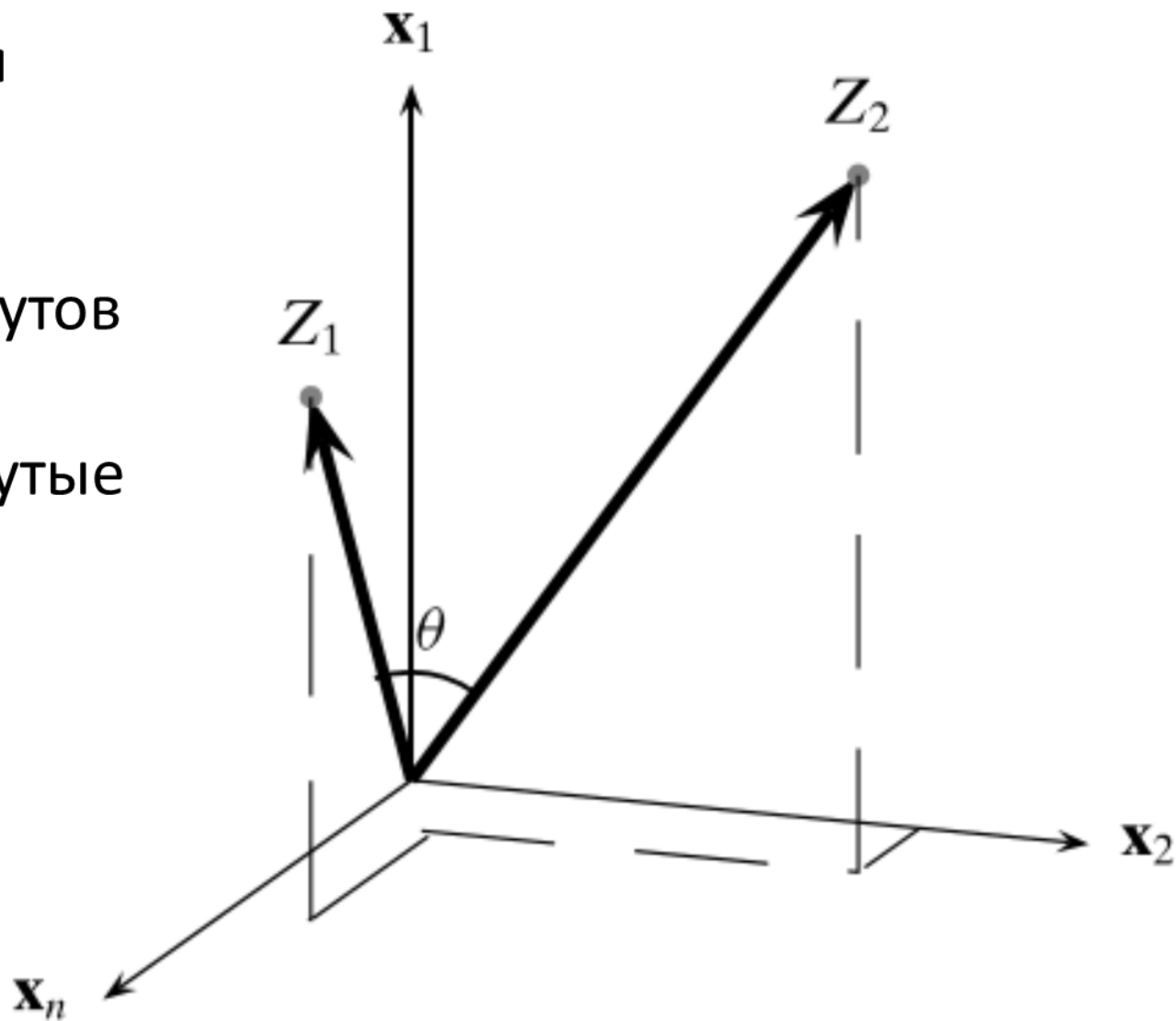
$$Z_1 = X_1 - \mathbf{1} \cdot \hat{\mu}_1 = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix} \quad Z_2 = X_2 - \mathbf{1} \cdot \hat{\mu}_2 = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}$$

Выборочная ковариация может быть записана так:

$$\hat{\sigma}_{12} = \frac{Z_1^T Z_2}{n}$$

Геометрическая интерпретация ковариации и корреляции.

Два центральных вектора атрибутов показаны (концептуально) в n -мерном пространстве R^n , натянутые на n точек.



Матрица ковариантности

Информация о дисперсии и ковариации для двух атрибутов X_1 и X_2 может быть обобщена в квадратной 2×2 ковариационной матрице:

Общая дисперсия двух атрибутов - сумма диагональных элементов Σ , которые также называются следом Σ :

Общая дисперсия двух атрибутов также обозначает ковариацию, в дополнении к дисперсии атрибутов и представлена в виде матрицы ковариации Σ , обозначаемой как $|\Sigma|$ or $\det(\Sigma)$.

Выборочная матрица ковариации представлена в виде:

$$\Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T]$$

$$= E \left[\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix} (X_1 - \mu_1 \quad X_2 - \mu_2) \right]$$

$$= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

$$var(\mathbf{D}) = tr(\Sigma) = \sigma_1^2 + \sigma_2^2$$

$$\hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}$$



Числовые атрибуты
Многофакторный анализ

В многофакторном анализе рассматриваются все d числовых атрибута $X_1, X_2, \dots, X_d$. Полные данные представлены в виде $n \times d$ матрицы:

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Данные строк - набор из n точек или векторов в d -мерном пространстве атрибутов:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^T \in \mathbb{R}^d$$

Данные столбцов - d точек или векторов в n -мерном пространстве:

$$X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T \in \mathbb{R}^n$$

С вероятностной точки зрения d атрибутов моделируются вектором случайных переменных

$\mathbf{X} = (X_1, X_2, \dots, X_d)^T$ и точки x_i считаются случайной выборкой, взятой из X , таким образом они независимы и распределены идентично X .

Математическое ожидание

Вектор многофакторного математического ожидания получается путем взятия среднего значения каждого атрибута:

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_d] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}$$

Выборочное математическое ожидание:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

Матрица ковариации

Обобщая до d-мерности, информация о ковариации представляется в виде $d \times d$ (квадратной) симметричной матрицы ковариации, которая задает ковариацию для каждой пары атрибутов:

$$\Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Диагональные элементы σ_i^2 определяют дисперсию для X_i , а элементы $\sigma_{ij} = \sigma_{ji}$ - ковариацию между парами X_i и X_j .

Матрица ковариации - положительно полуопределенная матрица

Σ является положительно полуопределенной матрицей, то есть $\mathbf{a}^T \Sigma \mathbf{a} \geq 0$ для любого d-мерного вектора $\mathbf{a}$.

При этом:

$$\begin{aligned}\mathbf{a}^T \Sigma \mathbf{a} &= \mathbf{a}^T E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] \mathbf{a} \\ &= E[\mathbf{a}^T (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \mathbf{a}] \\ &= E[Y^2] \\ &\geq 0\end{aligned}$$

где Y –случайная величина $Y = \mathbf{a}^T (\mathbf{X} - \boldsymbol{\mu}) = \sum_{i=1}^d a_i (X_i - \mu_i)$ и ожидаемый квадрат случайной переменной a – больше нуля. Так как Σ симметрична, то все собственные значения Σ реальные и не отрицательные. То есть $\mathbf{d}$ собственных значений Σ можно упорядочить от большего к меньшему:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0.$$

Как следствие, определитель Σ неотрицателен:

$$\det(\Sigma) = \prod_{i=1}^d \lambda_i \geq 0 \tag{2.27}$$

Полная и обобщенная дисперсия

Полная дисперсия представлена как след матрицы ковариации:

$$var(\mathbf{D}) = tr(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_d^2 \quad (2.28)$$

Обобщенная дисперсия - определитель ковариационной матрицы - $\det(\Sigma)$, (или $|\Sigma|$). Из уравнения. (2.27) $\Rightarrow \det(\Sigma) \geq 0$.

Выборочная матрица ковариации

Выборочная матрица ковариации определяется как:

$$\hat{\Sigma} = E[(\mathbf{X} - \hat{\mu})(\mathbf{X} - \hat{\mu})^T] = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \dots & \hat{\sigma}_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \dots & \hat{\sigma}_{2d} \\ \dots & \dots & \dots & \dots \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \dots & \hat{\sigma}_d^2 \end{pmatrix}$$

Ковариационная матрица может быть получена с помощью матричных операций. Пусть $\mathbf{Z}$ - центрированная матрицу данных, заданная как матрица центрированных векторов атрибутов $\mathbf{Z}_i = \mathbf{X}_i - \mathbf{1} \cdot \hat{\mu}_i$, where $\mathbf{1} \in \mathbb{R}^n$:

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \hat{\mu}^T = \begin{pmatrix} | & | & \cdots & | \\ \mathbf{Z}_1 & \mathbf{Z}_2 & \cdots & \mathbf{Z}_d \\ | & | & \cdots & | \end{pmatrix}$$

Или в терминах центрированных точек $\mathbf{z}_i = \mathbf{x}_i - \hat{\mu}$:

$$\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \hat{\mu}^T = \begin{pmatrix} \mathbf{x}_1^T - \hat{\mu}^T \\ \mathbf{x}_2^T - \hat{\mu}^T \\ \vdots \\ \mathbf{x}_n^T - \hat{\mu}^T \end{pmatrix} = \begin{pmatrix} - & \mathbf{z}_1^T & - \\ - & \mathbf{z}_2^T & - \\ & \vdots & \\ - & \mathbf{z}_n^T & - \end{pmatrix}$$

В матричной нотации выборочная ковариационная матрица может быть записана как:

$$\hat{\Sigma} = \frac{1}{n} (\mathbf{Z}^T \mathbf{Z}) = \frac{1}{n} \begin{pmatrix} \mathbf{Z}_1^T \mathbf{Z}_1 & \mathbf{Z}_1^T \mathbf{Z}_2 & \cdots & \mathbf{Z}_1^T \mathbf{Z}_d \\ \mathbf{Z}_2^T \mathbf{Z}_1 & \mathbf{Z}_2^T \mathbf{Z}_2 & \cdots & \mathbf{Z}_2^T \mathbf{Z}_d \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}_d^T \mathbf{Z}_1 & \mathbf{Z}_d^T \mathbf{Z}_2 & \cdots & \mathbf{Z}_d^T \mathbf{Z}_d \end{pmatrix} \quad (2.30)$$

Выборочная матрица ковариации задается как попарные внутренние или скалярные произведения центрированных векторов атрибутов, нормализованные по размеру выборки. В терминах центрированных точек $\mathbf{z}_i$ она также может быть записана как сумма матриц первого ранга, полученных как произведение каждой центрированной точки:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \cdot \mathbf{z}_i^T \quad (2.31)$$

Нормализация диапазона

Пусть X - атрибут и $x_1, x_2, \dots, x_n$ - случайная выборка из X . При нормализации диапазона каждое значение масштабируется на диапазон выборки $\hat{r}$ of X :

$$x'_i = \frac{x_i - \min_i \{x_i\}}{\hat{r}} = \frac{x_i - \min_i \{x_i\}}{\max_i \{x_i\} - \min_i \{x_i\}}$$

После преобразования новый атрибут принимает значения в диапазоне $[0,1]$.

Стандартная оценочная нормализация

В стандартной оценочной нормализации, называемой также z-нормализацией, каждое значение заменяется его z-оценкой:

$$x'_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

где $\hat{\mu}$ - выборочное мат. ожидание и $\hat{\sigma}^2$ - выборочная дисперсия X . После трансформации мат. ожидание новой величины станет $\hat{\mu}' = 0$, среднеквадратическое отклонение - $\hat{\sigma}' = 1$.



Числовые атрибуты

Нормальное распределение

Однофакторное нормальное распределение

Случайная величина X имеет нормальное распределение с мат. ожиданием μ и дисперсией σ^2 , если функция распределения плотности вероятности X представлена в виде:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

Распределение вероятности

Для данного интервала $[a, b]$ распределение вероятности при нормальном распределении внутри интервала представляется в виде:

$$P(a \leq x \leq b) = \int_a^b f(x|\mu, \sigma^2) dx$$

Распределение вероятности внутри k стандартных отклонений от мат. ожидания $[\mu - k\sigma, \mu + k\sigma]$

$$P(\mu - k\sigma \leq x \leq \mu + k\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu - k\sigma}^{\mu + k\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} dx$$

Через изменение переменной $z = (x-\mu)/\sigma$, мы получаем эквивалентную формулу в терминах стандартного нормального распределения:

$$\begin{aligned} P(-k \leq z \leq k) &= \frac{1}{\sqrt{2\pi}} \int_{-k}^k e^{-\frac{1}{2}z^2} dz \\ &= \frac{2}{\sqrt{2\pi}} \int_0^k e^{-\frac{1}{2}z^2} dz \end{aligned}$$

Последний шаг следует из факта, что $e^{-\frac{1}{2}z^2}$ симметрична, и следовательно интеграл для диапазона $[-k, k]$ равен двойному интегралу для диапазона $[0, k]$. В итоге через подстановку переменной $t = \frac{z}{\sqrt{2}}$, получаем:

$$P(-k \leq z \leq k) = 2 \cdot P(0 \leq t \leq k/\sqrt{2}) = \frac{2}{\sqrt{\pi}} \int_0^{k/\sqrt{2}} e^{-t^2} dt = \operatorname{erf}\left(k/\sqrt{2}\right) \quad (2.32)$$

где **erf** – Гауссовская функция ошибки, определенная как:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Многофакторное нормальное распределение

Представлено d-мерной векторной случайной переменной $\mathbf{X} = (X_1, X_2, \dots, X_d)^T$, X имеет многофакторное нормальное распределение с мат. ожиданием μ и матрицей ковариации Σ , если ее совместная многофакторная функция распределения плотности вероятности представлена в виде:

$$f(\mathbf{x}|\mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp \left\{ -\frac{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}{2} \right\} \quad (2.33)$$

где $|\Sigma|$ - определитель матрицы ковариации.

В терминах однофакторного анализа

$$(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$

определяет расстояние, называемое расстоянием Махаланобиса, от точки $\mathbf{x}$ до мат. ожидания $\boldsymbol{\mu}$ распределения, включающего всю дисперсионно-ковариационную информацию об атрибутах. Расстояние Махаланобиса является обобщением расстояния Евклида, так как если $\boldsymbol{\Sigma} = \mathbf{I}$, где $\mathbf{I}$ - $d \times d$ матрица-индикатор (с диагональными элементами – 1 и остальными элементами - 0), имеем:

$$(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{I}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) = \|\mathbf{x}_i - \boldsymbol{\mu}\|^2$$

Расстояние Евклида игнорирует информацию о ковариации между атрибутами, в то время как расстояние Махаланобиса принимает ее во внимание. Стандартное многофакторное нормальное распределение имеет параметры $\boldsymbol{\mu} = \mathbf{0}$ и $\boldsymbol{\Sigma} = \mathbf{I}$. Рисунок 2.6а отображает плотность вероятности стандартного двухфакторного ($d = 2$) нормального распределения с параметрами:

$$\boldsymbol{\mu} = \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

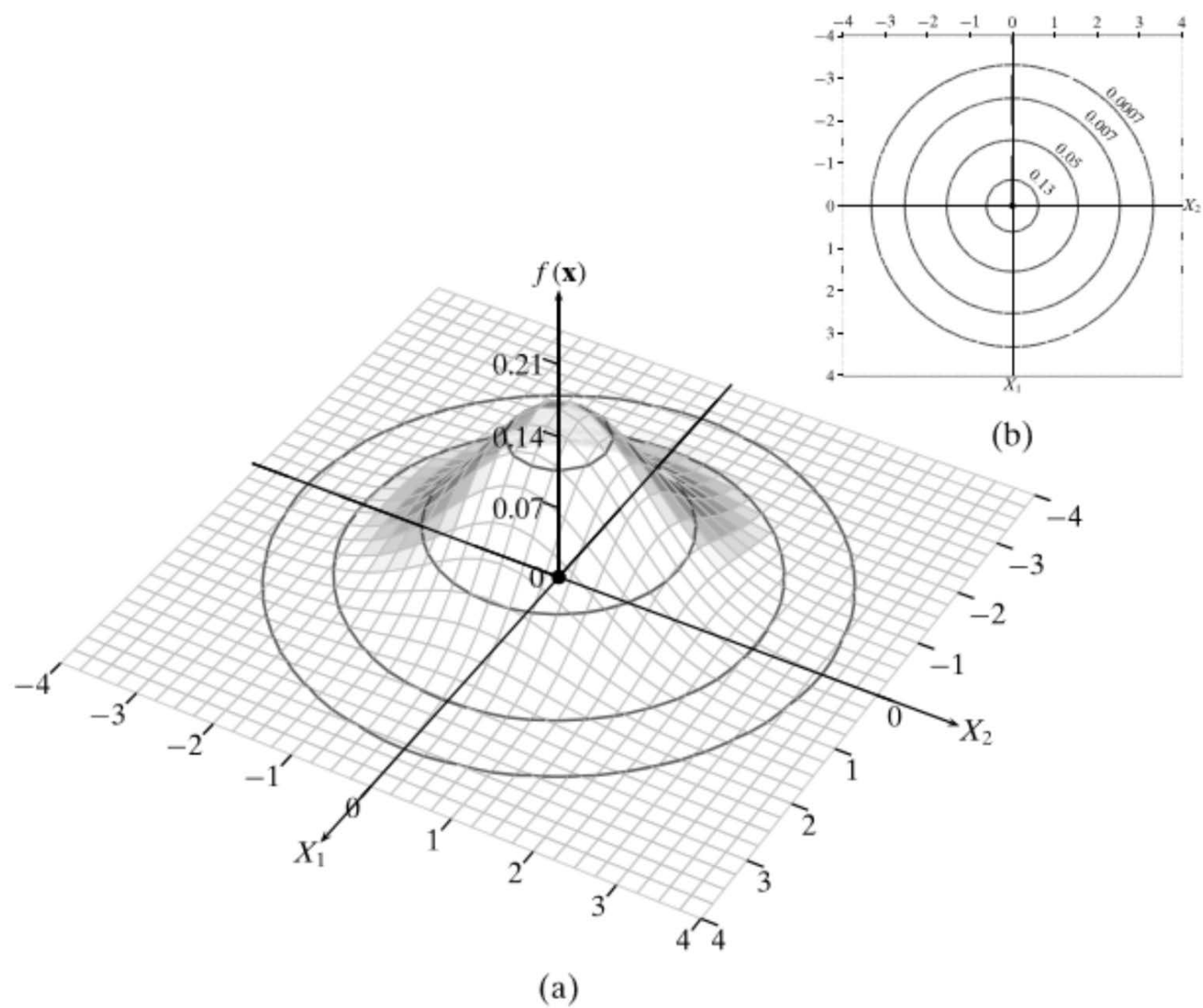


Figure 2.6. (a) Standard bivariate normal density and (b) its contour plot. Parameters: $\boldsymbol{\mu} = (0, 0)^T$, $\boldsymbol{\Sigma} = \mathbf{I}$.

Геометрия многофакторной нормали

Пусть μ - мат. ожидание и Σ - матрица ковариантности многомерного нормального распределения. Форма или геометрия нормального распределения становится ясной при рассмотрении собственного разложения ковариационной матрицы. Так как Σ это $d \times d$ симметричная положительная полуопределенная матрица, уравнение для собственного вектора Σ представлено в виде:

$$\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

Здесь λ_i это собственные значения Σ и вектор $\mathbf{u}_i \in \mathbb{R}^d$ это собственный вектор, относящийся к λ_i . Так как Σ симметричная положительная полуопределенная матрица, она имеет d реальных и не отрицательных собственных значений, которые можно расположить в порядке от наибольшего к наименьшему: $\lambda_1 \geq \lambda_2 \geq \dots \lambda_d \geq 0$.

Диагональная матрица Λ используется для записи собственных значений:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_d \end{pmatrix}$$

Собственные векторы - единичные (нормальные) и взаимно ортогональны, то есть они ортонормированы:

$$\mathbf{u}_i^T \mathbf{u}_i = 1 \quad \text{for all } i$$

$$\mathbf{u}_i^T \mathbf{u}_j = 0 \quad \text{for all } i \neq j$$

Собственные векторы можно объединить в ортогональную матрицу $\mathbf{U}$, определяемую как матрица с нормальными и ортогональными столбцами:

$$\mathbf{U} = \begin{pmatrix} | & | & \cdots & | \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_d \\ | & | & \cdots & | \end{pmatrix}$$

Тогда собственное разложение Σ можно компактно выразить следующим образом:

$$\Sigma = \mathbf{U} \Lambda \mathbf{U}^T$$

Полная и обобщенная дисперсия

Определитель ковариационной матрицы задается как $\det(\mathbf{\Sigma}) = \prod_{i=1}^d \lambda_i$.

Обобщенная дисперсия $\mathbf{\Sigma}$ - это произведение его собственных значений.

Так как след квадратной матрицы инвариантен к преобразованию подобия, например к смене базиса, то полная дисперсия $\text{var}(\mathbf{D})$ для набора данных $\mathbf{D}$ инвариантна, то есть

$$\text{var}(\mathbf{D}) = \text{tr}(\mathbf{\Sigma}) = \sum_{i=1}^d \sigma_i^2 = \sum_{i=1}^d \lambda_i = \text{tr}(\mathbf{\Lambda})$$

Другими словами:

$$\sigma_1^2 + \dots + \sigma_d^2 = \lambda_1 + \dots + \lambda_d.$$



Категориальные атрибуты
Однофакторный анализ

Данные состоят из значений для одного категориального атрибута X .

Пусть домен X состоит из m символьных значений $dom(X) = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$. Данные D - это $n \times 1$ символьная матрица, представленная как:

где каждый элемент - $x_i \in dom(X)$.

$$\mathbf{D} = \begin{pmatrix} X \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$



Категориальные атрибуты
Однофакторный анализ
Переменная Бернулли

Пусть категориальный атрибут X имеет домен $\{\alpha_1, \alpha_2\}$, где $m = 2$. X может быть представлена случайной величиной Бернулли, которая принимает два фиксированных значения - 1 и 0, согласно правилу:

$$X(v) = \begin{cases} 1 & \text{if } v = a_1 \\ 0 & \text{if } v = a_2 \end{cases}$$

Функция распределения вероятности (PMF) для X представлена в виде:

$$P(X = x) = f(x) = \begin{cases} p_1 & \text{if } x = 1 \\ p_0 & \text{if } x = 0 \end{cases}$$

где p_1 и p_0 параметры распределения, которые должны удовлетворять условию:

$$p_1 + p_0 = 1$$

Так как есть только один свободный параметр, обозначим $p_1 = p$, $\Rightarrow$
 $p_0 = 1 - p$. PMF случайной величины Бернулли X может быть записано в компактной форме:

$$P(X = x) = f(x) = p^x (1 - p)^{1-x}$$

При этом $P(X = 1) = p^1 (1 - p)^0 = p$ и $P(X = 0) = p^0 (1 - p)^1 = 1 - p$

Математическое ожидание и дисперсия

Мат. ожидание X : $\mu = E[X] = 1 \cdot p + 0 \cdot (1 - p) = p$

Дисперсия X :

$$\begin{aligned}\sigma^2 = \text{var}(X) &= E[X^2] - (E[X])^2 \\ &= (1^2 \cdot p + 0^2 \cdot (1 - p)) - p^2 = p - p^2 = p(1 - p)\end{aligned}\tag{3.1}$$

Выборочные мат. ожидание и дисперсия

Каждой символьной точке соответствует ее двоичное значение. И $\{x_1, x_2, \dots, x_n\}$ - случайная выборка из X , где каждый x_i распределен аналогично X . Выборочное среднее определяется так:

где n_1 число точек с $x_i = 1$ (равно числу появления символа a_1). Пусть $n_0 = n - n_1$ - количество точек с $x_i = 0$ в случайной выборке. Выборочная дисперсия определяется формулой:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{n_1}{n} = \hat{p}$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \\ &= \frac{n_1}{n} (1 - \hat{p})^2 + \frac{n - n_1}{n} (-\hat{p})^2 \\ &= \hat{p} (1 - \hat{p})^2 + (1 - \hat{p}) \hat{p}^2 \\ &= \hat{p} (1 - \hat{p}) (1 - \hat{p} + \hat{p}) \\ &= \hat{p} (1 - \hat{p}) \end{aligned}$$

Биномиальное распределение: Количество вхождений

Пусть X - переменная Бернулли, $\{x_1, x_2, \dots, x_n\}$ - случайная выборка размера n из X . Пусть N - случайная величина - количество вхождений символа a_1 (value $X = 1$) в выборке. Тогда N имеет биномиальное распределение:

$$f(N = n_1 | n, p) = \binom{n}{n_1} p^{n_1} (1 - p)^{n - n_1} \quad (3.3)$$

N - сумма n независимых случайных величин Бернулли x_i ,
распределенных также, как X :

$$N = \sum_{i=1}^n x_i$$

Среднее или ожидаемое число вхождений символа a_1 определяется как:

$$\mu_N = E[N] = E\left[\sum_{i=1}^n x_i\right] = \sum_{i=1}^n E[x_i] = \sum_{i=1}^n p = np$$

так как x_i - независимы, мат. ожидание N определяется как:

$$\sigma_N^2 = \text{var}(N) = \sum_{i=1}^n \text{var}(x_i) = \sum_{i=1}^n p(1 - p) = np(1 - p)$$



Категориальные атрибуты

Однофакторный анализ

Многомерная переменная Бернулли

Пусть X - категориальный атрибут с областью определения $\{a_1, a_2, \dots, a_m\}$. Мы можем представить X как m -размерную переменную Бернулли $\mathbf{X} = (A_1, A_2, \dots, A_m)^T$ где каждая A_i - переменная Бернулли с параметром p_i обозначающим вероятность наблюдения символа a_i . Так как X может принимать только одно символьное значение, если $X = a_i$, то $A_i = 1$, и $A_j = 0$ для всех $j \neq i$. Диапазон случайной величины X , таким образом представляет множество $\{0,1\}^m$ с дальнейшим ограничением - если $X = a_i$, то $X = e_i$, где e_i - i -й стандартный базисный вектор $e_i \in R^m$:

$$\mathbf{e}_i = (\overbrace{0, \dots, 0}^{i-1}, 1, \overbrace{0, \dots, 0}^{m-i})^T$$

Для e_i только те элементы равны 1 ($e_{ij} = 1$), где $i = j$, тогда как все остальные равны 0 ($e_{ij} = 0, \forall j \neq i$).

Многомерной переменная Бернулли - обобщение переменной Бернулли с двумя исходами до m исходов. Таким образом, категориальный атрибут X может быть представлен как многомерная переменная Бернулли X :

$$\mathbf{X}(v) = \mathbf{e}_i \text{ if } v = a_i$$

Диапазон $\mathbf{X}$ состоит из m различных векторов $\{e_1, e_2, \dots, e_m\}$, с PMF:

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = p_i$$

где p_i вероятность наблюдения значения a_i . Эти параметры должны удовлетворять условию:

$$\sum_{i=1}^m p_i = 1$$

PMF может быть записана компактно:

$$P(\mathbf{X} = \mathbf{e}_i) = f(\mathbf{e}_i) = \prod_{j=1}^m p_j^{e_{ij}}$$

Мат. ожидание

$$\mu = E[\mathbf{X}] = \sum_{i=1}^m \mathbf{e}_i f(\mathbf{e}_i) = \sum_{i=1}^m \mathbf{e}_i p_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} p_1 + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} p_m = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = \mathbf{p} \quad (3.5)$$

Выборочное среднее

Пусть каждая символическая точка $x_i \in D$ представлена переменной $x_i = X(x_i)$. Эти переменные $x_1, x_2, \dots, x_n$ представлены случайной выборкой из X . Выборочное среднее определяется путем подстановки в функцию распределения вероятности значения $1/n$ для каждой точки:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \sum_{i=1}^m \frac{n_i}{n} \mathbf{e}_i = \begin{pmatrix} n_1/n \\ n_2/n \\ \vdots \\ n_m/n \end{pmatrix} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_m \end{pmatrix} = \hat{\mathbf{p}} \quad (3.6)$$

где n_i число вхождений значений вектора $\mathbf{e}_i$ в выборке, что эквивалентно числу вхождений символа a_i . Кроме того:

$$\sum_{i=1}^m n_i = n$$

Так как X может принимать только m различных значений $\mathbf{e}_i$, и количество вхождений для каждого значения должно быть равно размеру выборки n .

Матрица ковариантности

Многомерная m -мерная переменная Бернулли - это вектор из m переменных Бернулли. Например, $\mathbf{X} = (A_1, A_2, \dots, A_m)^T$, где A_i - переменные Бернулли, связанные с символами a_i . Информация о дисперсии и ковариации между этими переменными Бернулли составляют ковариационную матрицу. Дисперсия по каждой переменной Бернулли A_i :

$$\sigma_i^2 = \text{var}(A_i) = p_i(1 - p_i)$$

Затем рассмотрим ковариацию между A_i и A_j :

$$\sigma_{ij} = E[A_i A_j] - E[A_i] \cdot E[A_j] = 0 - p_i p_j = -p_i p_j$$

так как $E[A_i A_j] = 0$, то A_i и A_j не могут быть одновременно равными 1 в момент времени, следовательно $A_i A_j = 0$. Этот же факт приводит к отрицательной связи между A_i и A_j .

Из предыдущих выражений для дисперсии и ковариации, ковариационная матрица размера $m \times m$ для X задается как:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \dots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_m \\ -p_1p_2 & p_2(1-p_2) & \dots & -p_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_m & -p_2p_m & \dots & p_m(1-p_m) \end{pmatrix}$$

P - $m \times m$ диагональная матрица:

$$\mathbf{P} = \text{diag}(\mathbf{p}) = \text{diag}(p_1, p_2, \dots, p_m) = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_m \end{pmatrix}$$

Или компактная запись матрицы ковариантности:

$$\Sigma = \mathbf{P} - \mathbf{p} \cdot \mathbf{p}^T \quad (3.8)$$

Выборочная матрица ковариантности

Выборочная матрица ковариантности может быть напрямую получена из (3.8):

$$\hat{\Sigma} = \hat{\mathbf{P}} - \hat{\mathbf{p}} \cdot \hat{\mathbf{p}}^T \quad (3.9)$$

где $\hat{\mathbf{P}} = \text{diag}(\hat{\mathbf{p}})$

и $\hat{\mathbf{p}} = \hat{\mu} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_m)^T$

определяют эмпирическую функцию распределения вероятности для X .

Мультиномиальное распределение: количество вхождений

Дана многомерная переменная Бернулли X и случайная выборка $\{x_1, x_2, \dots, x_n\}$ из X . Пусть N_i случайная переменная, связанная с числом вхождения символа a_i в выборку, и пусть $N = (N_1, N_2, \dots, N_m)^T$ - векторная случайная величина, соответствующая совместному распределению числа вхождений по всем символам. N имеет мультиномиальное распределение:

$$f(\mathbf{N} = (n_1, n_2, \dots, n_m) \mid \mathbf{p}) = \binom{n}{n_1 n_2 \dots n_m} \prod_{i=1}^m p_i^{n_i}$$

Мы видим, что это прямое обобщение биномиального распределения:

$$\binom{n}{n_1 n_2 \dots n_m} = \frac{n!}{n_1! n_2! \dots n_m!}$$

Что обозначает количество способов выбора n_i вхождения каждого символа a_i из выборки размера n , с $\sum_{i=1}^m n_i = n$

Среднее значение и матрица ковариации $\mathbf{N}$ задаются как n , умноженное на среднее значение и матрицу ковариации $\mathbf{X}$. То есть среднее значение $\mathbf{N}$ задается как

$$\boldsymbol{\mu}_{\mathbf{N}} = E[\mathbf{N}] = nE[\mathbf{X}] = n \cdot \boldsymbol{\mu} = n \cdot \mathbf{p} = \begin{pmatrix} np_1 \\ \vdots \\ np_m \end{pmatrix}$$

а его ковариационная матрица имеет вид

$$\boldsymbol{\Sigma}_{\mathbf{N}} = n \cdot (\mathbf{P} - \mathbf{p}\mathbf{p}^T) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_m \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_m \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_m & -np_2p_m & \cdots & np_m(1-p_m) \end{pmatrix}$$

Аналогичным образом выборочное среднее и ковариационная матрица для $\mathbf{N}$ задаются как

$$\hat{\boldsymbol{\mu}}_{\mathbf{N}} = n\hat{\mathbf{p}}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{N}} = n(\hat{\mathbf{P}} - \hat{\mathbf{p}}\hat{\mathbf{p}}^T)$$

Пусть данные содержат два категориальных атрибута - X_1 и X_2 , где

$$\text{dom}(X_1) = \{a_{11}, a_{12}, \dots, a_{1m_1}\}$$

$$\text{dom}(X_2) = \{a_{21}, a_{22}, \dots, a_{2m_2}\}$$

Даны n категориальных точек вида $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ где $x_{i1} \in \text{dom}(X_1)$ и $x_{i2} \in \text{dom}(X_2)$.

Таким образом, набор данных представляет собой символьную матрицу данных размером $n \times 2$:

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Мы можем представить X_1 и X_2 как многомерные переменные Бернулли $\mathbf{X}_1$ и $\mathbf{X}_2$ с размерностями m_1 и m_2 соответственно. Функция распределения вероятности для $\mathbf{X}_1$ и $\mathbf{X}_2$:

$$P(\mathbf{X}_1 = \mathbf{e}_{1i}) = f_1(\mathbf{e}_{1i}) = p_i^1 = \prod_{k=1}^{m_1} (p_i^1)^{e_{ik}^1}$$

$$P(\mathbf{X}_2 = \mathbf{e}_{2j}) = f_2(\mathbf{e}_{2j}) = p_j^2 = \prod_{k=1}^{m_2} (p_j^2)^{e_{jk}^2}$$

где $\mathbf{e}_{1i}$ - i -й стандартный базисный вектор в R^{m_1} (для атрибута X_1), чей k -й компонент - e_{ik}^1 , и $\mathbf{e}_{2j}$ - j -й стандартный базисный вектор в R^{m_2} (для атрибута X_2), чей k -й компонент - e_{jk}^2 . Параметр p_i^1 определяет вероятность наблюдения символа a_{1i} , и p_j^2 определяет вероятность наблюдения символа a_{2j} . Вместе они должны удовлетворять условиям: $\sum_{i=1}^{m_1} p_i^1 = 1$ и $\sum_{j=1}^{m_2} p_j^2 = 1$.

Совместное распределение X_1 и X_2 представлено как $d' = m_1 + m_2$ размерная векторная переменная $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, заданная как:

$$\mathbf{X}((v_1, v_2)^T) = \begin{pmatrix} \mathbf{X}_1(v_1) \\ \mathbf{X}_2(v_2) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1i} \\ \mathbf{e}_{2j} \end{pmatrix}$$

где $v_1 = a_{1i}$ и $v_2 = a_{2j}$. Диапазон $\mathbf{X}$ таким образом состоит из $m_1 \times m_2$ различных пар векторных значений $(\mathbf{e}_{1i}, \mathbf{e}_{2j})^T$, где $1 \leq i \leq m_1$ и $1 \leq j \leq m_2$. Совместная PMF:

$$P(\mathbf{X} = (\mathbf{e}_{1i}, \mathbf{e}_{2j})^T) = f(\mathbf{e}_{1i}, \mathbf{e}_{2j}) = p_{ij} = \prod_{r=1}^{m_1} \prod_{s=1}^{m_2} p_{ir}^{e_{ir}^1 \cdot e_{js}^2}$$



Категориальные атрибуты
Двухфакторный анализ

Пусть p_{ij} - вероятность наблюдения символьной пары (a_{1i}, a_{2j}) . И $\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} p_{ij} = 1$. Совместная PMF для X может быть представлена в виде $m_1 \times m_2$ матрицы:

$$\mathbf{P}_{12} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m_2} \\ p_{21} & p_{22} & \cdots & p_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_11} & p_{m_12} & \cdots & p_{m_1m_2} \end{pmatrix} \quad (3.10)$$

Мат. ожидание

Двухмерное мат. ожидание:

$$\mu = E[\mathbf{X}] = E\left[\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}\right] = \begin{pmatrix} E[\mathbf{X}_1] \\ E[\mathbf{X}_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{pmatrix}$$

Где $\mu_1 = \mathbf{p}_1 = (p_1^1, \dots, p_{m_1}^1)^T$ и $\mu_2 = \mathbf{p}_2 = (p_1^2, \dots, p_{m_2}^2)^T$ вектора мат. ожиданий для X_1 и X_2 . Вектора p_1 и p_2 - функции распределения вероятностей для X_1 и X_2 соответственно.

Выборочное мат. ожидание

Выборочное среднее также может быть обобщено постановкой в каждую точку значение вероятности - $1/n$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{m_1} n_i^1 \mathbf{e}_{1i} \\ \sum_{j=1}^{m_2} n_j^2 \mathbf{e}_{2j} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \\ n_1^2 \\ \vdots \\ n_{m_2}^2 \end{pmatrix} = \begin{pmatrix} \hat{p}_1^1 \\ \vdots \\ \hat{p}_{m_1}^1 \\ \hat{p}_1^2 \\ \vdots \\ \hat{p}_{m_2}^2 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \hat{\mathbf{p}}_2 \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_1 \\ \hat{\boldsymbol{\mu}}_2 \end{pmatrix}$$

Матрица ковариантности

Матрица ковариантности для X это $d' \times d' = (m_1 + m_2) \times (m_1 + m_2)$ матрица вида:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

где Σ_{11} это $m_1 \times m_1$ матрица ковариантности для X_1 , и Σ_{22} это $m_2 \times m_2$ матрица ковариантности для X_2 :

$$\Sigma_{11} = \mathbf{P}_1 - \mathbf{p}_1 \mathbf{p}_1^T$$

$$\Sigma_{22} = \mathbf{P}_2 - \mathbf{p}_2 \mathbf{p}_2^T$$

где $\mathbf{P}_1 = \text{diag}(\mathbf{p}_1)$ и $\mathbf{P}_2 = \text{diag}(\mathbf{p}_2)$. Далее, Σ_{12} - $m_1 \times m_2$ матрица ковариантности между X_1 и X_2 , представленная как:

Σ_{12} - $m_1 \times m_2$ матрица ковариантности между $\mathbf{X}_1$ и $\mathbf{X}_2$:

$$\begin{aligned}\Sigma_{12} &= E[(\mathbf{X}_1 - \boldsymbol{\mu}_1)(\mathbf{X}_2 - \boldsymbol{\mu}_2)^T] \\ &= E[\mathbf{X}_1 \mathbf{X}_2^T] - E[\mathbf{X}_1]E[\mathbf{X}_2]^T \\ &= \mathbf{P}_{12} - \boldsymbol{\mu}_1 \boldsymbol{\mu}_2^T \\ &= \mathbf{P}_{12} - \mathbf{p}_1 \mathbf{p}_2^T \\ &= \begin{pmatrix} p_{11} - p_1^1 p_1^2 & p_{12} - p_1^1 p_2^2 & \cdots & p_{1m_2} - p_1^1 p_{m_2}^2 \\ p_{21} - p_2^1 p_1^2 & p_{22} - p_2^1 p_2^2 & \cdots & p_{2m_2} - p_2^1 p_{m_2}^2 \\ \vdots & \vdots & \ddots & \vdots \\ p_{m_1 1} - p_{m_1}^1 p_1^2 & p_{m_1 2} - p_{m_1}^1 p_2^2 & \cdots & p_{m_1 m_2} - p_{m_1}^1 p_{m_2}^2 \end{pmatrix}\end{aligned}$$

где $\mathbf{P}_{12}$ представляет совместную PMF для $\mathbf{X}$.

Выборочная матрица ковариантности

Выборочная матрица ковариантности представлена в виде:

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{12}^T & \hat{\Sigma}_{22} \end{pmatrix}$$

где

$$\hat{\Sigma}_{11} = \hat{\mathbf{P}}_1 - \hat{\mathbf{p}}_1 \hat{\mathbf{p}}_1^T$$

$$\hat{\Sigma}_{22} = \hat{\mathbf{P}}_2 - \hat{\mathbf{p}}_2 \hat{\mathbf{p}}_2^T$$

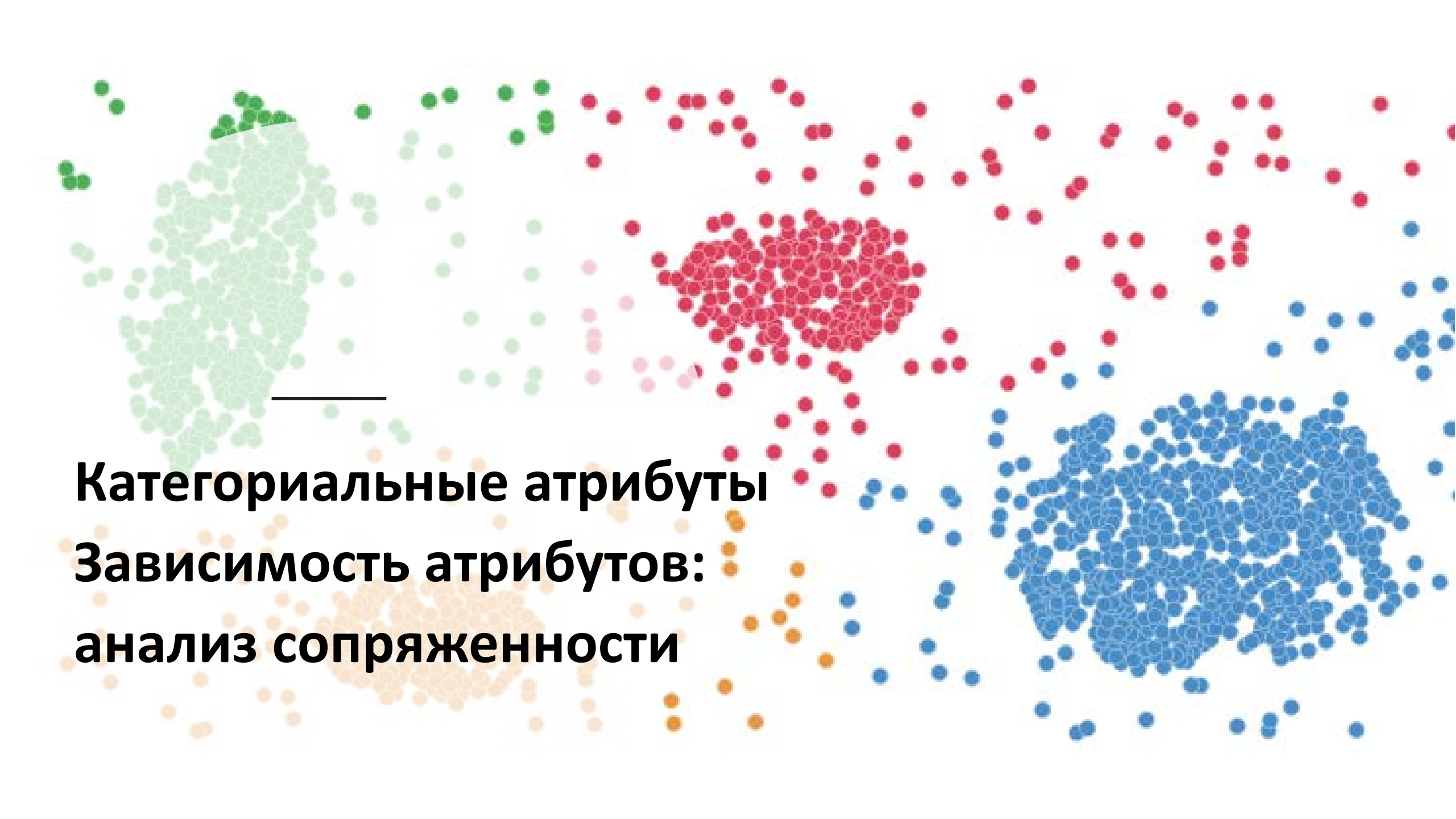
$$\hat{\Sigma}_{12} = \hat{\mathbf{P}}_{12} - \hat{\mathbf{p}}_1 \hat{\mathbf{p}}_2^T$$

Здесь $\hat{\mathbf{P}}_1 = \text{diag}(\hat{\mathbf{p}}_1)$ и $\hat{\mathbf{P}}_2 = \text{diag}(\hat{\mathbf{p}}_2)$, и $\hat{\mathbf{p}}_1$ и $\hat{\mathbf{p}}_2$ определяют эмпирическую функцию распределения вероятностей для X_1 и X_2 соответственно. $\hat{\mathbf{P}}_{12}$ определяет совместную PMF для X_1 и X_2 :

$$\hat{\mathbf{P}}_{12}(i, j) = \hat{f}(\mathbf{e}_{1i}, \mathbf{e}_{2j}) = \frac{1}{n} \sum_{k=1}^n I_{ij}(\mathbf{x}_k) = \frac{n_{ij}}{n} = \hat{p}_{ij} \quad (3.13)$$

где I_{ij} - переменная-индикатор:

$$I_{ij}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } \mathbf{x}_{k1} = \mathbf{e}_{1i} \text{ and } \mathbf{x}_{k2} = \mathbf{e}_{2j} \\ 0 & \text{otherwise} \end{cases}$$



Категориальные атрибуты
Зависимость атрибутов:
анализ сопряженности

Таблица сопряженности

Таблица сопряженности для X_1 и X_2 это $m_1 \times m_2$ матрица количеств наблюдения n_{ij} для всех пар значений (e_{1i}, e_{2j}) :

$$\mathbf{N}_{12} = n \cdot \hat{\mathbf{P}}_{12} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1m_2} \\ n_{21} & n_{22} & \cdots & n_{2m_2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m_1 1} & n_{m_1 2} & \cdots & n_{m_1 m_2} \end{pmatrix}$$

где $\hat{\mathbf{P}}_{12}$ эмпирическая совместная PMF для X_1 и X_2 . Таблица сопряженности затем дополняется маргинальными частотами по строкам и столбцам:

$$\mathbf{N}_1 = n \cdot \hat{\mathbf{p}}_1 = \begin{pmatrix} n_1^1 \\ \vdots \\ n_{m_1}^1 \end{pmatrix} \qquad \mathbf{N}_2 = n \cdot \hat{\mathbf{p}}_2 = \begin{pmatrix} n_1^2 \\ \vdots \\ n_{m_2}^2 \end{pmatrix}$$

Маргинальные частоты по строкам и столбцам, а также размер выборки удовлетворяют ограничениям:

$$n_i^1 = \sum_{j=1}^{m_2} n_{ij} \qquad n_j^2 = \sum_{i=1}^{m_1} n_{ij} \qquad n = \sum_{i=1}^{m_1} n_i^1 = \sum_{j=1}^{m_2} n_j^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij}$$

χ^2 - статистика и проверка гипотез

При нулевой гипотезе предполагается, что X_1 и X_2 независимы, и их совместная функция распределения вероятностей имеет вид:

$$\hat{p}_{ij} = \hat{p}_i^1 \cdot \hat{p}_j^2$$

Исходя из предположения о независимости, ожидаемая частота для каждой пары значений определяется как:

$$e_{ij} = n \cdot \hat{p}_{ij} = n \cdot \hat{p}_i^1 \cdot \hat{p}_j^2 = n \cdot \frac{n_i^1}{n} \cdot \frac{n_j^2}{n} = \frac{n_i^1 n_j^2}{n} \quad (3.14)$$

χ^2 - статистика определяет количественно разницу между наблюдаемыми и ожидаемыми значениями для каждой пары значений. Она определяется следующим образом:

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad (3.15)$$

Для χ^2 - статистики известно, что ее выборочное распределение определяется функцией плотности χ^2 с q степенями свободы:

$$f(x|q) = \frac{1}{2^{q/2}\Gamma(q/2)} x^{\frac{q}{2}-1} e^{-\frac{x}{2}} \quad (3.16)$$

где Гамма-функция Γ определена как:

$$\Gamma(k > 0) = \int_0^{\infty} x^{k-1} e^{-x} dx \quad (3.17)$$

Таким образом общее число степеней свободы равно:

$$\begin{aligned} q &= |dom(X_1)| \times |dom(X_2)| - (|dom(X_1)| + |dom(X_2)|) + 1 \\ &= m_1 m_2 - m_1 - m_2 + 1 \\ &= (m_1 - 1)(m_2 - 1) \end{aligned}$$

p-значение

p-значение статистики θ определяется как вероятность получения значения, по крайней мере, столь же экстремального, как наблюдаемое значение z при нулевой гипотезе, определяемой как

$$p\text{-value}(z) = P(\theta \geq z) = 1 - F(\theta)$$

где $F(\theta)$ – кумулятивная функция распределения статистики. Нулевая гипотеза не применима, если p-значение ниже некоторого уровня значимости - α . Для заданного уровня значимости α , значение проверочной статистики z , с p-значением $p\text{-value}(z) = \alpha$, называется критическим значением. Альтернативный тест для отклонения нулевой гипотезы - проверить, что $\chi^2 > z$, так как в этом случае p-значение наблюдаемого χ^2 значения ограничен α , так как, $p\text{-value}(\chi^2) \leq p\text{-value}(z) = \alpha$. Значение $1 - \alpha$ также называют уровнем уверенности.



Пусть данные содержат d категориальных атрибутов $\mathbf{X}_j$ ($1 \leq j \leq d$) где $\text{dom}(\mathbf{X}_j) = \{a_{j1}, a_{j2}, \dots, a_{jm_j}\}$. И даны n категориальных точек вида $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ где $x_{ij} \in \text{dom}(\mathbf{X}_j)$. Тогда набор данных представляет собой $n \times d$ символьную матрицу:

$$\mathbf{D} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_d \\ x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Каждый атрибут X_i представлен как m_i - размерная многомерная величина Бернулли X_i , и их совместная функция распределения - $d' = \sum_{j=1}^d m_j$ мерный вектор случайных величин:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_d \end{pmatrix}$$

Каждая категориальная точка данных $v_i = (v_1, v_2, \dots, v_d)^T$ представлена как d' -размерный двоичный вектор:
где $v_i = a_{ik_i}$, для k_i -го символа X_i .

Здесь e_{ik_i} - k_i стандартный базисный вектор в R^m .

$$\mathbf{X}(\mathbf{v}) = \begin{pmatrix} \mathbf{X}_1(v_1) \\ \vdots \\ \mathbf{X}_d(v_d) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_{1k_1} \\ \vdots \\ \mathbf{e}_{dk_d} \end{pmatrix}$$

Мат. ожидание

Обобщая двумерный случай, среднее и выборочное среднее для X представлены как:

$$\mu = E[\mathbf{X}] = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_d \end{pmatrix}$$

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_d \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{p}}_1 \\ \vdots \\ \hat{\mathbf{p}}_d \end{pmatrix}$$

Где $\mathbf{p}_i = (p_1, p_2, \dots, p_{m_i})^T$ - PMF для X_i , и $\hat{\mathbf{p}}_i = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_{m_i})^T$ - эмпирическая PMF для X_i .

Матрица ковариантности

Ковариационная матрица для X и ее оценка из выборки задаются как матрицы $d' \times d'$:

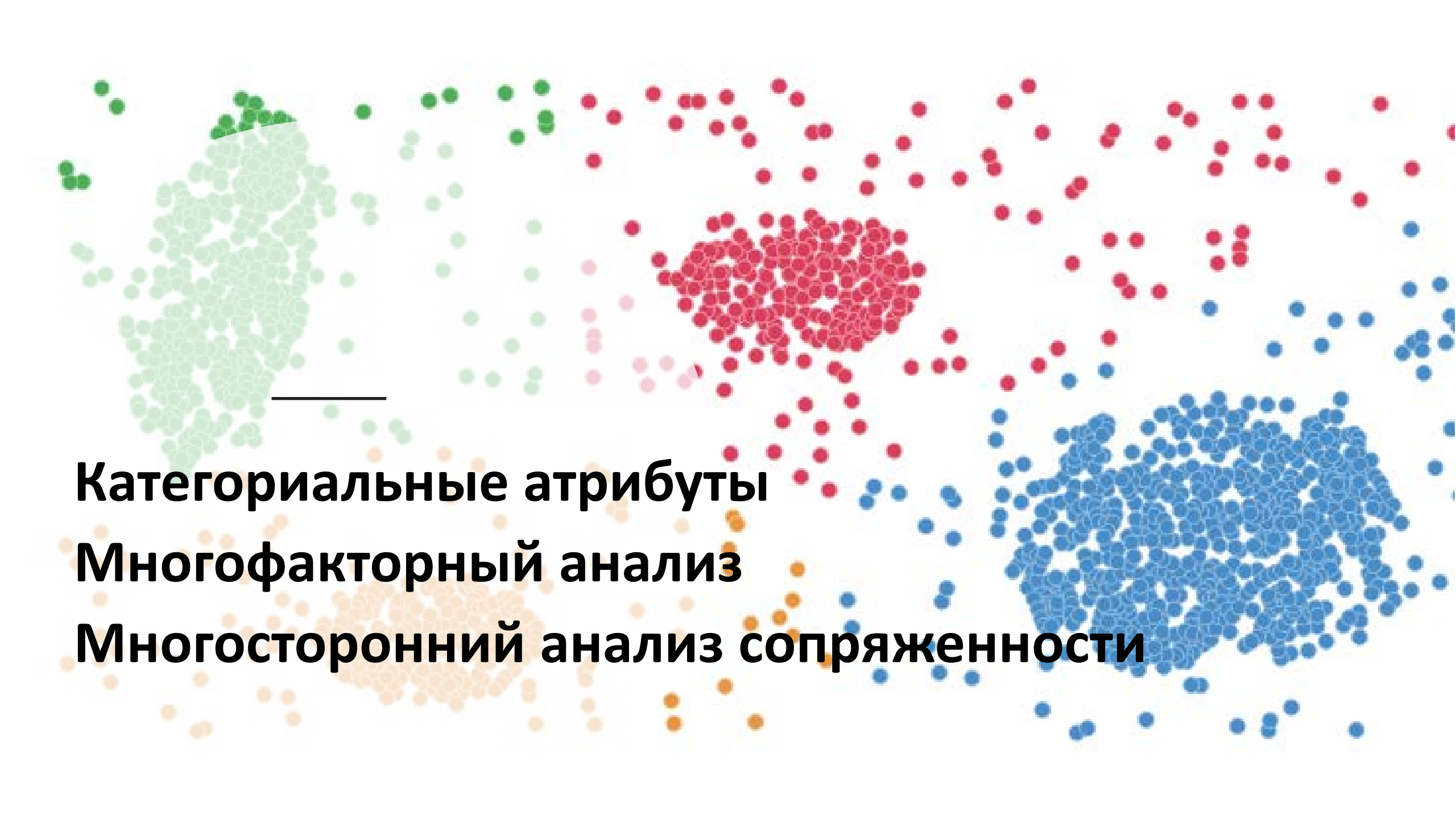
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1d} \\ \Sigma_{12}^T & \Sigma_{22} & \cdots & \Sigma_{2d} \\ \cdots & \cdots & \ddots & \cdots \\ \Sigma_{1d}^T & \Sigma_{2d}^T & \cdots & \Sigma_{dd} \end{pmatrix} \quad \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} & \cdots & \hat{\Sigma}_{1d} \\ \hat{\Sigma}_{12}^T & \hat{\Sigma}_{22} & \cdots & \hat{\Sigma}_{2d} \\ \cdots & \cdots & \ddots & \cdots \\ \hat{\Sigma}_{1d}^T & \hat{\Sigma}_{2d}^T & \cdots & \hat{\Sigma}_{dd} \end{pmatrix}$$

где $d' = \sum_{i=1}^d m_i$, и Σ_{ij} (and $\hat{\Sigma}_{ij}$) - $m_i \times m_j$ матрица ковариантности (и ее оценка)

Для атрибутов X_i и X_j :

$$\Sigma_{ij} = \mathbf{P}_{ij} - \mathbf{p}_i \mathbf{p}_j^T \quad \hat{\Sigma}_{ij} = \hat{\mathbf{P}}_{ij} - \hat{\mathbf{p}}_i \hat{\mathbf{p}}_j^T \quad (3.18)$$

Здесь $\mathbf{P}_{ij}$ - совместная PMF и $\hat{\mathbf{P}}_{ij}$ - эмпирическая совместная PMF для X_i и X_j .



A scatter plot illustrating data clustering. The plot features four distinct clusters of points, each represented by a different color: green, red, blue, and orange. The green cluster is located in the upper-left quadrant, the red cluster is in the upper-middle, the blue cluster is in the lower-right, and the orange cluster is in the lower-left. Each cluster consists of a dense core of points with some outliers scattered around. A horizontal line is drawn across the middle of the plot, separating the upper and lower clusters.

Категориальные атрибуты

Многофакторный анализ

Многосторонний анализ сопряженности

Для многостороннего анализа зависимости требуется определить эмпирическую совместную функцию распределения вероятности для X :

$$\hat{f}(\mathbf{e}_{1i_1}, \mathbf{e}_{2i_2}, \dots, \mathbf{e}_{di_d}) = \frac{1}{n} \sum_{k=1}^n I_{i_1 i_2 \dots i_d}(\mathbf{x}_k) = \frac{n_{i_1 i_2 \dots i_d}}{n} = \hat{p}_{i_1 i_2 \dots i_d}$$

где $I_{i_1 i_2 \dots i_d}$ - переменная-индикатор.

$$I_{i_1 i_2 \dots i_d}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } x_{k1} = \mathbf{e}_{1i_1}, x_{k2} = \mathbf{e}_{2i_2}, \dots, x_{kd} = \mathbf{e}_{di_d} \\ 0 & \text{otherwise} \end{cases}$$

Сумма $I_{i_1 i_2 \dots i_d}$ для всех n точек в выборке дает количество вхождений $n_{i_1 i_2 \dots i_d}$, символьного вектора $(a_{1i_1}, a_{2i_2}, \dots, a_{di_d})$. Деление числа вхождений на размер выборки дает вероятность наблюдения этих символов. Используя обозначения $i = (i_1, i_2, \dots, i_d)$ для обозначения индексного кортежа можно записать совместную эмпирическую PMF как d -мерную матрицу $\hat{\mathbf{P}}$ размера $m_1 \times m_2 \times \dots \times m_d = \prod_{i=1}^d m_i$, представленную в виде:

$$\hat{\mathbf{P}}(\mathbf{i}) = \{\hat{p}_{\mathbf{i}}\} \text{ for all index tuples } \mathbf{i}, \text{ with } 1 \leq i_1 \leq m_1, \dots, 1 \leq i_d \leq m_d$$

где $\hat{p}_i = \hat{p}_{i_1 i_2 \dots i_d}$. d-мерная таблица сопряженности выглядит так:

$$\mathbf{N} = n \times \hat{\mathbf{P}} = \{n_{\mathbf{i}}\} \text{ for all index tuples } \mathbf{i}, \text{ with } 1 \leq i_1 \leq m_1, \dots, 1 \leq i_d \leq m_d$$

где $n_i = n_{i_1 i_2 \dots i_d}$. Таблица сопряженности дополняется векторами маргинальной частоты N_i для всех d атрибутов X_i :

$$\mathbf{N}_i = n \hat{\mathbf{p}}_i = \begin{pmatrix} n_1^i \\ \vdots \\ n_{m_i}^i \end{pmatrix}$$

где $\hat{\mathbf{p}}_i$ - эмпирическая PMF для X_i .

χ^2 -проверка

Возможно d -стороннюю зависимость между d категориальными атрибутами, используя нулевую гипотезу H_0 что она d -сторонне независима. Альтернативная гипотеза H_1 что она не d -сторонне независима. При этом d -сторонний анализ сопряженности показывает, являются ли все d -атрибуты, взятые вместе, независимыми или нет. В общем случае возможно проведение k -стороннего анализа сопряженности, чтобы проверить, является ли какое-либо подмножество $k \leq d$ атрибутов независимым или нет. При нулевой гипотезе ожидаемое количество вхождений кортежа символа $(a_{1i_1}, a_{2i_2}, \dots, a_{di_d})$ представлено как:

$$e_i = n \cdot \hat{p}_i = n \cdot \prod_{j=1}^d \hat{p}_{ij}^j = \frac{n_{i_1}^1 n_{i_2}^2 \dots n_{i_d}^d}{n^{d-1}} \quad (3.19)$$

Статистика Хи-квадрат показывает разницу между наблюдаемым числом n_i и его ожидаемым значением e_i :

$$\chi^2 = \sum_{\mathbf{i}} \frac{(n_{\mathbf{i}} - e_{\mathbf{i}})^2}{e_{\mathbf{i}}} = \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \dots \sum_{i_d=1}^{m_d} \frac{(n_{i_1, i_2, \dots, i_d} - e_{i_1, i_2, \dots, i_d})^2}{e_{i_1, i_2, \dots, i_d}} \quad (3.20)$$

Общее число степеней свободы определяется как:

$$\begin{aligned} q &= \prod_{i=1}^d |dom(X_i)| - \sum_{i=1}^d |dom(X_i)| + (d - 1) \\ &= \left(\prod_{i=1}^d m_i \right) - \left(\sum_{i=1}^d m_i \right) + d - 1 \end{aligned} \quad (3.21)$$

Чтобы отклонить нулевую гипотезу, требуется проверить, является ли р-значение наблюдаемого χ^2 значения меньше желаемого уровня значимости α (например, $\alpha = 0,01$) с использованием плотности распределения Хи-квадрат с q степенями свободы.



Категориальные атрибуты
Многофакторный анализ
Расстояние и угол

При рассмотрении категориальных атрибутов как многомерных переменных Бернулли можно вычислить расстояние или угол между любыми двумя точками $\mathbf{x}_i$ и $\mathbf{x}_j$:

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{e}_{1i_1} \\ \vdots \\ \mathbf{e}_{di_d} \end{pmatrix} \quad \mathbf{x}_j = \begin{pmatrix} \mathbf{e}_{1j_1} \\ \vdots \\ \mathbf{e}_{dj_d} \end{pmatrix}$$

Меры расстояния и сходства зависят от количества совпадающих и несовпадающих значений (или символов) в атрибутах $\mathbf{X}_k$. Например, возможно вычислить количество совпадающих значений s с помощью скалярного произведения:

$$s = \mathbf{x}_i^T \mathbf{x}_j = \sum_{k=1}^d (\mathbf{e}_{ki_k})^T \mathbf{e}_{kj_k}$$

С другой стороны, количество несовпадений просто $d - s$. Также полезна норма каждой точки:

$$\|\mathbf{x}_i\|^2 = \mathbf{x}_i^T \mathbf{x}_i = d$$

Расстояние Евклида

Расстояние Евклида между $\mathbf{x}_i$ и $\mathbf{x}_j$ определяется как:

$$\delta(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\| = \sqrt{\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{x}_j + \mathbf{x}_j^T \mathbf{x}_j} = \sqrt{2(d - s)}$$

Расстояние Хемминга

Расстояние Хемминга между $\mathbf{x}_i$ и $\mathbf{x}_j$ как число несовпадающих значений:

$$\delta_H(\mathbf{x}_i, \mathbf{x}_j) = d - s = \frac{1}{2} \delta(\mathbf{x}_i, \mathbf{x}_j)^2$$

Косинусный коэффициент

Косинусный коэффициент угла между $\mathbf{x}_i$ и $\mathbf{x}_j$ определяется как:

$$\cos \theta = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|} = \frac{s}{d}$$

Коэффициент Жаккара

Коэффициент Жаккара - часто используемая мера сходства между двумя категориальными точками. Он определяется как отношение количества совпадающих значений к количеству несовпадающих, которые появляются в обоих x_i и x_j для всех d атрибутов:

$$J(\mathbf{x}_i, \mathbf{x}_j) = \frac{s}{2(d - s) + s} = \frac{s}{2d - s}$$



Категориальные атрибуты
Многофакторный анализ
Дискретизация

Дискретизация (бининг) преобразует числовые атрибуты в категориальные. Для дискретизации числового атрибута X и случайную выборку $\{x_i\}_{i=1}^n$ размера n , полученную из X , требуется разделить диапазон значений X на k последовательных интервалов, также называемых ячейками, путем нахождения $k - 1$ граничных значений $v_1, v_2, \dots, v_{k-1}$ которые дают k интервалов:

$$[x_{\min}, v_1], (v_1, v_2], \dots, (v_{k-1}, x_{\max}]$$

где экстремумы диапазона X задаются как $x_{\min} = \min_i \{x_i\}$ $x_{\max} = \max_i \{x_i\}$

Равноширокие интервалы

Самый простой подход бининга - разделить диапазон X на k интервалов одинаковой ширины.

Ширина интервала - это просто диапазон X , деленный на k :

Таким образом, граница i -го интервала задается как:

$$v_i = x_{\min} + i w, \text{ for } i = 1, \dots, k - 1$$

$$w = \frac{x_{\max} - x_{\min}}{k}$$

Равночастотные интервалы

При равночастотном бининге диапазон X делится на интервалы, содержащие (приблизительно) равное количество точек. При этом $\hat{F}^{-1}(q) = \min\{x \mid P(X \leq x) \geq q\}$, для $q \in [0,1]$. В частности, требуется, чтобы каждый интервал содержал $1/k$ распределенной вероятности; поэтому границы интервалов задаются следующим образом:

$$v_i = \hat{F}^{-1}(i/k) \text{ for } i = 1, \dots, k-1$$