

Многомерные статистические методы 2021

Формулы по курсу. Часть I

Архипова Марина Юрьевна,
профессор департамента статистики и анализа данных
Золотарев Антон Олегович, БСТ182

19 января 2021 г.

Содержание

1	Предварительный анализ данных	3
1.1	Диагностика выбросов	3
1.2	Визуализация выбросов и распределения	4
1.2.1	Ящичковая диаграмма	4
1.2.2	Stemplot/Листовая диаграмма	4
1.2.3	Dotplot/Точечная диаграмма	4
2	Корреляционный анализ	5
2.1	Парные коэффициенты корреляции	5
2.2	Частные коэффициенты корреляции	6
2.3	Множественные коэффициенты корреляции	6
3	Регрессионный анализ	8
3.1	Отклонение выходов модели от реальных значений наблюдений	9
3.2	MSE и дисперсия коэффициентов	10
3.3	Проверка гипотезы о значимости коэффициентов	11
3.4	Доверительные интервалы	11
4	Метод главных компонент	12
5	Кластерный анализ	14
5.1	Методы вычисления расстояния между наблюдениями	14
5.2	Методы вычисления расстояния между группами наблюдений	14
5.3	Функционалы качества	15
5.4	Иерархические методы классификации (дендрограммы)	15
5.5	Метод классификации через k-средних	16
6	Дискриминантный анализ	18
7	Решающие деревья	18
7.1	Задача классификации	18
7.2	Задача регрессии	18
8	Секретные темы, название которых вы узнаете позже...	18
	Математико-статистические таблицы распределений	

1 Предварительный анализ данных

Коэффициент вариации:

$$V_s = \frac{S}{\bar{x}} \cdot 100\%$$

Относительное линейное отклонение:

$$K_d = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})}{\bar{x}} \cdot 100\%$$

Квартильное отклонение:

$$d_k = \frac{Q_3 - Q_1}{2}$$

Относительный показатель квартильной вариации:

$$K_Q = \frac{d_k}{Me} = \frac{Q_3 - Q_1}{2 \cdot Q_2}$$

Интерквартильное отклонение:

$$IQR = Q_3 - Q_1$$

1.1 Диагностика выбросов

Правило 3σ (если $X \sim \mathcal{N}(\mu, \sigma)$, то $P\{X \notin (\mu - 3\sigma; \mu + 3\sigma)\} = 0.0027$.):

$$[\bar{x} - 3 \cdot S \leq X \leq \bar{x} + 3 \cdot S]$$

Правило Чебышёва с параметрами (правило $k\sigma$):

$$[\bar{x} - k \cdot S \leq X \leq \bar{x} + k \cdot S]$$

Правило $k \cdot IQR$:

$$[Q_1 - k \cdot IQR \leq X \leq Q_3 + k \cdot IQR]$$

Если наблюдение находится вне рамок указанных интервалов, то оно считается выбросом по соответствующему правилу.

1.2 Визуализация выбросов и распределения

1.2.1 Ящичковая диаграмма

Линия на ящике - медиана;

Границы ящика - Q_1 и Q_3 ;

Границы усов через правило $k \cdot IQR$;

Точки за пределами усов - выбросы

[Пример графика](#)

1.2.2 Stemplot/Листовая диаграмма

Корни дерева - самый левый разряд числа, количество листьев эквивалентно количеству наблюдений в стебле/разряде. Если числа приведены, например, от 1 до 25, до у первого стебля корнем вершины будет 0, у второго - 1, у третьего - 2.

[Пример построения диаграммы](#)

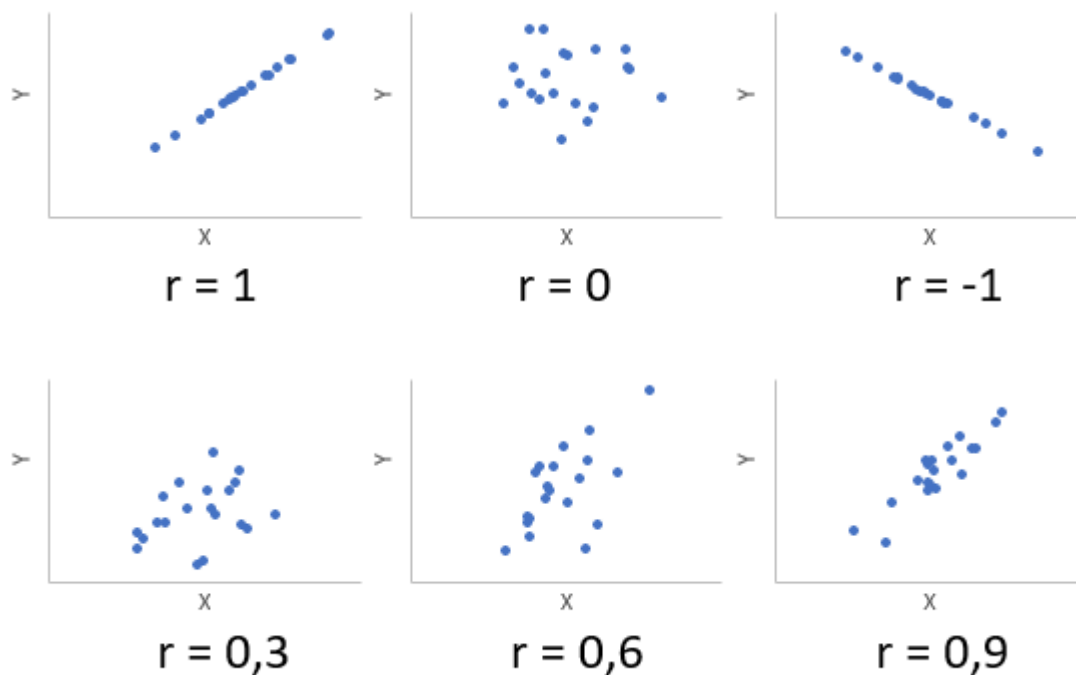
[Подробнее о диаграмме](#)

1.2.3 Dotplot/Точечная диаграмма

По оси X значения переменной, по оси Y - их частота встречаемости. Если повторений нет, то, возможно, надо поделить на какое-то число, чтобы наблюдения "сгруппировались".

[Пример построения диаграммы](#)

2 Корреляционный анализ



2.1 Парные коэффициенты корреляции

$$r_{xy} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{S_x \cdot S_y}$$

Проверка значимости парного коэффициента корреляции (через критерий Стьюдента):

$$t = \frac{r \cdot \sqrt{n - l - 2}}{\sqrt{1 - r^2}}$$

r - оценка парного коэффициента, l - число фиксируемых переменных, n - количество наблюдений. Критическое значение находится через таблицу распределения Стьюдента ([Ссылка на все таблицы от Миронкиной Ю.Н.](#)). Если $|t| > t$, то парный коэффициент корреляции значимо отличается от нуля с вероятностью ошибки первого рода α .

Проверка значимости парного коэффициента корреляции (через критерий Фишера-Йейтса) осуществляется через сравнение найденного наблюдаемого значения коэффициента корреляции и критического значения, которое можно найти через [таблицу Фишера-Йейтса](#)

Интервальные оценки парного коэффициента корреляции:

1. Переход к статистике Фишера:

$$Z' = \frac{1}{2} \cdot \ln \frac{1+r}{1-r}$$

2. Находим доверительный интервал для статистики Фишера по всем знакомой и известной формуле из математической статистики:

$$P \left(Z' - t_\gamma \sqrt{\frac{1}{n-l-3}} \leq Z \leq Z' + t_\gamma \sqrt{\frac{1}{n-l-3}} \right) = \gamma$$

3. Осуществляем обратное преобразование Фишера [по таблице](#) и получаем ДИ для нашего коэффициента корреляции (можно воспользоваться функцией "ФИШЕРОБР" в Excel). Обращаем ваше внимание на то, что используемая нами статистика Фишера является нечётной функцией, то есть $Z'(-r) = -Z'(r)$.

2.2 Частные коэффициенты корреляции

$$\rho_{xy/z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} = -\frac{A_{12}}{\sqrt{A_{11} \cdot A_{22}}}$$

Построение доверительных интервалов и проверка значимости частных коэффициентов корреляции производится через те же формулы, которые были приведены для парных коэффициентов корреляции, с одним уточнением: для парных $l=0$, для частных l есть число фиксируемых переменных, например, в трехмерной модели $l = 1$, в k -мерной $l = k - 2$.

2.3 Множественные коэффициенты корреляции

$$r_{1/2,3} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2 \cdot r_{12} \cdot r_{13} \cdot r_{23}}{1 - r_{23}^2}} = \sqrt{1 - \frac{|R|}{A_{11}}}$$

$|R|$ - определитель матрицы парных коэффициентов корреляции, A_{ij} - алгебраическое дополнение элемента r_{ij} корреляционной матрицы R .

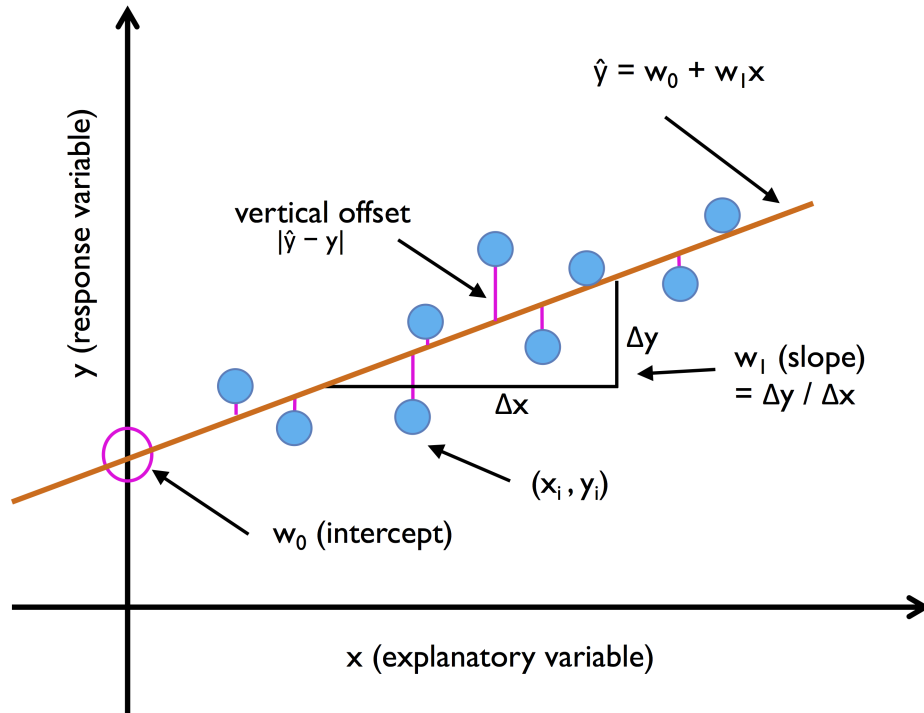
Множественный коэффициент корреляции, возведённый в квадрат, называется множественным коэффициентом детерминации и играет крайне важную роль в регрессионном анализе: он отражает, какая доля дисперсии целевой переменной может быть объяснена влиянием фиксируемых переменных (2, 3 в примере выше). То есть это своего рода верхняя граница качества регрессионной модели, показывающая также качество отобранных для анализа переменных.

Проверка значимости множественного коэффициента корреляции через F-критерий Фишера-Снедекора:

$$F_{\text{набл}} = \frac{\frac{1}{k-1} r_{1/2, \dots, k}^2}{\frac{1}{n-k} (1 - r_{1/2, \dots, k}^2)}$$

В рамках задач курса на ручной счёт вам предстоит работать лишь с трёхмерными моделями, где $k = 3$.

3 Регрессионный анализ



Рассмотрим частный случай модели двумерной регрессии (для задач на ручной счёт этого будет достаточно):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

где y_i, x_i — значение зависимой и независимой переменной для наблюдения i (наблюдаемые величины), β_0, β_1 — коэффициенты уравнения регрессии, $\varepsilon_i \sim IID(0; \sigma^2)$ — случайная ошибка для i -го наблюдения (ненаблюдаемые величины). Параметры модели для оценивания: $\beta_0, \beta_1, \sigma^2$.

Оценки регрессионных коэффициентов b_0, b_1 могут быть получены методом наименьших квадратов (МНК):

$$L = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \xrightarrow{b_0, b_1} \min$$

и при выполнении условий Гаусса-Маркова оценки МНК являются эффективными в классе линейных несмещенных оценок.

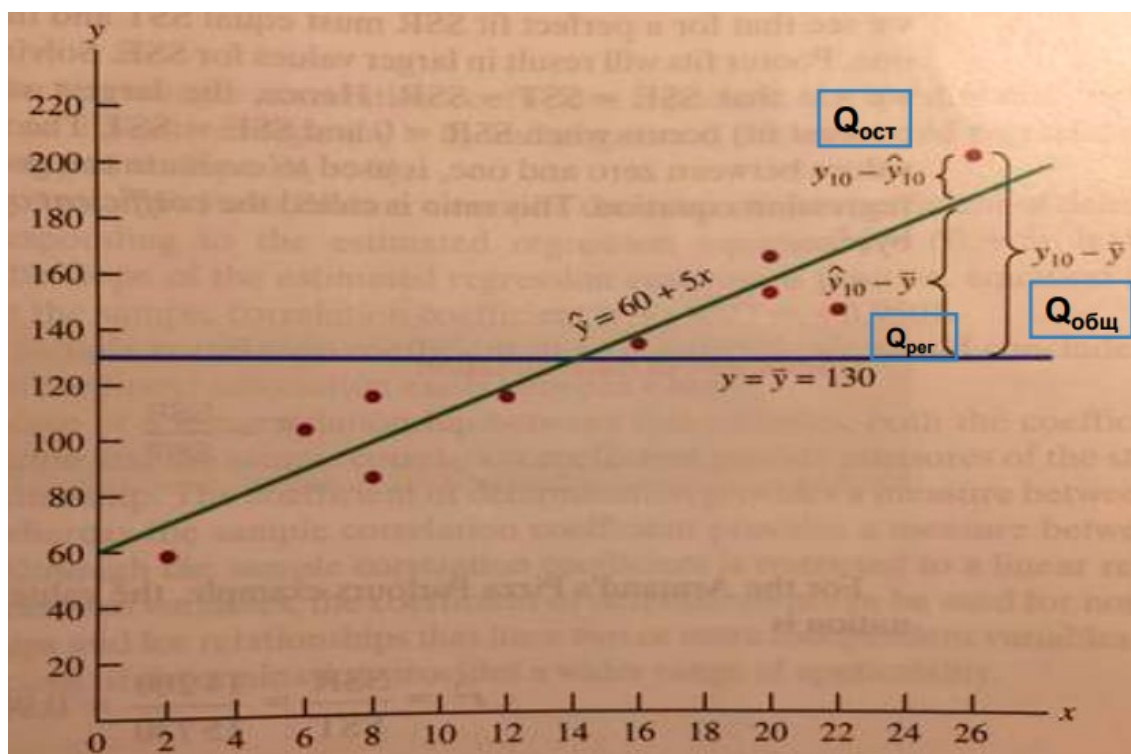
В результате решения оптимизационной задачи, мы получаем следующие формулы для нахождения оценок коэффициентов регрессии b_0, b_1 :

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

$$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{s_x^2}$$

s_x^2 здесь - это дисперсия объясняющей переменной. Смысловая нагрузка двух найденных коэффициентов такова: b_0 - точка пересечения линии регрессии с осью целевой переменной, необъясняемая признаками константа; b_1 - наклон линии регрессии, характеризует силу и направление влияния объясняющей переменной на целевую.

3.1 Отклонение выходов модели от реальных значений наблюдений



Квадраты отклонений целевой переменной (отсылка к МС: дисперсия, умноженная на количество наблюдений) в регрессионном анализе обозначаются как $Q_{\text{общ}}$ или TSS, разбиваясь на две величины: Q_R (воздействие

объясняющей переменной, также обозначается RSS) и $Q_{\text{ост}}$ (необъяснённые в модели различия целевой переменной, также обозначается ESS). Формально:

$$TSS = RSS + ESS$$

$$TSS = Q_{\text{общ}} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$RSS = Q_{\text{рег}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$ESS = Q_{\text{ост}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

При этом следует заметить, что отношение $\frac{RSS}{TSS}$ представляет из себя коэффициент детерминации (в случае многомерной (хотя бы трёхмерной) регрессии уместно вспомнить про квадрат множественного коэффициента корреляции и проверить идентичность с коэффициентом детерминации):

$$r^2 = \frac{RSS}{TSS} = \frac{TSS - ESS}{TSS}$$

3.2 MSE и дисперсия коэффициентов

Остаточное среднеквадратическое отклонение:

$$s = \sqrt{MSE} = \sqrt{\frac{ESS}{n-2}}$$

По аналогии с курсом математической статистики, мы можем оценить разброс найденных коэффициентов регрессии, посчитав их дисперсию:

$$s_{b_1} = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{s}{\sqrt{(n-1) \cdot s_x^2}}$$

$$s_{b_0}^2 = s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$$

3.3 Проверка гипотезы о значимости коэффициентов

Найденные СКО коэффициентов очень важны для проверки значимости регрессионной модели (значимо ли влияют объясняющие переменные на целевую?). Проверка гипотезы о незначимости коэффициента регрессии:

$$t_{\text{набл}} = \frac{b_1}{s_{b_1}}$$
$$t_{\text{кр}}(\alpha, \nu = n - 2)$$

Если наблюдаемое значение больше критического, то гипотеза о незначимости отклоняется, то есть с вероятностью ошибки первого рода α существует статистически значимая линейная связь между объясняющей и целевой переменными.

Если у нас более одной объясняющей переменной, то необходимо применить F-test для проверки гипотезы:

$$F_{\text{набл}} = \frac{MSR}{MSE} = \frac{\frac{RSS}{k}}{\frac{ESS}{n-2}} = \frac{RSS \cdot (n-2)}{ESS \cdot k}$$

$$F_{\text{кр}}(\alpha, \nu_1 = 1, \nu_2 = n - 2)$$

где $MSR = \frac{RSS}{k}$ - среднеквадратичное отклонение, объясняемой в рамках регрессионной модели, k - количество объясняющих переменных.

3.4 Доверительные интервалы

Также значимость коэффициентов регрессии можно вычислять через построение доверительных интервалов:

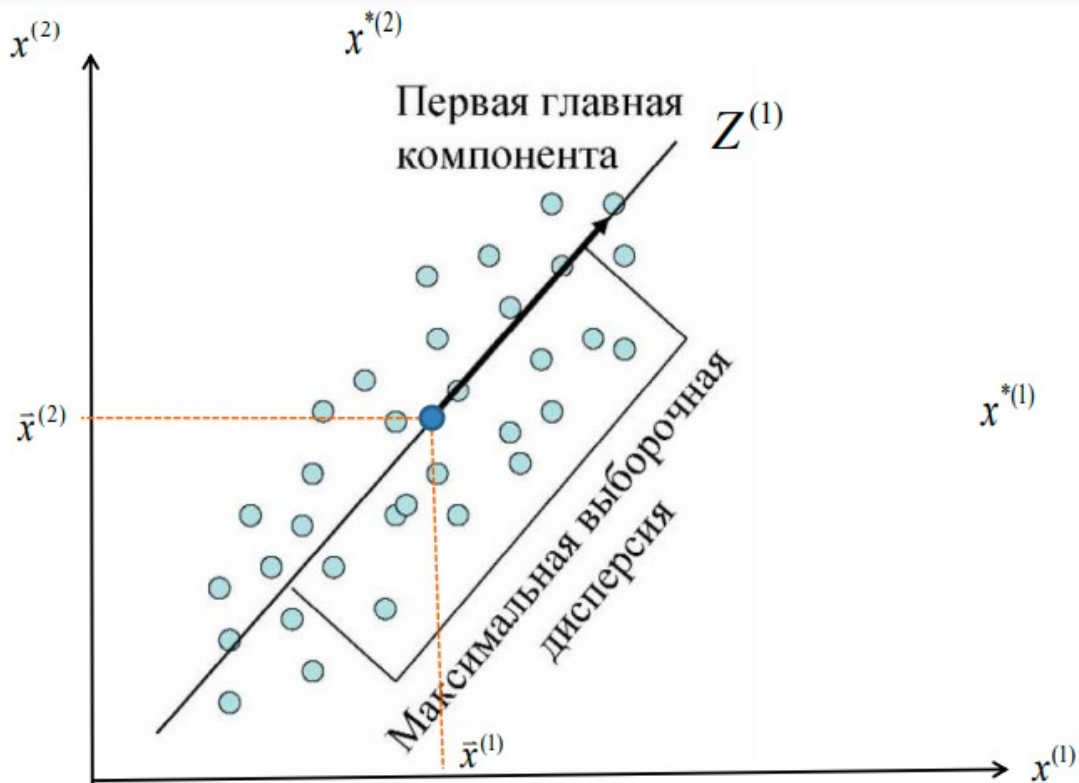
$$P(b_1 - t_{\alpha} s_{b_1} \leq \beta_1 \leq b_1 + t_{\alpha} s_{b_1}) = \gamma$$

$$P(b_0 - t_{\alpha} s_{b_0} \leq \beta_0 \leq b_0 + t_{\alpha} s_{b_0}) = \gamma$$

t_{α} - из таблицы распределения Стьюдента для заданного α и $\nu = n - 2$

При выполнении компьютерной работы крайне важным условием для проведения регрессионного анализа через МНК является проверка модели на гомоскедастичность - мы должны удостовериться в постоянстве(одинаковости) дисперсии рассматриваемых признаков!

4 Метод главных компонент



Этапы проведения МГК:

(А) Стандартизация переменных:

$$X^* = \frac{X - \bar{X}}{S}$$

\bar{X} - вектор средних для каждой объясняющей переменной

S - вектор СКО для каждой объясняющей переменной

(В) Расчет корреляционной матрицы:

$R = \frac{1}{n} X^{*T} X^*$, матрица R должна обладать размерностью $k \times k$, где k - количество объясняющих переменных.

(С) Поиск собственных значений корреляционной матрицы (рассмотрен случай для 2 объясняющих переменных):

$$\det \begin{pmatrix} 1 - \lambda & r_{12} \\ r_{12} & 1 - \lambda \end{pmatrix} = (1 - \lambda)^2 - r_{12}^2 = 0$$

Легко вывести, что для случая двух объясняющих переменных мы имеем следующие собственные значения: $\lambda_1 = 1 + r_{12}$, $\lambda_2 = 1 - r_{12}$

Поиск собственных векторов:

$$\lambda_1 = 1 + r_{12}$$

$$\begin{pmatrix} 1 - 1 - r_{12} & r_{12} \\ r_{12} & 1 - 1 - r_{12} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0, \quad l'_1 = \begin{pmatrix} ? \\ ? \end{pmatrix}$$

$$\lambda_2 = 1 - r_{12}$$

$$\begin{pmatrix} 1 - 1 + r_{12} & r_{12} \\ r_{12} & 1 - 1 + r_{12} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0, \quad l'_2 = \begin{pmatrix} ? \\ ? \end{pmatrix}$$

(D) Построение ортогональной матрицы собственных векторов:

$$L = (l_1 \ l_2)$$

$$l_i = \frac{l'_i}{\sqrt{(l'_i)^T l'_i}}$$

(E) Построение матрицы факторных нагрузок:

$$A = L\Lambda^{1/2} = (A_1 \ A_2)$$

(F) Построение матрицы нормированных значений главных компонент:

$$Z = X^*(A^T)^{-1}$$

Напомним, что обратная матрица для размерности 2×2 предполагает смену мест значений на главной диагонали и изменение знаков на побочной, при этом каждое из значений необходимо поделить на определитель исходной матрицы.

(G) Поиск коэффициентов регрессии, построенной на главные компоненты:

$$\hat{\beta} = (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T Y$$

\tilde{Z} отличается от найденной пунктом ранее матрицы Z лишь добавлением слева столбца из единиц, обозначающих константу.

Удельный вклад 1-й компоненты: $\frac{\lambda_1}{\sum \lambda_i} = \frac{1+r_{12}}{2}$, удельный вклад 2-й компоненты: $\frac{\lambda_2}{\sum \lambda_i} = \frac{1-r_{12}}{2}$.

5 Кластерный анализ

5.1 Методы вычисления расстояния между наблюдениями

Евклидово расстояние:

$$d_E(x_i, x_j) = \sqrt{\sum_{l=1}^k (x_{il} - x_{jl})^2}$$

Взвешенное Евклидово расстояние:

$$d_{WE}(x_i, x_j) = \sqrt{\sum_{l=1}^k w_l (x_i^{(l)} - x_j^{(l)})^2}$$

Где $\sum_{l=1}^k w_l = 1$ k - количество признаков, в пространстве которых находятся наблюдения

Расстояние по метрике городских кварталов (Манхэттенское, Хеммингово):

$$d_T(x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| = \sum_{l=1}^k |x_i^{(l)} - x_j^{(l)}|$$

Расстояние Махаланобиса:

$$d_M = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}$$

Расстояние Минковского:

$$d_T(x_i, x_j) = \left(\sum_{l=1}^k |x_i^{(l)} - x_j^{(l)}|^p \right)^{1/p}$$

Большее количество методов вычисления расстояния между объектами можно посмотреть в [Энциклопедическом словаре расстояний](#).

5.2 Методы вычисления расстояния между группами наблюдений

- “ближнего соседа”;
 $d_{min}(S_l, S_m) = \min_{x_i \in S_l, x_j \in S_m} d(x_i, x_j)$

- “дальнего соседа”;

$$d_{max}(S_l, S_m) = \max_{x_i \in S_l, x_j \in S_m} d(x_i, x_j)$$
- средней связи;

$$d_{average}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{x_i \in S_l, x_j \in S_m} d(x_i, x_j)$$
- центра тяжести.

$$d_{center}(S_l, S_m) = d(\bar{x}_l, \bar{x}_m)$$

5.3 Функционалы качества

- Сумма внутриклассовых дисперсий:

$$Q_1(S) = \sum_{l=1}^p \sum_{O_i \in S_l} d^2(O_i, \bar{X}(l)) \rightarrow \min,$$

p – число классов;

S_l – l -ый класс в классификации S ;

$\bar{X}(l)$ – центр класса S_l .

- Сумма попарных внутриклассовых расстояний между объектами:

$$Q_2(S) = \sum_{l=1}^p \sum_{O_i, O_j \in S_l} d^2(O_i, O_j) \rightarrow \min$$

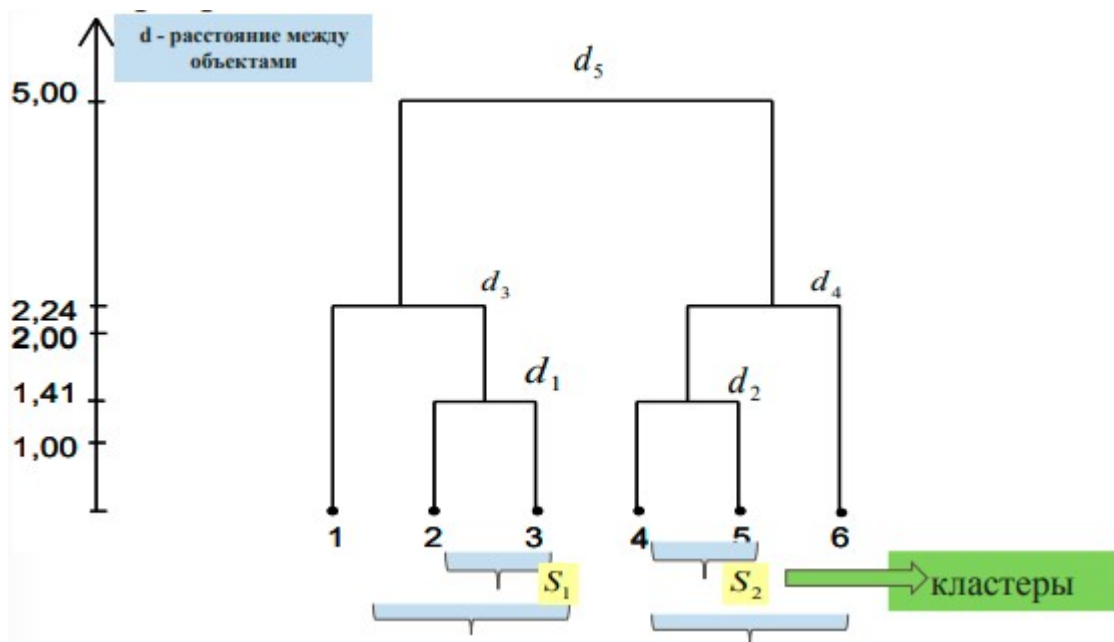
- Обобщённая внутриклассовая дисперсия:

$$Q_3(S) = \sum_{l=1}^p \sum_{j=1 \in k} S_j^2(l) \rightarrow \min,$$

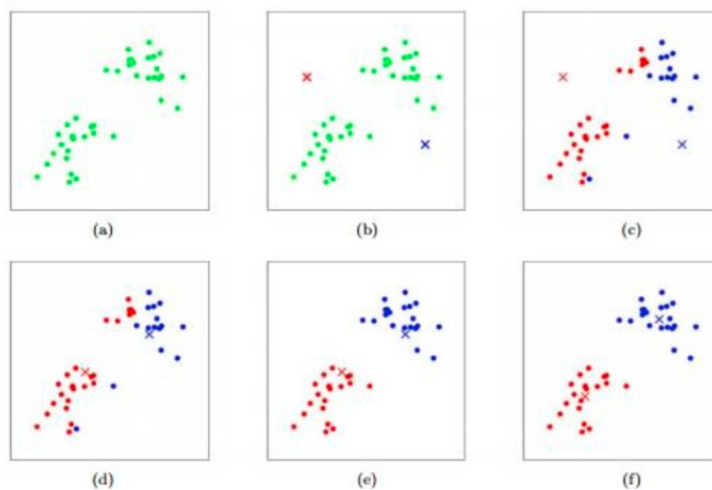
$S_j^2(l)$ – оценка дисперсии j -ого признака l -ого класса.

5.4 Иерархические методы классификации (дендрограммы)

При необходимости кластеризации относительно небольшого числа наблюдений будет уместно представить визуализацию пошагового объединения наблюдений в кластеры через дендрограмму. По оси ОУ на ней представлено расстояние между сгруппированными объектами, по оси ОХ – номера объединяемых наблюдений:



5.5 Метод классификации через k-средних



При необходимости кластеризации большого числа наблюдений следует заранее определить оптимальное число кластеров (в машинном обучении число кластеров называется гиперпараметром) и вычислять принадлежность наблюдений к тому или иному классу через итеративные алгоритмы, наиболее популярным из которых является алгоритм k-средних (k-means):

Как применяется метод?

- (A) Выбрать случайно заданные координаты центров кластеров (c_1, c_2, \dots, c_K) для того количества кластеров, которое мы избрали оптимальным
- (B) Отнести каждый объект к ближайшему из случайно на предыдущем шаге заданных центров:

$$y_i = \arg \min d(x_i, c_j)$$

- (C) Найти центр тяжести для каждой совокупности наблюдений, объединённых на данной итерации в один кластер, и переместить в этот центр тяжести центр каждого из кластеров (для каждого кластера разный центр тяжести!):

$$c_j = \frac{\sum_{i=1}^n x_i \cdot I[y_i = j]}{\sum_{i=1}^n I[y_i = j]}$$

- (D) Повторять два предыдущих шага до момента, пока на новой итерации наблюдения не перестанут переходить из одного кластера в другой

- 6 Дискриминантный анализ
- 7 Решающие деревья
 - 7.1 Задача классификации
 - 7.2 Задача регрессии
- 8 Секретные темы, название которых вы узнаете позже...