Статистическая кластеризация

Алексей Дзюба / ПИЯФ НИЦ КИ

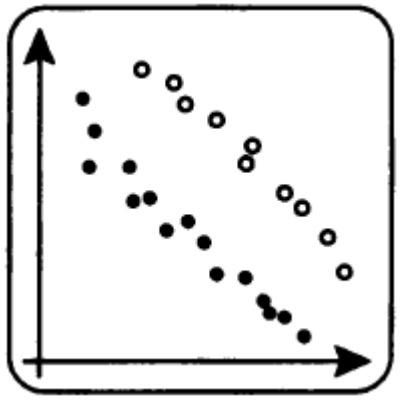
(учебник М.Б. Лагутин «Наглядная математическая статистика», БИНОМ, М., 2009, Глава 19)

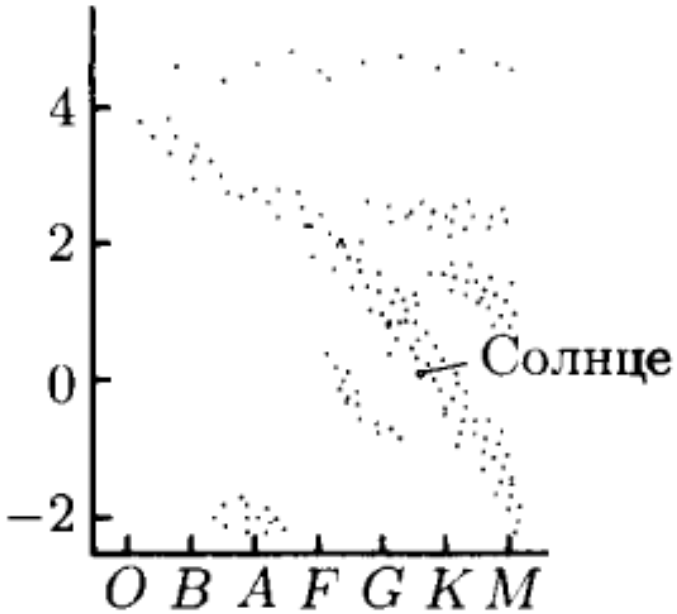
Вводные замечания

- В практической части курса мы будем рассматривать алгоритмы машинного обучения для **задачи классификации**
  - Есть размеченная на классы выборка (кортеж по нескольким параметрам), на основе которой нужно построить алгоритм для определения класса объекта по параметрам

- В этой презентации мы рассмотрим **задачу кластеризации**, то есть разбиение множества объектов на компактные группы

Эффект существенной многомерности



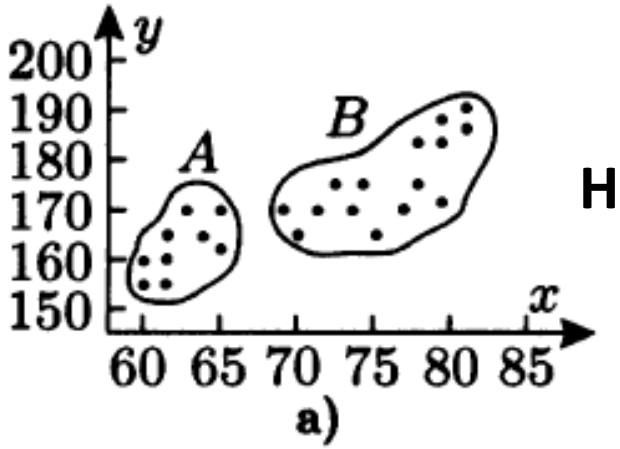


Так, точки и кружки на рис. 1 почти не отличаются друг от друга по каждой из координат в отдельности, но очевидным образом разделяются по новому признаку — сумме координат.

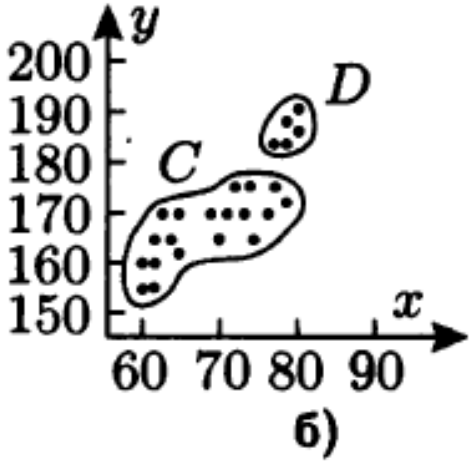
«С давних пор астрономы знали о различной светимости звезд, т. е. о различной их «истинной яркости». В конце XIX в. были открыты также различные спектральные классы звезд, попросту говоря — различный цвет их излучения (от красного до голубого). До 1913 г. эти характеристики существовали в представлении ученых раздельно, но вот (независимо друг от друга) датский астроном Герцшпрунг и американец Ресселл сопоставили их между собой и построили двумерную проекцию объектов-звезд на плоскость признаков спектр — светимость. Результаты оказались неожиданными

Астрономы увидели, что звезды не распределены в пространстве этих признаков равномерно, а образуют несколько ярко выраженных кластеров, причем стало возможным предсказать эволюцию звезд по значениям их основных характеристик. С тех пор диаграмма Герцшпрунга—Ресселла стала одним из важных инструментов в работе современных астрономов.»

Зависимость от выбора масштаба шкалы



а)



б)

Пример:

A – девушки, B – юноши

D – высокие юноши, C – все остальные

Нормировка расстояний (для одномерных наблюдений):

N1) $Z_i' = (Z_i - Z_{min})/(Z_{max} - Z_{min})$.

N2) $Z_i' = (Z_i - \overline{Z})/S$, где $\overline{Z} = \frac{1}{n}\sum Z_i$ — среднее арифметическое,

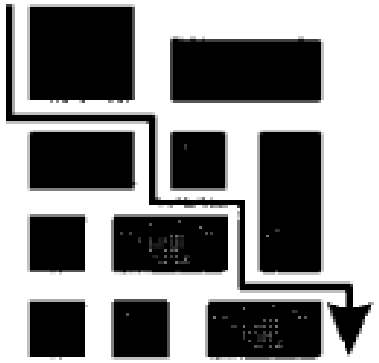
$S^2 = \frac{1}{n}\sum (Z_i - \overline{Z})^2$ — выборочная дисперсия.

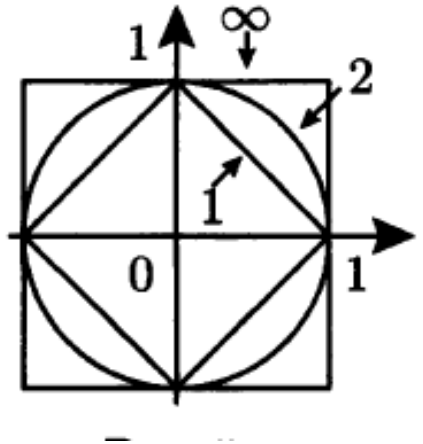
N3) $Z_i' = (Z_i - MED)/MAD$, где MED — выборочная медиана (см. § 2 гл. 7), MAD[**)] — *(нормированная) медиана абсолютных отклонений* от MED:

$$MAD = \frac{1}{\Phi^{-1}(3/4)}\, MED\{|Z_i - MED|,\ i = 1, \ldots, n\},$$

где $\Phi^{-1}(x)$ — функция, обратная к функции распределения закона $\mathcal{N}(0,1)$.[***)] Такое преобразование менее подвержено влиянию выделяющихся значений Z_i.

Метрика (выбор расстояния между объектами)





Единичные шары для различных метрик

D1) *Метрика города* (рис. 4)*): $d_{ij} = \sum_{l=1}^{m} |x_{il} - x_{jl}|$. При использовании метрики города хорошо выделяются классы, имеющие вид «облака», вытянутого вдоль оси некоторого признака.
В случае, когда координаты объектов принимают только значения 0 и 1, это расстояние равно количеству несовпадающих координат, т. е. длине пути по ребрам единичного m-мерного куба из одной вершины в другую (*метрика Хемминга*).

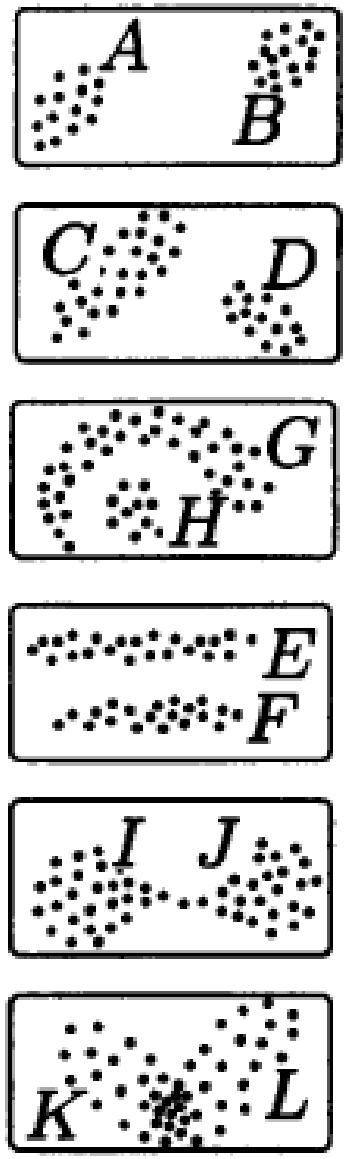
D2) *Евклидова метрика*: $d_{ij} = \left(\sum_{l=1}^{m} (x_{il} - x_{jl})^2 \right)^{1/2}$.

D3) *Метрика Чебышёва*: $d_{ij} = \max_{1 \leqslant l \leqslant m} |x_{il} - x_{jl}|$.

Все три расстояния являются частными случаями (соответственно при $p = 1$, 2 и ∞) так называемого *расстояния Минковского*

$$d_{ij} = \left(\sum_{l=1}^{m} |x_{il} - x_{jl}|^p \right)^{1/p}.$$

Типы классов



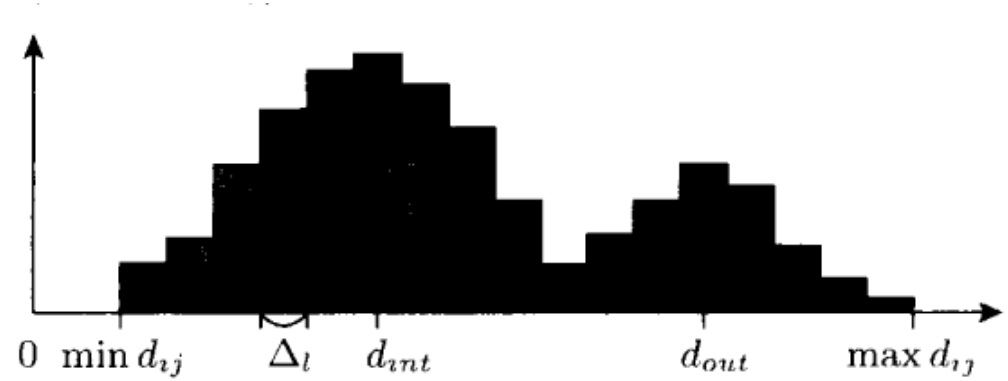
C1) Класс типа ядра [60] (в [56, с. 235] такой класс называется *сгущением*). Все расстояния между объектами внутри класса меньше любого из расстояний между объектами класса и остальной частью множества объектов. На рис. 6 сгущениями являются A и B. Остальные пары множеств не разделяются с помощью этого определения.

C2) Кластер (*сгущение в среднем* [56]). Среднее расстояние внутри класса меньше среднего расстояния объектов класса до всех остальных. Множества C и D теперь разделяются, но у E (G) среднее внутреннее расстояние больше, чем среднее расстояние между E и F (G и H).

C3) Класс типа ленты [60] (*слабое сгущение* [56]). Существует $\tau > 0$ такое, что для любого $\boldsymbol{x}_i$ из класса S найдется такой объект $\boldsymbol{x}_j \in S$, что $d_{ij} \leqslant \tau$, а для всех $\boldsymbol{x}_k \notin S$ справедливо неравенство $d_{ik} > \tau$. В смысле этого определения на рис. 6 разделяются все пары множеств кроме I и J, K и L.

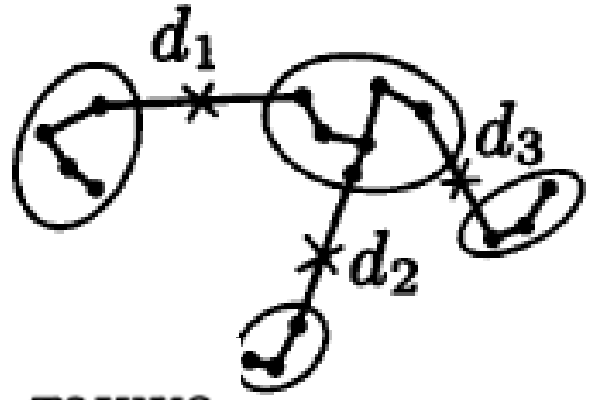
C4) Класс с центром. Существует порог $R > 0$ и некоторая точка $\boldsymbol{x}^*$ в пространстве, занимаемом объектами класса S (в частности, элемент этого множества) такие, что все объекты из S и только они содержатся в шаре радиуса R с центром в $\boldsymbol{x}^*$. Часто в качестве $\boldsymbol{x}^*$ выступает центр масс класса S, т. е. координаты центра определяются как средние значения признаков у объектов класса. Множества I и J являются классами с центром, а E, F и G — нет.

Эвристические методы



А1) Связные компоненты. Все объекты разбиваются на классы *типа ленты, или слабого сгущения* (тип С3 в § 1), где задаваемый параметр $\tau \in (\min d_{ij}, \max d_{ij})$. В этой постановке задача классификации эквивалентна нахождению связных компонент графа (вершины графа i и j соединены ребром, если $d_{ij} \leqslant \tau$). (Алгоритм выделения связных компонент методом *поиска в глубину* излагается в § 9.) Для выбора величины τ полезно построить *гистограмму межобъектных расстояний* (высота прямоугольника над промежутком Δ_l на рис. 7 пропорциональна количеству d_{ij} в Δ_l). При хорошей структурированности данных гистограмма, как правило, имеет два выделяющихся максимума: при $d_{ij} \approx d_{int}$ (*типичное внутриклассовое расстояние*) и при $d_{ij} \approx d_{out}$ (*типичное межклассовое расстояние*). Часто удачным выбором τ оказывается значение $(d_{int} + d_{out})/2$.

Эвристические методы



A2) Кратчайший незамкнутый путь (КНП). Его также называют *минимальным покрывающим деревом* или *каркасом*. Соединяются ребром две ближайшие точки, затем среди оставшихся отыскивается точка, ближайшая к любой из уже соединенных точек, и присоединяется к ним и т. д. до исчерпания всех точек. Р. Прим в 1957 г. доказал, что построенный таким способом граф имеет минимальную общую длину ребер среди всевозможных соединений, связывающих все вершины (см. [28, с. 60]).

В найденном КНП затем отбрасывают $k-1$ самых длинных дуг и получают k классов (рис. 8).*) Метод позволяет выделять классы произвольной формы.

Эвристиче
методы

А3) Метод k-средних[*)] предназначен для выделения классов типа С4 («класс с центром»). Приведем **два варианта**:

(a) *Алгоритм Г. Болла и Д. Холла* (1965 г., см. [52, с. 110]). Случайно выбираются k объектов (*эталонов*); каждый объект присоединяется к ближайшему эталону (тем самым образуются k классов); в качестве новых эталонов принимаются центры масс классов.[**)] После пересчета объекты снова распределяются по ближайшим эталонам и т. д. Критерием окончания алгоритма служит стабилизация центров масс всех классов.

Вместо случайно выбираемых эталонов лучше использовать k наиболее удаленных объектов: сначала отыскиваются два самых удаленных друг от друга объекта, затем l-й эталон ($l = 3, \ldots, k$) определяется как наиболее удаленный в среднем от уже имеющихся.

(b) *Алгоритм Дж. Мак-Кина* (1967 г., см. [52, с. 98]). Он отличается от метода Болла и Холла тем, что при просмотре списка объектов пересчет центра масс класса происходит после присоединения к нему каждого очередного объекта.

Эвристические методы

А4) Алгоритм «Форель». Случайный объект объявляется центром класса; все объекты, находящиеся от него на расстоянии не большем R, входят в первый класс. В нем определяется центр масс, который объявляется новым центром класса и т. д. до стабилизации центра. Затем все объекты, попавшие в первый класс изымаются, и процедура повторяется с новым случайным центром.

Можно скомбинировать алгоритмы А4 и А2 ([52, с. 67)]: при небольшом R по алгоритму А4 находят $k' > k$ классов; их центры соединяют КНП, из которого удаляют $k - 1$ самых длинных ребер и получают k классов. При этом образуются классы более сложной формы, чем m-мерные шары (рис. 10). Здесь важна идея двухэтапности классификации: сначала выделить заведомо компактные маленькие группы, затем произвести их объединение. Так можно успешно классифицировать довольно большие массивы информации (сотни объектов).

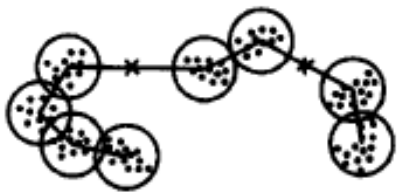
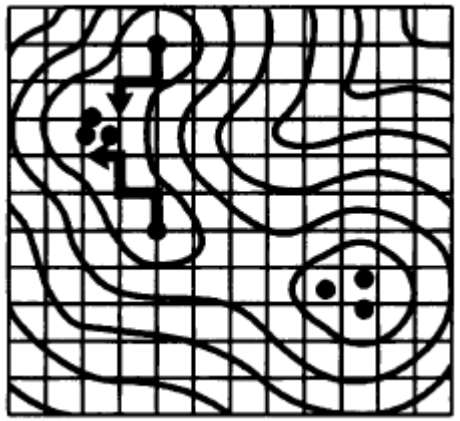


Рис. 10

Эвристические методы



A5) Метод потенциальных ям. Предположим, что каждый объект $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$ создает вокруг себя поле притяжения с некоторой весовой функцией, например, гладким *квартическим ядром*

$$W_i(\boldsymbol{x}) = \left[1 - (r_i/R)^2\right]^2 I_{\{r_i \leqslant R\}},$$

где $r_i = |\boldsymbol{x} - \boldsymbol{x}_i|$, а параметр $R > 0$ задает эффективный размер области притяжения. Все вместе объекты создают потенциальное поле $U(\boldsymbol{x}) = -\sum W_i(\boldsymbol{x})$. Классам соответствуют потенциальные ямы: объект $\boldsymbol{x}_i$ относится к яме, в которую он «скатывается» при свободном движении. Практически приходится, стартовав с $\boldsymbol{x}_i$, запускать некоторый алгоритм (локальной) минимизации.

Иерархические методы

Общая схема этих процедур такова: сначала каждый объект считается отдельным классом; на первом шаге объединяются два ближайших объекта, которые образуют новый класс (если сразу несколько объектов (классов) одинаково близки, то выбирается одна случайная пара); вычисляются *меры отдаленности* ρ (см. ниже)*) от этого класса до всех остальных классов, и размерность матрицы межклассовых мер отдаленности сокращается на единицу; шаги процедуры повторяются до тех пор, пока все объекты не объединятся в один класс.

Иерархические методы

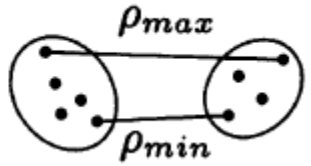
P1) Метод «ближнего соседа»: $\rho_{min} = \min\limits_{\boldsymbol{x}_i \in S_k,\, \boldsymbol{x}_j \in S_l} d_{ij}$,

P2) Метод «дальнего соседа»: $\rho_{max} = \max\limits_{\boldsymbol{x}_i \in S_k,\, \boldsymbol{x}_j \in S_l} d_{ij}$.

P3) Метод средней связи: $\rho_{ave} = \dfrac{1}{n_k n_l} \sum\limits_{\boldsymbol{x}_i \in S_k} \sum\limits_{\boldsymbol{x}_j \in S_l} d_{ij}$ (здесь n_k и n_l — количества объектов в классах S_k и S_l).

P4) Метод центров масс: $\rho_{center} = |\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}}_l|^2$, где $\overline{\boldsymbol{x}}_k$ и $\overline{\boldsymbol{x}}_l$ обозначают центры масс k-го и l-го классов.

P5) Метод Уорда*): $\rho_W = \dfrac{n_k n_l}{n_k + n_l} |\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{x}}_l|^2$.



Номер	Название	C_1	C_2	C_3	C_4
P1	Ближний сосед	$1/2$	$1/2$	0	$-1/2$
P2	Дальний сосед	$1/2$	$1/2$	0	$1/2$
P3	Средняя связь	$\frac{n_1}{n_1+n_2}$	$\frac{n_2}{n_1+n_2}$	0	0
P4	Центры масс	$\frac{n_1}{n_1+n_2}$	$\frac{n_2}{n_1+n_2}$	$-\frac{n_1 n_2}{(n_1+n_2)^2}$	0
P5	Метод Уорда	$\frac{n_0+n_1}{n_0+n_1+n_2}$	$\frac{n_0+n_2}{n_0+n_1+n_2}$	$-\frac{n_0}{n_0+n_1+n_2}$	0

удобно воспользоваться общей для методов P1—P5 **формулой Г. Ланса и У. Уильямса:**

$$\rho(S_0, S_1 \cup S_2) = C_1\rho_{01} + C_2\rho_{02} + C_3\rho_{12} + C_4|\rho_{01} - \rho_{02}|,$$

Сравнение методов по данным [52]

52. *Мандель И. Д.* Кластерный анализ. — М.: Финансы и статистика, 1988

Изложим кратко **основные выводы**. Наилучшей (почти идеальной) по восстанавливаемости разбиения проявила себя иерархическая процедура P5 («метод Уорда»). Следом за ней идут процедура P2 («дальнего соседа») и алгоритм A3 (метод k-средних Болла и Д. Холла) (случайный выбор эталонов в алгоритме A3 показал себя как крайне неудачный). Самой плóхой оказалась процедура P1 («ближнего соседа»).

По уровню устойчивости к шуму лидерами стали алгоритмы A3 и P5. Замыкает список снова P1.

В случае, когда классы имеют сложную форму, скажем, относятся к типу C3 из § 1 («класс типа ленты или слабое сгущение»), именно алгоритмы P1, A1 и A2 позволят правильно произвести разбиение.

ММП-оценка для многомерной нормальной модели

$$p(\boldsymbol{x},\boldsymbol{\theta}) = (2\pi)^{-m/2}\,(\det \boldsymbol{\Sigma})^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

Положим $\overline{\boldsymbol{x}} = (\boldsymbol{x}_1 + \ldots + \boldsymbol{x}_n)/n$.

Введем *выборочную ковариационную матрицу*

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T.$$

В силу независимости случайных векторов $\boldsymbol{\xi}_i$ логарифм правдоподобия (§ 4 гл. 9) имеет вид

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i,\boldsymbol{\theta}) =$$

$$= -\frac{1}{2}\left[n \ln \det \boldsymbol{\Sigma} + \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right] + \text{const}\,.$$

Минимизируя логарифм L, можно получить

Таким образом, оценками максимального правдоподобия параметров $\boldsymbol{\mu}$ и $\boldsymbol{\Sigma}$ являются, соответственно, статистики $\overline{\boldsymbol{x}}$ и $\widehat{\boldsymbol{\Sigma}}$.

Двумерный случай

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

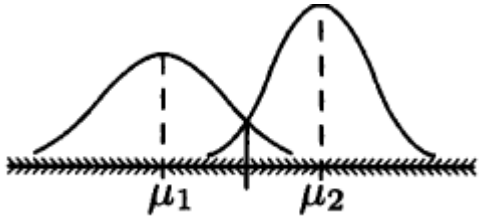
Разделение многомерных нормальных величин (дискриминантный анализ)

Для того, чтобы это сделать, воспользуемся статистической моделью случайного выбора единственного наблюдения из некоторого многомерного нормального закона, причем заранее известно, что этот закон является одним из k заданных законов $\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$, где $l = 1, \ldots, k$.

В соответствии с **принципом максимального правдоподобия** будем считать *областью притяжения* закона $\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ $(l = 1, \ldots, k)$ множество таких точек $\boldsymbol{x} \in \mathbb{R}^m$, где плотность распределения $\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ больше других. Это равносильно тому, что величина

$$h_l(\boldsymbol{x}) = \ln \det \boldsymbol{\Sigma}_l + (\boldsymbol{x} - \boldsymbol{\mu}_l)^T \boldsymbol{\Sigma}_l^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_l)$$

имеет *наименьшее значение* среди $h_1, \ldots, h_k$ (см. формулу (8)).



Одномерный случай для двух классов

В случае выбора между двумя законами метод минимизирует сумму ошибок первого и второго рода

Представление результатов

После проведения классификации важно в удобной форме представить ее результаты. Приведем список важнейших (согласно [52, с. 159]) **характеристик классификации**.

1. Распределение номеров объектов по номерам классов.
2. Гистограмма межобъектных расстояний (подобная изображенной на рис. 7).
3. Средние внутриклассовые расстояния.
4. Матрица средних межклассовых расстояний.
5. Визуальное представление данных на плоскости двух (в пространстве трех) «наиболее информативных» признаков.
6. Дендрограмма для иерархических процедур.
7. Средние значения и размахи во всех классах для каждого признака.