

Многомерный анализ данных: классификация и кластерный анализ

для магистров СПбГУ, ПМИ, статмод

Голяндиной Н.Э. (черновой вариант, просьба сообщать об опечатках)

Первоначальный набор теста — выпуск статмода 2017 (магистратура)

Собрано 22 ноября 2021 г. в 20:56

1 Классификация

1.1 Общий подход к классификации через апостериорные вероятности

Общая подход к классификации: строятся классифицирующие функции f_i , такие что классификация проводится так: индивид с признаками \mathbf{x} относится к группе с максимальным значением на нем классифицирующей функции: $\arg \max_i f_i(\mathbf{x})$.

Откуда берутся эти классифицирующие функции? Естественная идея взять в качестве f_i вероятность (ее оценку) принадлежности к i -му классу. Пусть ξ – дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\mathcal{P}(\eta \mid \xi = A_i) = \mathcal{P}_i$ и имеет плотность $p_i(\mathbf{x})$. Тогда было бы логично взять $f_i = p_i$. Для практического применения надо было бы оценить плотности, либо непараметрически (например, по числу точек, попавших в дельта-окрестность — типа метода ближайших соседей), либо параметрически (если известно, что распределение нормальное, тогда просто оцениваем векторы средних и ковариационные матрицы).

Более сложный подход — через апостериорные вероятности. Если у нас есть априорное знание вероятности того, что индивид из того или иного класса, то мы можем его учесть. Введем понятие класса $C_i = \{\xi = A_i\}$. Чтобы классифицировать наблюдение \mathbf{x} , необходимо найти

$$\arg \max P(\xi = A_i \mid \eta = \mathbf{x}) = \arg \max P(C_i \mid \mathbf{x}).$$

Пусть известны априорные вероятности принадлежности нового наблюдения к i -му классу $\pi_i = P(C_i)$. Тогда апостериорные вероятности по формуле Байеса будут иметь вид

$$P(C_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid C_i) \pi_i}{\sum_{j=1}^k P(\mathbf{x} \mid C_j) \pi_j}.$$

Поэтому в качестве классифицирующих функций берут

$$f_i(\mathbf{x}) = \frac{p_i(\mathbf{x}) \pi_i}{\sum_{j=1}^k p_j(\mathbf{x}) \pi_j}.$$

Так как знаменатель у всех f_i одинаковый, его можно отбросить, и итоговые классифицирующие функции будут выглядеть как $f_i(\mathbf{x}) = P(\mathbf{x} \mid C_i) \pi_i = p_i(\mathbf{x}) \pi_i$.

Как выбрать априорные вероятности?

1. Равномерно, $\forall i \in 1 : k \pi_i = 1 / k$.
2. По соотношениям в обучающей выборке: $\pi_i = n_i / \sum_{j=1}^k n_j$.
3. На основе другой дополнительной информации о данных (результаты предыдущих исследований, etc.)

Свойство. Построенный метод классификации $\text{predict}(\mathbf{x}) = \arg \max_i \pi_i p_i(\mathbf{x})$ минимизирует среднюю апостериорную ошибку:

$$\sum_{i=1}^k \pi_i P(\text{predict}(\mathbf{x}) \neq i \mid C_i).$$

Видно, что можно с помощью априорных вероятностей формально задавать важность ошибочных классификаций для разных классов.

1.2 Линейный и квадратичный дискриминантный анализ для классификации

1.2.1 LDA

Модель: ξ — дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\mathcal{P}(\eta \mid \xi = A_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Тогда плотность в точке \mathbf{x}

$$p_i(\mathbf{x}) = p(\mathbf{x} \mid \xi = A_i) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right),$$

и классифицирующая функция $f_i(\mathbf{x}) = \pi_i p(\mathbf{x} \mid \xi = A_i)$, где π_i — априорная вероятность наблюдения попасть в i -ю группу. Для упрощения вычислений можно переписать классифицирующую функцию через возрастающее монотонное преобразование как

$$g_i(\mathbf{x}) = \log f_i(\mathbf{x}) = \log \pi_i - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i).$$

Сократив часть, не зависящую от номера класса, получаем линейные классифицирующие функции

$$h_i(\mathbf{x}) = -\frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \log \pi_i.$$

Замечание. Если две группы, то гипотеза о равенстве многомерных мат.оэж. $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (различие значимо, если гипотеза отвергается, а только тогда имеет смысл проводить классификацию) в модели LDA проверяется с помощью критерия Хотеллинга (Hotelling). Если групп несколько, то есть разные критерии, например, критерия Wilk's Lambda или Roy's greatest root (эти же критерии используются в MANOVA, Multivariate ANalysis Of VAriance). Они отличаются мощностью против разного расположения групп.

Немного про канонические переменные В LDA есть так называемые канонические переменные. Идея похожа на АГК, только оптимизационная задача другая. Аналогично, на основе исходных признаков (признаки центрируются) строятся новые признаки как линейные комбинации исходных признаков. Только первая каноническая переменная — это такая линейная комбинация исходных признаков, по которой группа максимально отличаются (отличие измеряется на основе ANOVA, по статистике критерия Фишера). Вторая линейная комбинация должна быть ортогональна первой и приводит к максимальному различию среди ортогональных линейных комбинаций. И т.д. Удобно смотреть на данные в плоскости первой и второй канонических переменных (иногда это называют roots).

Приведем формулу, как находятся коэффициенты линейной комбинации (канонические коэффициенты) для получения канонических переменных. Неудивительно, что экстремальная задача приводит к поиску собственных векторов некоторой матрицы.

Имеет место разложение выборочной ковариационной матрицы, умноженной на n (индивид $\mathbf{y}_{ij} \in \mathbb{R}^p$ — j -й индивид из i -й группы):

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}})(\mathbf{y}_{ij} - \bar{\mathbf{y}})^T = \sum_{i=1}^k n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^T + \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T = \mathbf{H} + \mathbf{E}. \quad (1)$$

Должно быть ясно, что это лишь многомерное обобщение разложения выборочной дисперсии (умноженной на n).

Первое слагаемое отвечает за равенство средних (неотличимые группы), его назовем **H** от слов hypothesis, а второе — за отклонение данных в каждой группе от своего среднего, его назовем **E** от слова error.

В этих обозначениях, канонические коэффициенты являются собственными векторами матрицы $\mathbf{E}^{-1}\mathbf{H}$. А собственные числа λ_i (упорядоченные по убыванию вместе с собственными векторами) этой матрицы отражают то, насколько группы хорошо разделяются по соответствующей канонической переменной. Число ненулевых собственных чисел $s \leq \min(n, k - 1)$.

Критерии для проверки гипотезы о том, что группы не разделимы ($H_0 : \boldsymbol{\mu}_1 = \dots = \boldsymbol{\mu}_k$), являются комбинацией этих собственных чисел. Например, статистика критерия Wilks' Lambda

$$\Lambda = \prod_{i=1}^s \frac{1}{1 + \lambda_i}$$

(с какой стороны критическая область?). А статистика критерия Roy's greatest root имеет вид

$$r_1^2 = \frac{\lambda_1}{1 + \lambda_1}$$

(с какой стороны критическая область?).

Как понять, против какого расположения группы мощнее один, а против какого — другой?

1.2.2 QDA

Модель: ξ — дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\mathcal{P}(\eta \mid \xi = A_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Тогда плотность в точке \mathbf{x}

$$p(\mathbf{x} \mid \xi = A_i) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right),$$

и классифицирующая функция $f_i(\mathbf{x}) = \pi_i p(\mathbf{x} \mid \xi = A_i)$. Применяем возрастающее монотонное преобразование и оставляем в классифицирующей функции только члены, отличающиеся в разных группах:

$$g_i(\mathbf{x}) = \log f_i(\mathbf{x}) = \log \pi_i - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i),$$

получаем квадратично зависящую от \mathbf{x} классифицирующую функцию.

Замечание. Если две группы, то гипотеза о равенстве многомерных мат.ож. $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (различие значимо, если гипотеза отвергается, а только тогда имеет смысл проводить классификацию) в модели QDA тоже проверяется с помощью критерия Хотеллинга (Hotelling), но с отдельно оцененными ковариационными матрицами (критерий асимптотический). Если групп несколько, то тут уже критерий сложно построить.

1.3 Классификация в случае двух классов

Если всего два класса, то можно построить границу между классами, приравняв классифицирующие функции.

1.3.1 LDA

Приравняв $h_1(x) = h_2(x)$, получим разделяющую гиперплоскость. Разделяющая два класса гиперплоскость имеет вид

$$\begin{aligned}\{\mathbf{x} : h_1(\mathbf{x}) = h_2(\mathbf{x})\} = \\ = \{\mathbf{x} : -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \log(\pi_1/\pi_2) = 0\}.\end{aligned}$$

От соотношения между априорными вероятностями зависит положение границы относительно классов (к какому она ближе). Видно, что априорные вероятности влияют только на сдвиг разделяющей гиперплоскости.

Заметим, что классификацию можно записать как сравнение $-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$ с некоторым порогом ($-\log(\pi_1/\pi_2)$), который зависит от априорных вероятностей (или весов ошибок для разных классов, смотря как на это смотреть).

1.3.2 QDA

В данном случае, разделяющая поверхность имеет вид квадратичной поверхности, может состоять из двух гиперболоидом, может иметь форму эллипса.

1.3.3 Картинки

Здесь мы обсуждали число параметров в моделях, возможный overfitting (переподгонку). Использовали слова — обобщающая способность алгоритма.

1.4 Качество классификации

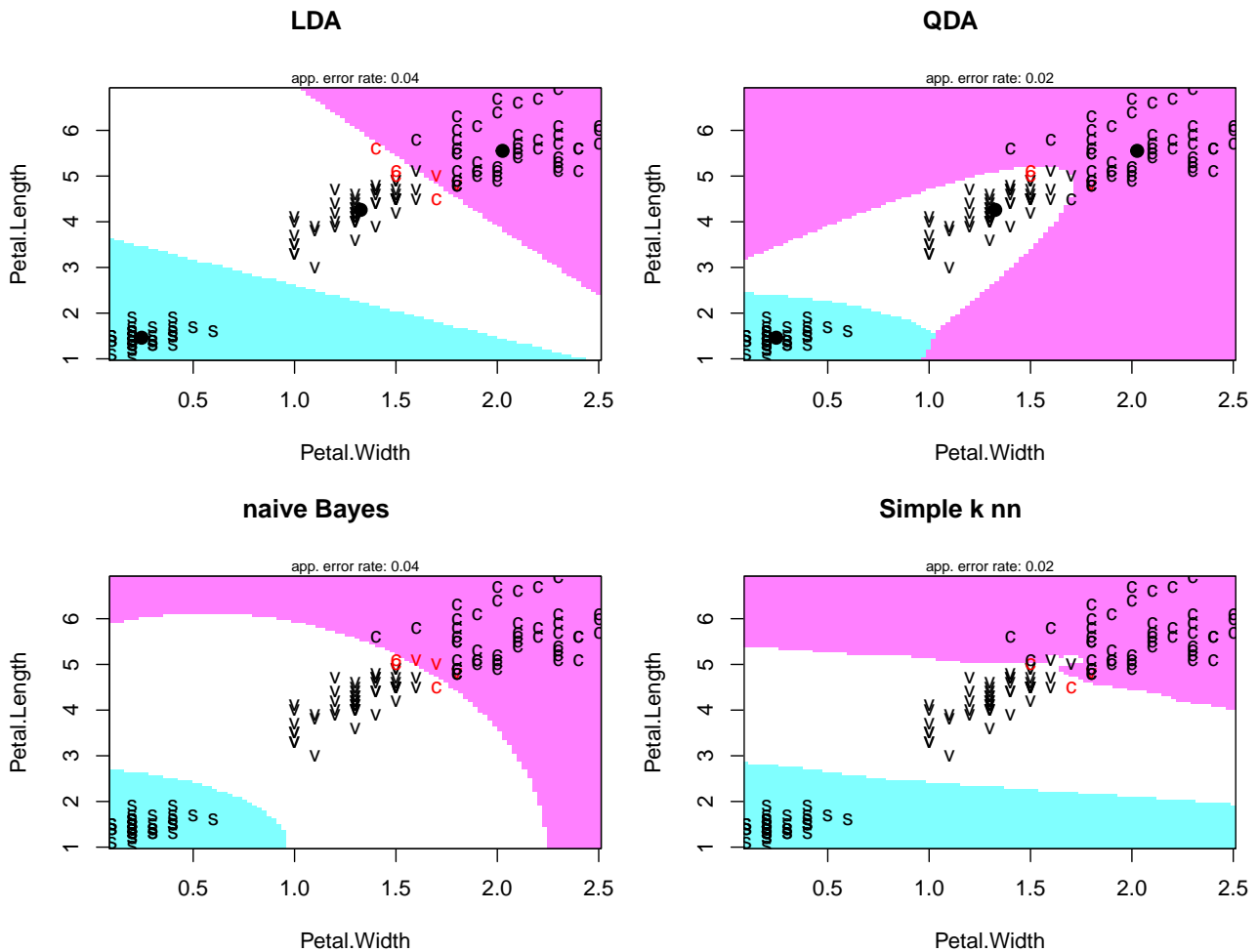
1.4.1 Ошибки классификации

Качество классификации измеряется ошибками классификации (доля неправильно классифицированных объектов). n_{ij} — число объектов из класса i , отнесенных к классу j . В соответствующей матрице классификации на диагонали стоят правильно классифицированные объекты, вне диагонали — ошибки.

На самом деле, нельзя проверять качество предсказания на тех данных, на которых это предсказание строилось. Поэтому используют кросс-валидацию (скользящий контроль). Например, каждое наблюдение по очереди исключается из выборки, классифицирующее правило строится без него и с помощью этого правила индивид классифицируется. Строится аналогичная таблица из n_{ij} . В ней ошибок будет, вообще говоря, больше.

Здесь обсуждали, что имеет смысл смотреть на ошибки без кросс-валидации и с ней. Если разница существенная, то это говорит о переподргонке используемой модели. Вероятно, она не очень хорошая; например, слишком много параметров.

Замечание. Нельзя путать классификацию и различие групп. Группы могут значительно различаться, классификация может быть при этом бессмысленной (ошибок чуть меньше 50%).



1.4.2 ROC и AUC

wikipedia ROC-кривая (англ. receiver operating characteristic, рабочая характеристика приёмника) — график, позволяющий оценить качество бинарной классификации, отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак, (англ. true positive rate, TPR, называемой чувствительностью алгоритма классификации) и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак (англ. false positive rate, FPR, величина $1 - \text{FPR}$ называется специфичностью алгоритма классификации) при варьировании порога решающего правила.

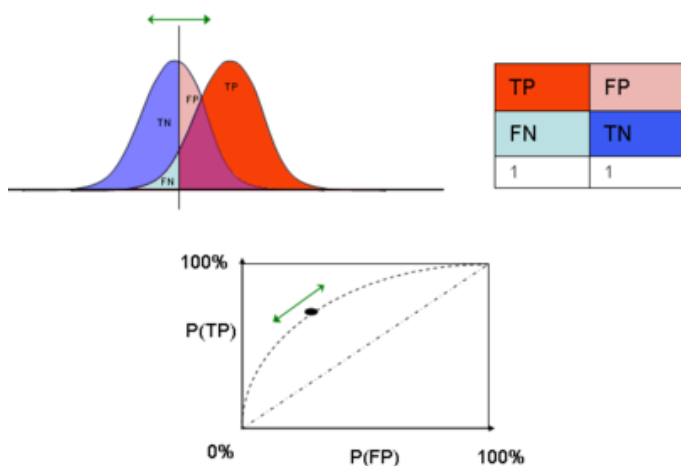
Также известна как кривая ошибок. Анализ классификаций с применением ROC-кривых называется ROC-анализом.

Количественную интерпретацию ROC даёт показатель AUC (англ. area under ROC curve, площадь под ROC-кривой) — площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию). Значение менее 0,5 говорит, что классификатор действует с точностью до наоборот: если положительные назвать отрицательными и наоборот, классификатор будет работать лучше.

мои комментарии Если кто-то хорошо представляет себе, как выглядит график зависимости мощности от ошибки первого рода, то это именно такой график. Меняется уровень значимости (как порог отвергнуть - не отвергнуть) и по оси x откладывается ошибка пер-

вого рода, она же false positive rate $FP/(TN+FP)$, а по оси y откладывается мощность, она же true positive rate $TP/(TP+FN)$ (слово positive означает, что нулевая гипотеза отвергнута в пользу второй, альтернативной, гипотезы, а в случае классификации, что элемент классифицируется как относящийся ко второму классу).

Таким образом, меняем порог/параметр для метода классификации (пример параметра — априорная вероятность π_1) и по оси x откладываем долю неправильно классифицированных элементов из первого класса ($n_{12}/(n_{11} + n_{12})$, FPR), а по оси y — долю правильно классифицированных элементов из второго класса ($n_{22}/(n_{22} + n_{21})$, TPR).



Пусть классы имеют вид 4,6,8,10,12 первый и 1,3,5,7 второй. Опишем ROC-кривую. Пусть к первому классу мы относим, если число больше порога γ . Для $\gamma < 1$ мы находимся в точке $(0, 0)$. При $1 < \gamma < 3$ мы перескакиваем в точку $(0, 0.25)$. При $3 < \gamma < 4$ мы перескакиваем в точку $(0, 0.5)$. При $4 < \gamma < 5$ мы перескакиваем в точку $(0.2, 0.5)$. При $5 < \gamma < 6$ мы перескакиваем в точку $(0.2, 0.75)$. При $6 < \gamma < 7$ мы перескакиваем в точку $(0.4, 0.75)$. При $7 < \gamma < 8$ мы перескакиваем в точку $(0.4, 1)$. Дальше мы при $x = 1$ последовательно перескакиваем по y в 0.6, 0.8 и при $\gamma > 12$ попадаем в точку $(1, 1)$.

2 Кластерный анализ

2.1 Кластерный анализ

Цель кластерного анализа — разбить индивиды на кластеры, т.е., на группы, между которыми, в некотором смысле, расстояние больше, чем между точками внутри. Задача не формализована и, можно сказать, плохо поставлены, поэтому решается плохо.

Вообще, кластерный анализ — это ‘обучение без учителя’. Это означает, что вы не сможете формально проверить правильность результата.

Единственный вариант поставить задачу четко — это предположить какую-то статистическую модель данных и в ней находить параметры, например, по методу максимального правдоподобия (model-based clustering).

Все остальные методы — эвристические с плохо определенным (хорошо-плохо) результатом.

2.2 Кластерный анализ, пример model-based подхода

Предположим, что многомерная выборка — неоднородная. Но в отличие от дискриминантного анализа у нас нет признака, объясняющего эту неоднородность, и задачей является ее выявить. Тип классификации, когда есть модель, называется model-based clustering. Например, пусть наша выборка из смеси k нормальных распределений. Таким образом ее плотность имеет вид

$$p(x) = \pi_1 p(x, \mu_1, \Sigma_1) + \dots + \pi_k p(x, \mu_k, \Sigma_k), \quad (2)$$

где

$$p(x, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2} \sqrt{|\Sigma_i|}} \exp \left(-\frac{1}{2} (x - \mu_i) \Sigma_i^{-1} (x - \mu_i)^T \right) \quad (3)$$

Эта задача решается методом максимального правдоподобия. Можно выписать функцию правдоподобия (выпишите), но она имеет сложный вид и искать ее максимум по такому большому числу параметров очень непросто. Для нахождения этого максимума используется так называемый ЕМ-алгоритм (Expectation - Maximization). Мы не будем здесь его обсуждать.

2.3 Кластерный анализ: k -means, k -means++

Хотим искать кластеры C_1, \dots, C_k минимизируя следующий функционал

$$\sum_{i=1}^k \sum_{j \in C_i} \|x_j - \mu_i\|^2 \quad (4)$$

по разбиению всего пространства индивидов на C_j и по всем μ_i . Можно делать это по следующему алгоритму:

1. Выбираем случайно μ_1, \dots, μ_k .
2. C_j — кластер, содержащий точки, которые лежат к μ_j ближе, чем к остальным μ_i .
3. Для каждого C_j пересчитываем центр μ_j как выборочное среднее элементов из этого кластера.
4. Делаем 2 и 3 пока алгоритм не сойдется.

Проблема метода в том, что у такого функционала много локальных минимумов, и алгоритм может сойтись в значение, далекое от истинного. Метод k -means++ повторяет алгоритм, приведенный выше, но начальные значения выбираются не случайно, а следующим образом

1. Выбираем случайным образом первый центр μ_1 .
2. Считаем расстояние от всех точек до ближайшего центра $\{\rho_i\}$. После чего выбираем x_i как новый центр с вероятностью, пропорциональной ρ_i .
3. Пока количество центров меньше, чем k , повторяем процедуру.

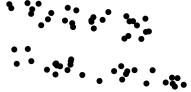

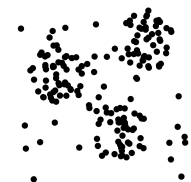

Результат функционала в k -means для данной процедуры выбора начальных центров запишем как $J(\{C_j\}, \{\mu_j\})$. Известно, что при некоторых условиях на форму кластеров

$$\frac{\mathbb{E}(J(\{C_j\}, \{\mu_j\}))}{J_{min}} = O(\ln k),$$

т.е. результат, в среднем, довольно близко к настоящему минимуму.

Замечание. *Есть результаты, что если к данным применить анализ главных компонент, то пространство, натянутое на первые $k - 1$ главных векторов, при некоторых условиях будет близко к пространству, проходящему, через центры кластеров. Поэтому часто с помощью АГК уменьшают число признаков и потом применяют процедуру кластерного анализа.*

2.4 «Плохие» кластерные структуры

1.  ленточные кластеры. Внутрикластерные расстояния могут быть больше межкластерных;
2.  перекрывающиеся кластеры;
3.  кластеры, соединяющиеся перемычками и накладывающиеся на фон из редко расположенных объектов;
4.  кластеры могут отсутствовать.

Здесь мы обсуждали, что практически невозможно придумать определение кластера (не статистическое), при котором все эти кластеры будут ему удовлетворять. Вариант смеси нормальных распределений, возможно, подойдет во всех случаях.

Еще обсуждали вопрос, что для данных, где реально обособленных кластеров может и не быть (например, последняя картинка), часто кластеризацией называют сегментацию — просто нарезку на части с описанием каждого сегмента на основе значений признаков.

2.5 Иерархический кластерный анализ

2.5.1 Расстояние между точками ρ

Сначала нужно задать, как мы будем измерять расстояние между точками.

Самое стандартное — евклидово расстояние: $\rho(x, y) = (\sum_i (x_i - y_i)^2)^{1/2}$.

Расстояние городских кварталов (манхэттенское расстояние): $\rho(x, y) = \sum_i |x_i - y_i|$.

Расстояние Чебышёва: $\rho(x, y) = \max_i |x_i - y_i|$.

Процент несогласия (эта мера используется в тех случаях, когда данные являются категориальными): $\rho(x, y) = (\#\{i : x_i \neq y_i\})/i$.

Особый случай, если кластеризуются признаки, а не индивиды (а какая разница — такой кластерный анализ не статистическая процедура, ему все равно), то логично в качестве расстояния рассматривать корреляции. Например, 1 минус модуль корреляции или 1 минус просто корреляция, что правильнее по смыслу для задачи.

Замечание. Важно либо исходно стандартизовать признаки, либо измерять расстояние специальным образом. Например, использовать расстояние Махаланобиса вместо обычного евклидова, если есть предположения о форме распределения точек внутри кластера.

2.5.2 Примеры межкластерных расстояний

Правила слияния кластеров (linkage rule) основывается на расстояниях между кластерами.

Расстояние ближнего соседа (single linkage, кластеры в виде цепочек):

$$R^n(U, V) = \min_{u \in U, v \in V} \rho(u, v), \quad U, V \subset X;$$

расстояние дальнего соседа (complete linkage, кластеры ближе к шарикам):

$$R^l(U, V) = \max_{u \in U, v \in V} \rho(u, v);$$

групповое среднее расстояние:

$$R^g(U, V) = \frac{1}{|U||V|} \sum_{u \in U} \sum_{v \in V} \rho(u, v);$$

расстояние между центрами:

$$R^c(U, V) = \rho^2 \left(\sum_{u \in U} \frac{u}{|U|}, \sum_{v \in V} \frac{v}{|V|} \right);$$

расстояние Уорда:

$$R^w(U, V) = \frac{|U||V|}{|U| + |V|} R^c(U, V).$$

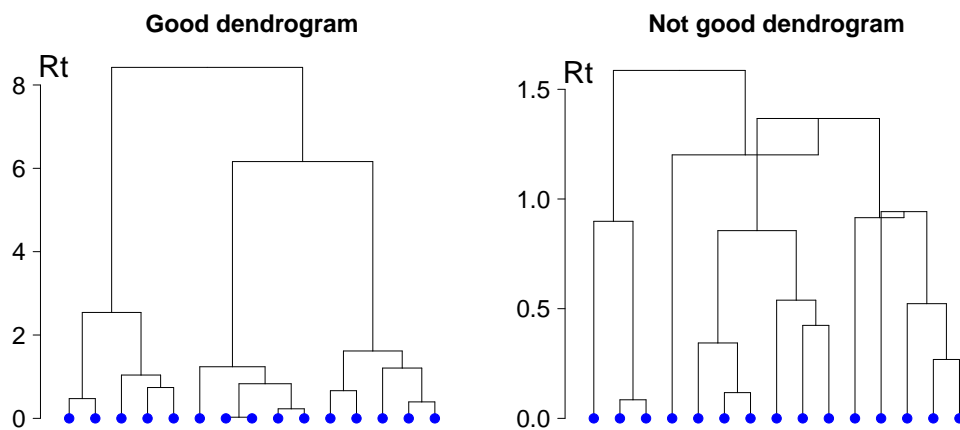
2.5.3 Алгоритм агломеративной иерархической кластеризации

1. Сначала все кластеры одноэлементные: $C_1 = \{\{x_1\}, \dots, \{x_l\}\}$; $R_1 = 0$;
 $\forall i \neq j$ вычислить $R(\{x_i\}, \{x_j\})$;
2. для всех $t = 2, \dots, l$ (t — номер итерации)

3. найти в C_{t-1} два ближайших кластера:
 $(U, V) = \arg \min_{U \neq V} R(U, V);$
 $R_t = R(U, V);$
4. слить их в один кластер:
 $W = U \cup V;$
 $C_t = C_{t-1} \cup W \setminus \{U, V\};$
5. для всех $S \in C_t \setminus W$
6. вычислить $R(W, S);$

2.5.4 Визуализация кластерной структуры

Определение. Дендрограмма — деревоподобный график, отражающий процесс последовательных слияний и структуру кластеров.



После построения дерева можно его разрезать на поддеревья по заданному расстоянию между кластерами и получить сами кластеры. Разрез делается там, где долго не было объединения кластеров (длинная ветка у дерева).

Но долго-недолго — это субъективно и зависит от выбранного расстояния. Если расстояние в квадрате, то дальние ветки искусственно удлиняются.