Пакет 'отримле'

29 мая 2021 г.

Тип Пакет

Название Надежная кластеризация на основе моделей

Описание Выполняет надежный кластерный анализ с учетом выбросов и шума, которые невозможно подобрать.
тед любым кластером. Данные моделируются смесью гауссовых распределений.
и шумовой составляющей, которая представляет собой неправильное равномерное распределение, охватывающее всю Европу.
клидовое пространство. Параметры оцениваются по (псевдо) максимальной вероятности. Это подходит-
с помощью алгоритма EM-типа. См. Коретто и Хен-
ниг (2016) <doi: 10.1080 / 01621459.2015.1100996> и Коретто и Хен-
ниг (2017) <https://jmlr.org/papers/v18/16-382.html>.

Версия 2.0

Дата 2021-05-28

Импортирует статистику, утилиты, графику, grDevices, mvtnorm, parallel, foreach,
doParallel, robustbase, mclust

Лицензия GPL (> = 2)

LazyData ИСТИНА

Требуется компиляция нет

Автор Пьетро Коретто [aut, cre] (Домашняя страница:
<https://pietro-coretto.github.io>),
Кристиан Хенниг [aut] (Домашняя страница:
<https://www.unibo.it/sitoweb/christian.hennig/en>)

Сопровождающий Pietro Coretto <pcoretto@unisa.it>

Репозиторий CRAN

Дата / Публикация 2021-05-29 06:30:02 UTC

Задокументированные темы R:

денежная купюра Данные о швейцарских банкнотах

Описание

Данные из таблиц 1.1 и 1.2 (стр. 5-8) Flury и Riedwyl (1988). Есть шесть измерений
Сделано на 200 швейцарских банкнотах (старые швейцарские 1000 франков). Банкноты относятся к двум классам
одинаковый размер: подлинный и поддельный.

Применение

данные (банкнота)

Формат

Data.frame размером 200x7 со следующими переменными:

Класс фактор с классами: подлинный, поддельный

Длина Длина банкноты (мм)

Левая ширина левого края (мм)

Правая ширина правого края (мм)

Нижняя ширина нижнего поля (мм)

Верхняя ширина верхнего поля (мм)

Диагональ Длина диагонали (мм)

Источник

Флури, Б. и Ридвил, Х. (1988). Многомерная статистика: практический подход. Лондон: Чепмен
& Зал.

генератор.отримле Генерирует случайные данные из выходной модели OTRIMLE

Описание

При этом используются данные и выходные данные otrimle или rimle для создания нового искусственного набора данных размера исходных данных с использованием шума и пропорций кластера из выходных данных кластеризации. Кластеры затем генерируется из многомерных нормальных распределений с параметрами, оцененными с помощью otrimle, шум генерируется передискретизацией из того, что оценивается как компонент шума с заданными весами апостериорной вероятностью того, что все наблюдения являются шумом. См. Хенниг и Коретто (2021 г.).

Применение

```
generator.otrimle (данные, подходят)
```

Аргументы

данные	что-то, что можно принудить к матрице. Набор данных.
соответствовать	выходной объект отримле или обода.

Значение

Список с данными компонентов, кластеризация.

данные	матрица сгенерированных данных.
CS	вектор целых чисел. Индикатор кластеризации.

Авторы)

Кристиан Хенниг <christian.hennig@unibo.it> https://www.unibo.it/sitoweb/christian .
hennig / en /

Рекомендации

Хенниг, К. и П. Коретто (2021 г.). Подход адекватности для определения количества кластеров для OTRIMLE кластеризация на основе робастной смеси Гаусса. Появиться в Австралии и Новой Зеландии Статистический журнал, https://arxiv.org/abs/2009.00921.

Смотрите также

kerndensp , kerndensmeasure , otrimle , ободок

Примеры

```
данные (банкнота)
selectdata <- c (1: 30,101: 110,117: 136,160: 161)
set.seed (555566)
x <- банкнота [selectdata, 5: 7]
ox <- otrimle (x, G = 2, ncores = 1)
str (generator.otrimle (x, ox))
```

Описание

Вычисляет начальное назначение кластера на основе комбинации шума на основе ближайшего соседа.
и агломеративная иерархическая кластеризация на основе критериев максимального правдоподобия для Gaus-
сианские модели смесей.

Применение

```
InitClust (данные, G, k = 3, knnd.trim = 0,5, modelName = VVV)
```

Аргументы

данные	Числовой вектор, матрица или фрейм данных наблюдений. Строки соответствуют наблюдения и столбцы соответствуют переменным. Категориальные переменные и NA значения не допускаются.
грамм	Целое число, определяющее количество кластеров.
k	Целое число, определяющее количество рассматриваемых ближайших соседей на каждую используемую точку. для шага шумоподавления (см. Подробности).
knnd.trim	Число в [0,1), которое определяет долю точек, инициализированных как шум. Обычно knnd.trim <= 0,5 (см. Подробности).
название модели	Строка символов, указывающая модель ковариации, которая будет использоваться. Возможные модели находятся: "E": равная дисперсия (одномерная) "V": сферический, с переменной дисперсией (одномерный) «EII»: сферический, равного объема «VII»: сферическая, неодинакового объема. "EEE": эллипсоид, равный объем, форма и ориентация. «VVV»: эллипсоид, переменного объема, формы и ориентации (по умолчанию). Смотрите подробности.

Подробности

Инициализация основана на Coretto and Hennig (2017). Сначала выполняются два шага:

Шаг 1 (шаг шумоподавления): для каждой точки данных вычислить расстояние до ближайших соседей (k-NND).
Все точки с k-NND больше, чем (1-knnd.trim) -квантиль k-NND, инициализируются как
шум. Интерпретация k заключается в том, что (k-1), но не k, точки, расположенные близко друг к другу, все еще могут интерпретироваться как шум или выбросы

Шаг 2 (этап кластеризации): выполните иерархическую кластеризацию на основе модели (MBHC), предложенную в
Фрейли (1998). Этот шаг выполняется с помощью hc. Входной аргумент modelName передается hc.
См. Подробности о hc для получения более подробной информации.

Если на предыдущем шаге 2 не удалось предоставить кластеры G, каждый из которых содержит не менее 2 различных точек данных, выполняется заменена классической иерархической кластеризацией, реализованной в hclust. Наконец, если hclust не
укажите допустимый раздел, пробуется до десяти случайных разделов.

Значение

Целочисленный вектор, определяющий начальное назначение кластера, где 0 обозначает шум / выбросы.

Рекомендации

Фрейли, К. (1998). Алгоритмы гауссовской иерархической кластеризации на основе моделей. SIAM Журнал на Научные вычисления 20: 270-281.

П. Коретто и К. Хенниг (2017). Последовательность, отказоустойчивость и алгоритмы для надежной неправильная кластеризация максимального правдоподобия. Журнал исследований в области машинного обучения, Vol. 18 (142), С. 1-39. https://jmlr.org/papers/v18/16-382.html

Авторы)

Пьетро Коретто <pcoretto@unisa.it> https://pietro-coretto.github.io

Смотрите также

hc

Примеры

```
## Загрузить данные о швейцарских банкнотах
данные (банкнота)
x <- банкнота [, - 1]

## Начальные кластеры с аргументами по умолчанию
init <- InitClust (данные = x, G = 2)
печать (инициализация)

## Выполнить отримле
a <- otrimle (data = x, G = 2, initial = init,
                        logicd = c (-Inf, -50, -10), ncores = 1)
plot (a, what = "кластеризация", data = x)
```


Kerndenscluster Агрегированное расстояние до эллиптической унимодальной плотности по кластерам

Описание

Это вызывает kerndensp для вычисления и агрегирования распределения на основе плотности и главных компонентов. соотношения между многомерными данными и одномодальным эллиптическим распределением относительно среднего значения данных для всех кластеры в кластеризации на основе смеси, генерируемые otrimle или rimle. Для использования в отримлеге.

Применение

```
kerndenscluster (x, подгонка, maxq = qnorm (0,9995), kernn = 100)
```

Аргументы

Икс	что-то, что можно принудить к матрице. Набор данных.
соответствовать	выходной объект отримле или обода.
maxq	положительное числовое. Одномерные плотности оцениваются между средним (x) -maxq и mean (x) + maxq.
Kernn	целое число. Количество точек, в которых оценивается одномерная плотность, входной параметр плотности n. Это должно быть ровно.

Подробности

См. Hennig and Coretto (2021), Sec. 4.2. kerndenscluster вызывает kerndensp для всех кластеров и

агрегирует полученные меры в виде корня из суммы квадратов.

Значение

Список с компонентами ddpi, ddpm, measure.

ddpi	список выходов kerndensp для всех кластеров.
ddpm	вектор мер-компонентов ddpi.
мера	Окончательный результат агрегирования.

Авторы)

Кристиан Хенниг <christian.hennig@unibo.it> https://www.unibo.it/sitoweb/christian . hennig / en /

Рекомендации

Хенниг, К. и П. Коретто (2021 г.). Подход адекватности для определения количества кластеров для OTRIMLE кластеризация на основе робастной смеси Гаусса. Появиться в Австралии и Новой Зеландии Статистический журнал, https://arxiv.org/abs/2009.00921.

Смотрите также

kerndensp , kerndensmeasure , otrimle , ободок

Примеры

```
данные (банкнота)
selectdata <- c (1: 30,101: 110,117: 136,160: 161)
set.seed (555566)
x <- банкнота [selectdata, 5: 7]
ox <- otrimle (x, G = 2, ncores = 1)
kerndenscluster (x, ox) $ мера
```

Kerndensmeasure Статистическое измерение близости к симметричному унимодальному распределению

Описание

Расстояние на основе плотности между одномерными данными и унимодальным симметричным распределением около среднее значение данных основано на Pons (2013, стр.79), адаптировано Хеннигом и Коретто (2021), см. подробности.

Применение

```
kerndensmeasure (x, weights = rep (1, nrow (as.matrix (x))), maxq = qnorm (0,9995),
                 kernn = 100)
```

Аргументы

Икс	вектор. Одномерный набор данных.
веса	неотрицательный вектор. Относительные веса наблюдений (будут стандартизированы до поддерживать до одного внутри).
maxq	плотности оцениваются между средним (x) -maxq и средним (x) + maxq.
Kernn	целое число. Количество точек, в которых оценивается плотность, входной параметр n

от плотности. Это должно быть ровно.

Подробности

Функция плотности используется для вычисления оценки плотности ядра на основе данных. Кернн значения плотности затем упорядочиваются от наибольшего до наименьшего. Начиная от самого большого до наименьшее, формируются пары из двух значений (наибольшее и наибольшее наибольшее, третье и четвертое наибольшее и скоро). Каждая пара заменяется двумя копиями среднего двух значений. Затем с каждой стороны среднее значение каждой копии размещается от самого большого до самого маленького, и это дает симметричный плотность около среднего. Среднеквадратичная разница между этой плотностью и исходной плотностью вычисляется.

Значение

Список с компонентами cp, cpx, measure.

cp	вектор сгенерированных значений симметричной плотности от наибольшего к наименьшему (всего один копия, sp kernn / 2 значения).
cpx	y-компонента плотности-выход.
мера	среднеквадратичная разница между плотностями.

Авторы)

Кристиан Хенниг <christian.hennig@unibo.it> https://www.unibo.it/sitoweb/christian . hennig / en /

Рекомендации

Хенниг, К. и П. Коретто (2021 г.). Подход адекватности для определения количества кластеров для OTRIMLE кластеризация на основе робастной смеси Гаусса. Появиться в Австралии и Новой Зеландии Статистический журнал, https://arxiv.org/abs/2009.00921.

Понс, О. (2013). Статистическая проверка непараметрических гипотез: асимптотическая теория. Мировая наука tific, Сингапур.

Смотрите также

Kerndensp

Примеры

```
set.seed (124578)
x <- runif (20)
str (kerndensmeasure (x))
```

Kerndensp Близость многомерного распределения к эллиптическому одномодальному распределению тион

Описание

Расстояние на основе плотности и главных компонентов между многомерными данными и одномодальным эл. липтическое распределение о средних данных, см. Hennig and Coretto (2021). Для использования в kerndenscluster.

Применение

```
kerndensp (x, weights = rep (1, nrow (as.matrix (x))), siglist, maxq = qnorm (0,9995),
          kernn = 100)
```


Аргументы

Икс	что-то, что можно принудить к матрице. Набор данных.
веса	неотрицательный вектор. Относительные веса наблюдений (будут стандартизированы до поддерживать до одного внутри).
сиглист	список с компонентами cov (ковариационная матрица), center (среднее) и n.obs (num-количество наблюдений).
maxq	положительное числовое. Одномерные плотности оцениваются между средним (x) -maxq и mean (x) + maxq.
Kernn	целое число. Количество точек, в которых оценивается одномерная плотность, входной параметр плотности n. Это должно быть ровно.

Подробности

См. Hennig and Coretto (2021), Sec. 4.2. kerndensmeasure выполняется на основных компонентах Икс. Полученные показатели стандартизируются с помощью kmeanfun и ksdfun, а затем агрегируются как среднее значение.

квадрат положительных значений, см. Hennig and Coretto (2021). PCS вычисляется princomp и всегда будет использовать siglist, а не статистику, вычисленную по x.

Значение

Список с компонентами cml, cm, pca, stanmeasure, measure.

cml	Список выходных данных kerndensmeasure для основных компонентов.
см	вектор компонент меры керненсера для главного компонента ненцы.
stanmeasure	вектор компонент стандартизированной меры керненсера для принципа ципальные компоненты.
pca	вывод princomp.
мера	Окончательный результат агрегирования.

Авторы)

Кристиан Хенниг <christian.hennig@unibo.it> https://www.unibo.it/sitoweb/christian . hennig / en /

Рекомендации

Хенниг, К. и П. Коретто (2021 г.). Подход адекватности для определения количества кластеров для OTRIMLE кластеризация на основе робастной смеси Гаусса. Появиться в Австралии и Новой Зеландии Статистический журнал, https://arxiv.org/abs/2009.00921.

Смотрите также

kerndensmeasure , kerndenscluster

Примеры

```
set.seed (124578)
x <- cbind (runif (20), runif (20))
siglist <- список (cov = cov (x), center = colMeans (x), n.obs = 20)
kerndensp (x, siglist = siglist) $ мера
```

Kmeanfun Среднее и стандартное отклонение статистики унимодальности

Описание

Эти функции аппроксимируют среднее значение и стандартное отклонение вычисленной статистики унимодальности. by kerndensmeasure, предполагающий стандартные гауссовские данные, зависящие от количества наблюдений. Они были выбраны на основе моделирования с 74 различными значениями n. Используется для стандартных isation в kerndensp.

Применение

```
kmeanfun (сущ)
ksdfun (сущ.)
```

Аргументы

n	целое число. Количество наблюдений.

Значение

Полученное среднее (kmeanfun) или стандартное отклонение (ksdfun).

Авторы)

Кристиан Хенниг <christian.hennig@unibo.it> https://www.unibo.it/sitoweb/christian . hennig / en /

Смотрите также

Kerndensp

Примеры

```
кмеанфун (50)
ksdfun (50)
```

Описание

otrimle ищет G кластеры приблизительно гауссовой формы с / без шума / выбросов. В
настройка метода, контролирующего уровень шума, выбирается адаптивно на основе данных.

Применение

```
otrimle (данные, G, начальное = NULL, logicd = NULL, npr.max = 0,5, erc = 20,
beta = 0, iter.max = 500, tol = 1e-06, ncores = NULL, monitor = TRUE)
```

Аргументы

данные	Числовой вектор, матрица или фрейм данных наблюдений. Строки соответствуют наблюдения и столбцы соответствуют переменным. Категориальные переменные и NA значения не допускаются.
грамм	Целое число, определяющее количество кластеров.
исходный	Целочисленный вектор, определяющий начальное назначение кластера, где 0 обозначает шум / выбросы. Если NULL (по умолчанию) инициализация выполняется с помощью InitClust.
логичный	Вектор, определяющий сетку логарифмических (icd) значений, где icd обозначает неправильное совпадение. постоянная плотность. Если logicd = NULL, рассматривается сетка по умолчанию. Чистый гауссовский Подгонка модели смеси (полученная, когда log (icd) = - Inf) включена в значение по умолчанию. путь поиска.
npr.max	Число в [0,1), определяющее максимальную долю шума / выбросов. Этот определяет ограничение доли шума. Если npr.max = 0, одно решение без вычисляется шумовая составляющая (соответствующая логикеd = -Inf.
эрк	Число> = 1, указывающее максимально допустимое соотношение между внутри кластера собственные значения ковариационной матрицы. Это определяет ограничение на собственное отношение. erc = 1 обеспечивает сферические кластеры с равными ковариационными матрицами. Большой erc позволяет для больших межкластерных расхождений ковариации. Чтобы облегчить При настройке erc предлагается масштабировать столбцы данных (см. масштаб) в любое время единицы измерения различных переменных совершенно несовместимы.
бета	Неотрицательная константа. Это коэффициент бета-штрафа, введенный в Коретто и Хенниг (2016).
iter.max	Целочисленное значение, определяющее максимальное количество итераций, разрешенных в лежащий в основе ECM-алгоритм.
тол	Критерий остановки для лежащего в основе ECM-алгоритма. Итерация ECM останавливается если два последовательных неправильных значения логарифма правдоподобия находятся в пределах допуска.
ncores	целочисленное значение, определяющее количество ядер, используемых для параллельных вычислений. Когда ncores = NULL (по умолчанию), количество доступных ядер r определяется, и (r-1) из них используются (см. подробности). Если ncores = 1, параллельный бэкэнд не запускается.
монитор	логично. Если ИСТИНА, сообщения о ходе выполнения печатаются на экране.

Подробности

Функция otrimle вычисляет решение OTRIMLE на основе ECM-алгоритма (ожидание
алгоритм условной максимизации), предложенный в Coretto and Hennig (2017).

Критерий otrimle минимизируется по логической сетке логарифмических (icd) значений с использованием параллельных вычислений.
ing на основе foreach. Обратите внимание, что, в зависимости от настройки BLAS / LAPACK, увеличение ncores
может не привести к желаемому сокращению времени вычислений. Последнее происходит при оптимизации линейного
используются процедуры алгебры (например, OpenBLAS, Intel Math Kernel Library (MKL) и т. д.). Эти оптимальные

В миниатюрных разделяемых библиотеках уже реализована многопоточность. Следовательно, в этом случае увеличение ncores может лишь незначительно сократить время вычислений.

Иногда могут быть наборы данных, для которых функция не дает решения, основанного на де-
аргументы вины. Это соответствует code = 0 и flag = 1 или flag = 2 на выходе (см. Раздел «Значение»).
ниже). Обычно это происходит, когда возникают некоторые (или все) из следующих обстоятельств: (i) erc is
слишком большой; (ii) npr.max слишком велик; (iii) выбор начального разбиения. Что касается (i) и (ii), это
невозможно дать числовые ссылки, потому что, являются ли эти числа слишком большими / маленькими, сильно
зависит от размера выборки и размерности данных. Ссылки, приведенные ниже, объясняют
соотношение между этими величинами.

Предлагается попробовать следующее всякий раз, когда возникает ошибка code = 0. Установите логический диапазон
достаточно широкий (например, logicd = seq (-500, -5, length = 50)), выберите erc = 1 и низкий выбор npr.max
(например, npr.max = 0,02). Следите за решением с помощью графика профилирования критериев (сюжет.). Ас-
в соответствии с логикой изменения графика профилирования критерия и увеличения erc и npr.max до точки
когда получен "четкий" минимум в графике профилирования критериев. Если эта стратегия не сработает,
Предлагается поэкспериментировать с другими начальными разделами (см. начальный выше).

TBA: Кристиан может добавить сюда кое-что о бета-версии.

Объект pi, возвращаемый функцией rimle (см. Value), соответствует вектору параметров pi
в базовой псевдомодели (1), определенной в Coretto and Hennig (2017). С logicd = -Inf
функция Римла аппроксимирует MLE для модели простой гауссовой смеси с собственным соотношением
ковариационная регуляризация, в этом случае первый элемент вектора pi устанавливается в ноль, потому что
шумовая составляющая не учитывается. В общем, для выборки iid из моделей конечной смеси
В контексте эти параметры pi определяют ожидаемые пропорции кластеров. Из-за доли шума
ограничение в RIMLE, бывают ситуации, когда это соединение может не произойти на практике. В
последнее, вероятно, произойдет, когда как logicd, так и npr.max велики. Таким образом, расчетная ожидаемая
пропорции кластеров указываются в объекте экспропорции на выходе обода, и это
вычисляется на основе неправильных апостериорных вероятностей, указанных в тау. Увидеть Коретто и Хеннига
(2017) для более подробного обсуждения этого вопроса.

Более ранняя приближенная версия алгоритма была первоначально предложена Коретто и Хеннигом.
(2016). Программное обеспечение для исходной версии алгоритма можно найти в дополнительном материале.
риалы Коретто и Хеннига (2016).

Значение

Объект S3 класса otrimle, обеспечивающий оптимальное (по критерию OTRIMLE)
кластеризация. Компоненты вывода следующие:

код	Целочисленный индикатор сходимости. code = 0, если решение не найдено (см. Подробности); code = 1, если при оптимальном значении icd соответствующий EM-алгоритм не сходились в пределах em.iter.max; code = 2 сходимость полностью достигнута.

флаг	Строка символов, содержащая один или несколько флагов, связанных с итерацией EM в оптимальный icd. flag = 1, если не удалось предотвратить численное вырождение несобственных апостериорных вероятностей (значение тау ниже). flag = 2, если принудительное исполнение ограничение доли шума не удалось по числовым причинам. flag = 3, если шум ограничение пропорции было успешно применено хотя бы один раз. flag = 4, если ограничение на собственное отношение было успешно применено хотя бы один раз.
iter	Количество итераций, выполненных в базовом EM-алгоритме при оптимальном icd.
логичный	Результирующее значение оптимального журнала (icd).
илоглик	Результирующее значение неправдоподобной вероятности.
критерий	Итоговое значение критерия OTRIMLE.
Пи	Расчетный вектор параметров pi базовой псевдомодели (см. De-хвосты).
иметь в виду	Матрица размерности ncol (данные) x G, содержащая средние параметры каждого кластер (по столбцам).

cov	Массив размером ncol (данные) x ncol (данные) x G, содержащий ковариационную ма- матрицы каждого кластера.
тау	Матрица размерности nrow (data) x {1 + G}, где tau [i, 1 + j] - оценка (несобственная) апостериорная вероятность того, что i-е наблюдение принадлежит j-му кластеру - тер. tau [i, 1] - это оценочная (несобственная) апостериорная вероятность того, что i-е наблюдение ция относится к шумовой составляющей.
smd	Матрица размерности nrow (данные) x G, где smd [i, j] - квадрат Маха- расстояние lanobis данных [i,] от среднего [, j] согласно cov [,, j].
кластер	Вектор целых чисел, обозначающий назначения кластера для каждого наблюдения. Это 0 для наблюдения, относящиеся к шуму / выбросам.
размер	Вектор целых чисел с размерами (количеством) каждого кластера.
экспропорция	Вектор предполагаемых ожидаемых пропорций кластеров (см. Подробности).
оптимизация	Data.frame с профилированием оптимизации OTRIMLE. Для каждого значения log (icd) исследуемый алгоритмом data.frame хранит логику, критерий, илоглик, код, флаг (определено выше), а enpr обозначает ожидаемую долю шума.

Рекомендации

Коретто, П. и К. Хенниг (2016). Надежная неправильная максимальная вероятность: настройка, вычисление и сравнение с другими методами робастной гауссовой кластеризации. Журнал американской статистики Ассоциация, Vol. 111 (516), стр. 1648–1659. DOI: 10.1080 / 01621459.2015.1100996

П. Коретто и К. Хенниг (2017). Последовательность, отказоустойчивость и алгоритмы для надежной неправильная кластеризация максимального правдоподобия. Журнал исследований в области машинного обучения, Vol. 18 (142), С. 1-39. https://jmlr.org/papers/v18/16-382.html

Авторы)

Пьетро Коретто <pcoretto@unisa.it> https://pietro-coretto.github.io

Смотрите также

plot.otrimle , InitClust , ободок,

Примеры

```
## Загрузить данные о швейцарских банкнотах
данные (банкнота)
x <- банкнота [, - 1]

## Выполните отримле кластеризацию с аргументами по умолчанию
set.seed (1)
a <- otrimle (data = x, G = 2, logicd = c (-Inf, -50, -10), ncores = 1)

## Кластеризация графиков
plot (a, data = x, what = "кластеризация")

## Постройте профилирование критерия OTRIMLE
сюжет (a, what = "критерий")

## Постройте неправильное профилирование логарифмической вероятности
сюжет (a, what = "iloglik")

## PP график кластерного эмпирического взвешенного квадрата Махаланобиса
## расстояния относительно целевого распределения pchisq (, df = ncol (data))
сюжет (a, what = "fit")
plot (a, what = "fit", cluster = 1)
```

```
## Выполните то же отримле, что и раньше, с ненулевым штрафом
set.seed (1)
b <- otrimle (data = x, G = 2, beta = 0.5, logicd = c (-Inf, -50, -10), ncores = 1)

## Кластеризация графиков
plot (b, data = x, what = "кластеризация")

## Постройте профилирование критерия OTRIMLE
сюжет (b, what = "критерий")

## Постройте неправильное профилирование логарифмической вероятности
сюжет (b, what = "iloglik")

## PP график кластерного эмпирического взвешенного квадрата Махаланобиса
## расстояния относительно целевого распределения pchisq (, df = ncol (data))
сюжет (b, what = "fit")
plot (b, what = "fit", cluster = 1)
```



```
## Не работать:
## Выполните тот же пример, используя более тонкую сетку по умолчанию logicd
## значения с использованием нескольких ядер
##
a <- otrimle (data = x, G = 2)

## Проверьте критерий otrimle-vs-logicd
сюжет (a, что = критерий)

## Минимум достигается при $ logicd = -9, и исследуя $ оптимизацию, это
## видно, что интервал [-12,5, -4] ограничивает оптимальную
## решение. Мы ищем с более мелкой сеткой, расположенной около минимума
##
b <- otrimle (данные = x, G = 2, logicd = seq (-12,5, -4, length.out = 25))

## Проверьте критерий otrimle-vs-logicd
сюжет (б, что = критерий)

## Проверьте разницу между двумя кластерами
таблица (A = a $ cluster, B = b $ cluster)

## Проверить различия в оценочных параметрах
##
colSums (abs (a $ mean - b $ mean))                    ## Расстояние L1 для средних векторов
применить ({a $ cov-b $ cov}, 3, norm, type = "F")     ## Расстояние Фробениуса для ковариаций
c (Шум = abs (a $ npr-b $ npr), abs (a $ cpr-b $ cpr)) ## Абсолютная разница в пропорциях

## Конец (не запускается)
```

отримлег OTRIMLE для диапазона количества кластеров с кластерами на основе плотности-статистика качества

Описание

Вычисляет оптимально настроенную робастную неправильную кластеризацию максимального правдоподобия (OTRIMLE), см. отримлевместе со статистикой качества кластера Q на основе плотности (Hennig and Coretto 2021) для

диапазон значений количества кластеров.

Применение

otrimleg (набор данных, G = 1: 6, многоядерный = ИСТИНА, ncores = detectCores (логический = ЛОЖЬ) -1,
erc = 20, beta0 = 0, fixlogicd = NULL, monitor = 1, dmaxq = qnorm (0,9995))

Аргументы

набор данных	что-то, что может быть преобразовано в матрицу переменных, умноженных на наблюдения. В набор данных.
грамм	вектор целых чисел (обычно начиная с 1). Количество рассматриваемых кластеров эред.

многоядерный	логично. Если TRUE, параллельные вычисления используются через функцию mclapply из пакет параллельный; прочтите там предупреждения, если вы собираетесь использовать это; это не сработает в Windows.
ncores	целое число. Количество ядер для распараллеливания.
эрк	Число больше или равное единице, указывающее максимально допустимое соотношение между между собственными значениями внутрикластерной ковариационной матрицы. См. Отримле.
beta0	Неотрицательная константа, штрафной член за шум, который передается в otrimle как бета., см. документацию там.
fixlogicd	числовое значение NULL. Значение логарифма неправильной логической постоянной плотности, см. ободок, который запускается вместо otrimle, если он не равен NULL. NULL означает, что отримле определяет это по данным.
монитор	0 или 1. Если 1, на экран выводятся сообщения о ходе выполнения.
dmaxq	числовой. Передано как maxq в kerndensmeasure. Интервал, рассматриваемый для одномерная оценка плотности (-maxq, maxq).

Подробности

Для оценки количества кластеров это должно быть вызвано otrimlesimg,. Выход
otrimleg не предназначен для непосредственного использования для оценки количества кластеров, см. Hennig and Коретто (2021 г.).

Значение

otrimleg возвращает список, содержащий компоненты solution, iloglik, ibic, criterion, logicd, noiseprob, denscrit, ddpm.

Все это списки или векторы, номер компонента которых является количеством кластеров.

решение	список объектов вывода отримле или обода.
илоглик	вектор неправильных значений правдоподобия из otrimle.
ибик	вектор неправильных значений BIC (маленький - хорошо), вычисленный из iloglik и количество параметров. Обратите внимание, что поведение неправильной вероятности несовместимо со стандартным использованием BIC, поэтому это экспериментальный и не стоит доверять при выборе количества кластеров.
критерий	вектор значений критерия OTRIMLE, см. отримле.
noiseprob	вектор расчетных пропорций шума, экспропорция [1] из отримле.
denscrit	вектор статистики качества кластера Q на основе плотности (Hennig and Coretto 2021) как обеспечивается компонентом меры ядра kerndensmeasure.
ddpm	список вектора кластерных мер качества кластеров на основе плотности как про- с помощью компонента ddpm программы kerndensmeasure.

Авторы)

Кристиан Хенниг <christian.hennig@unibo.it> https://www.unibo.it/sitoweb/christian . hennig / en /

Рекомендации

Коретто, П. и К. Хенниг (2016). Надежная неправильная максимальная вероятность: настройка, вычисление и сравнение с другими методами робастной гауссовой кластеризации. Журнал американской статистики Ассоциация, Vol. 111 (516), стр. 1648–1659. DOI: 10.1080 / 01621459.2015.1100996

П. Коретто и К. Хенниг (2017). Последовательность, отказоустойчивость и алгоритмы для надежной неправильная кластеризация максимального правдоподобия. Журнал исследований в области машинного обучения, Vol. 18 (142), С. 1-39. https://jmlr.org/papers/v18/16-382.html

Хенниг, К. и П. Коретто (2021 г.). Подход адекватности для определения количества кластеров для OTRIMLE кластеризация на основе робастной смеси Гаусса. Появиться в Австралии и Новой Зеландии Статистический журнал, https://arxiv.org/abs/2009.00921.

Смотрите также

otrimle , rimle , otrimlesimg , kerndensmeasure

Примеры

```
данные (банкнота)
selectdata <- c (1: 30,101: 110,117: 136,160: 161)
x <- банкнота [selectdata, 5: 7]
obanknote <- otrimleg (x, G = 1: 2, multicore = FALSE)
```

отримлесимг Подход адекватности количества кластеров для OTRIMLE

Описание

otrimlesimg вычисляет оптимально настроенную робастную неправильную кластеризацию максимального правдоподобия (OTRIMLE), см. otrimle для диапазона значений количества кластеров, а также для искусственных наборов данных simu- рассчитано по параметрам модели, оцененным на исходных данных. Резюме-методы присутствуют и оцените результаты так, чтобы наименьшее адекватное количество кластеров можно было найти как наименьшее. для которых значение статистики качества кластера Q на основе плотности исходных данных совместимо с его распределением на искусственных наборах данных с тем же количеством кластеров, см. Hennig и Coretto 2021 для подробностей.

Применение

```
otrimlesimg (набор данных, G = 1: 6, многоядерный = ИСТИНА,
ncores = detectCores (логический = FALSE) -1, erc = 20, beta0 = 0, simruns = 20,
sim.est.logicd = ЛОЖЬ,
монитор = 1)

## Метод S3 для класса otrimlesimgdens
сводка (объект, noisepenalty = 0,05, sdcutoff = 2
, ...)
```

```
## Метод S3 для класса summary.otrimlesimgdens
печать (x, ...)

## Метод S3 для класса summary.otrimlesimgdens
plot (x, plot = "критерий", penx = NULL,
peny = NULL, pencex = 1, cutoff = TRUE, ylim = NULL, ...)
```

Аргументы

набор данных	что-то, что может быть преобразовано в матрицу переменных, умноженных на наблюдения. В набор данных.
грамм	вектор целых чисел (обычно начиная с 1). Количество рассматриваемых кластеров эред.
многоядерный	логично. Если TRUE, параллельные вычисления используются через функцию mclapply из пакет параллельный; прочтите там предупреждения, если вы собираетесь использовать это; это не сработает в Windows.
ncores	целое число. Количество ядер для распараллеливания.
эрк	Число больше или равное единице, указывающее максимально допустимое соотношение между между собственными значениями внутрикластерной ковариационной матрицы. См. Отримле.
beta0	Неотрицательная константа, штрафной член за шум, который передается в otrimle как бета., см. документацию там.
симруны	целое число. Количество реплик искусственных наборов данных, взятых из каждой модели.
sim.est.logicd	логический. Если ИСТИНА, логарифм неправильной логики постоянной плотности, см. отримле, переоценивается при запуске otrimle на искусственных наборах данных. Oth- В противном случае значение, оцененное на исходных данных, принимается фиксированным. ИСТИНА требует гораздо больше времени вычислений, но можно рассматривать как генерирующие более реалистичные вариация.
монитор	0 или 1. Если 1, на экран выводятся сообщения о ходе выполнения.
шум	число от 0 до 1. p_0 в Хеннига и Коретто (2021 г.); обычно маленький. Метод предпочитает рассматривать долю <= noiseprenalty точек как выбросы. для добавления кластера.
sdcutoff	числовой. c в формуле (7) Хеннига и Коретто (2021 г.). Кластеризация лечится как адекватный, если его значение измерения качества кластера на основе плотности Q откалибровано (т. е. среднее / стандартное отклонение) по значениям в искусственных наборах данных, сгенерированных из расчетной модели составляет <= sdcutoff.
участок	«критерий» или «шум», см. подробности.
ручка	FALSE, NULL или числовой. x-координата, откуда простота упорядочивания дана кластеризация (как тест на графике). Если ЛОЖЬ, это не добавляется к графику. Если NULL, по умолчанию делается предположение о хорошей позиции (что не всегда работает. хорошо).
копейка	NULL или числовой. x-координата, откуда простота упорядочивания кластеров- Теринг дан (как тест на графике). Если NULL, по умолчанию делается предположение для хорошего положение (что не всегда работает).

пенекс	числовой. Коэффициент увеличения (параметр сех передается в легенду) для простота заказа, см. параметр penx.
отрезать	логично. Если ИСТИНА, "критерий" -пункт показывает значение отсечения, ниже которого количество кластеров адекватное, см. подробности.
Илим	вектор двух числовых значений, диапазон оси Y, который будет передан для построения графика. Если NULL, диапазон выбирается автоматически (но может отличаться от графика по умолчанию).
объект	объект класса otrimlesimgdens, полученный при вызове otrimlesimg
Икс	объект класса summary.otrimlesimgdens, полученный в результате вызова summary функция над объектом класса otrimlesimgdens, полученным в результате вызова отримлесимг.
...	необязательные параметры для передачи на график.

Подробности

Метод полностью описан у Хеннига и Коретто (2021). Необходимые константы настройки для выбор оптимального количества кластеров, наименьшего процента дополнительного шума, который пользователь готов обменять на добавление еще одного кластера (p_0 в документе, noisepenalty здесь) и критическое значение (c в документе, sdcutoff здесь) для соответствия стандартизированной плотности на основе качества мера Q передаются в итоговую функцию, которая требуется для выбора наилучшего (простейшего адекватное) количество кластеров.

Функция построения графика plot.summary.otrimlesimgdens может создавать два графика. Если сюжет = "критерий", стандартизованная мера качества кластера Q на основе плотности нанесена на график в зависимости от количества кластеров. Значения для смоделированных искусственных наборов данных являются точками, значения для исходного набора данных даны. как тип линии. Если cutoff = "TRUE", критические значения (см. Выше) добавляются в виде красных крестиков; число кластеров является адекватным, если значение исходных данных ниже критического значения, т. е. Q не значительно больше, чем для искусственных наборов данных, созданных на основе подобранной модели. Используя penx, упорядоченные номера кластеров от простейших до наименее простых также могут быть указаны на графике, где просто определяется как количество кластеров плюс расчетная доля шума, разделенная шумовым штрафом, см. выше. Выбранное количество кластеров является наиболее простым и адекватным, т. Е. что предпочтительны небольшое количество кластеров и низкая доля шума.

Если plot = "noise", доля шума (черный) и простота (красный) наносятся на график против числа-бер кластеров.

Значение

otrimlesimg возвращает список типа «otrimlesimgdens», содержащий компоненты result, simresult, simruns.

результат	выходной объект otrimleg (список результатов по исходным данным) запускается с параметром ters, предоставленные otrimlesimg.
simresult	список симуляций длины выходных объектов отримлега для всех имитируемых артефактов. сиальные наборы данных.
симруны	входной параметр simruns.

summary.otrimlesimgdens возвращает список типа «summary.otrimlesimgdens» с компонентами. G, simeval, ssimruns, npr, nprdiff, logicd, denscrit, peng, penorder, bestG, sdcutoff, bestresult, cluster. симруны

грамм	otrmlesimg входной параметр G (количество кластеров).
Симевал	список с компонентами denscrit, meandens, sddens, standens, errors, определены ниже.
ssimruns	otrmlesimg входной параметр simruns.
энергетический ядерный реактор	вектор расчетных долей шума на исходных данных для всех чисел кластеры, экспропорции [1] от отримле .

nprdiff	вектор для всех количеств кластеров различий между оцененными наименьшими кластерами-пропорция ter и доля шума на исходных данных.
логичный	вектор журналов неправильных значений постоянной плотности на исходных данных для всех количество кластеров.
denscrit	вектор по всем числам кластеров статистики качества кластера на основе плотности Q на исходных данных, предоставленных компонентом меры ядра kerndensmeasure.
Пэн	вектор значений простоты (см. Подробности) по всем номерам кластеров.
наказание	простота порядка количества кластеров.
bestG	наилучшее (т.е. наиболее простое адекватное) количество кластеров.
sdcutoff	входной параметр sdcutoff.
результат	вывод отримле для лучшего числа кластеров bestG.
кластер	вектор кластеризации для наилучшего числа кластеров bestG. 0 соответствует шуму / выбросам.

Компоненты simeval компонента вывода summary.otrimlesimgdens:

denscritmatrix	максимальное количество кластеров, умноженное на simruns матрицу denscrit-векторов для всех кластеризация на смоделированных данных.
извилистый	вектор по количеству кластеров робастной оценки среднего значения плотности над смоделированными наборами данных, вычисленными с помощью scaleTau2 .
sddens	вектор по количеству кластеров робастной оценки стандартного отклонения denscrit по смоделированным наборам данных, вычисленный с помощью scaleTau2.
Standens	вектор по количеству кластеров плотности исходных данных, стандартизированных по меанденс и сденс.
ошибки	вектор по количеству кластеров, количество раз, которое отримле привело к ошибка.
	plot.summary.otrimlesimgdens вернет результат номинальной() перед любым-
	вещь была изменена функцией сюжета.

Авторы)

Кристиан Хенниг <christian.hennig@unibo.it> https://www.unibo.it/sitoweb/christian . hennig / en /

Рекомендации

Коретто, П. и К. Хенниг (2016). Надежная неправильная максимальная вероятность: настройка, вычисление и сравнение с другими методами робастной гауссовой кластеризации. Журнал американской статистики Ассоциация, Vol. 111 (516), стр. 1648–1659. DOI: 10.1080 / 01621459.2015.1100996

П. Коретто и К. Хенниг (2017). Последовательность, отказоустойчивость и алгоритмы для надежной неправильная кластеризация максимального правдоподобия. Журнал исследований в области машинного обучения, Vol. 18 (142), С. 1-39. https://jmlr.org/papers/v18/16-382.html

Хенниг, К. и П. Коретто (2021 г.). Подход адекватности для определения количества кластеров для OTRIMLE кластеризация на основе робастной смеси Гаусса. Появиться в Австралии и Новой Зеландии Статистический журнал, https://arxiv.org/abs/2009.00921.

Смотрите также

otrimle , rimle , otrimleg , kerndensmeasure

Примеры

```
## otrimlesimg требует больших ресурсов компьютера, поэтому только небольшая часть данных
## используется для скорости.
данные (банкнота)
selectdata <- c (1: 30,101: 110,117: 136,160: 161)
set.seed (555566)
x <- банкнота [selectdata, 5: 7]

## simruns = 2 выбрано для скорости. На практике это не рекомендуется.
obanknote <- otrimlesimg (x, G = 1: 2, multicore = FALSE, simruns = 2, monitor = 0)
sobanknote <- резюме (obanknote)
печать (банкнота)
сюжет (банкнота, сюжет = "критерий", penx = 1,4)
сюжет (sobanknote, plot = "noise", penx = 1.4)
plot (x, col = sobanknote $ cluster + 1, pch = c ("N", "1", "2") [sobanknote $ cluster + 1])
```

сюжет. Методы построения для объектов OTRIMLE

Описание

Постройте надежные результаты кластеризации на основе модели: диаграмма рассеяния с информацией о кластеризации, оптимизация профилирование и кластерная подгонка.

Применение

```
## S3 метод для класса otrimle
plot (x, what = c («критерий», «илоглик», «соответствие», «кластеризация»),
      данные = NULL, поля = NULL, кластер = NULL, ...)
```


Аргументы

Икс	Вывод из отримле
какие	Тип графика. Это может быть одно из следующих значений: «критерий» (по умолчанию), «илоглик», «подгонка», «кластеризация». Смотрите подробности.
данные	Вектор данных, матрица или data.frame (или какое-то их преобразование), используемый для получение отримле объекта. Это актуально, только если what = "clustering".
поля	Вектор целых чисел, обозначающий переменные (количество столбцов данных), которые должны быть используется для парного графика, если what = "кластеризация". Когда поля = NULL, устанавливается значение 1: ncol (данные) (по умолчанию).
кластер	Целое число, обозначающее кластер, для которого возвращается подходящий график. Это только актуально, если что = "подходит".
...	дальнейшие аргументы, переданные другим методам или от них.

Значение

Если что = "критерий" График с профилированием критерия оптимизации OTRIMLE. Критерий at log (icd) = - всегда отображается Inf.

If what = "iloglik" График с профилированием неправильной функции логарифма правдоподобия вдоль путь поиска для оптимизации OTRIMLE.

If what = "fit" График PP (график вероятности-вероятности) взвешенного эмпирического распределения функция расстояний Махаланобиса наблюдений от центров скоплений относительно цели получить распространение. Целевое распределение - это распределение хи-квадрат со степенями свободы

dom равно ncol (данные). Веса задаются неправильными апостериорными вероятностями. Если cluster = NULL PP-графики создаются для всех кластеров, в противном случае кластер выбирает один PP сюжет временами.

If what = "clustering" Парная диаграмма рассеяния с членством в кластерах. Баллы, присвоенные Компонент шум / выбросы обозначен знаком +.

Рекомендации

Коретто, П. и К. Хенниг (2016). Надежная неправильная максимальная вероятность: настройка, вычисление и сравнение с другими методами робастной гауссовой кластеризации. Журнал американской статистики Ассоциация, Vol. 111 (516), стр. 1648–1659. DOI: 10.1080 / 01621459.2015.1100996

П. Коретто и К. Хенниг (2017). Последовательность, отказоустойчивость и алгоритмы для надежной неправильная кластеризация максимального правдоподобия. Журнал исследований в области машинного обучения, Vol. 18 (142), С. 1-39. https://jmlr.org/papers/v18/16-382.html

Авторы)

Пьетро Коретто <pcoretto@unisa.it> https://pietro-coretto.github.io

Смотрите также

сюжет.

Примеры

```
## Загрузить данные о швейцарских банкнотах
данные (банкнота)
x <- банкнота [, - 1]

## Выполните отримле кластеризацию на небольшой сетке логических значений
a <- otrimle (data = x, G = 2, logicd = c (-Inf, -50, -10), ncores = 1)
печать (а)

## Кластеризация графиков
plot (a, data = x, what = "кластеризация")

## Постройте кластеризацию по выбранным полям
plot (a, data = x, what = "clustering", margins = 4: 6)

## Постройте кластеризацию по первым двум основным компонентам
z <- масштаб (x)% *% собственные (кор (x), симметричный = ИСТИНА) $ векторов
colnames (z) <- paste ("ПК", 1: ncol (z), sep = "")
plot (a, data = z, what = "кластеризация", поля = 1: 2)

## Постройте профилирование критерия OTRIMLE
сюжет (a, what = "критерий")

## Постройте неправильное профилирование логарифмической вероятности
сюжет (a, what = "iloglik")

## Подобрать график для всех кластеров
сюжет (a, what = "fit")

## Подобрать график для кластера 1
plot (a, what = "fit", cluster = 1)



## Не работать:
## Выполните тот же пример, используя более тонкую сетку по умолчанию logicd
## значения с использованием нескольких ядер
```

```
##
a <- otrimle (data = x, G = 2)

## Проверьте критерий otrimle-vs-logicd
сюжет (a, что = критерий)

## Минимум достигается при $ logicd = -9, и исследуя $ оптимизацию, это
## видно, что интервал [-12,5, -4] ограничивает оптимальную
## решение. Мы ищем с более мелкой сеткой, расположенной около минимума
##
b <- otrimle (данные = x, G = 2, logicd = seq (-12,5, -4, length.out = 25))

## Проверьте критерий otrimle-vs-logicd
сюжет (б, что = критерий)
```



```
## Проверьте разницу между двумя кластерами
таблица (A = a $ cluster, B = b $ cluster)

## Проверить различия в оценочных параметрах
##
colSums (abs (a $ mean - b $ mean))                    ## Расстояние L1 для средних векторов
применить ({a $ cov-b $ cov}, 3, norm, type = "F")     ## Расстояние Фробениуса для ковариаций
c (Шум = abs (a $ npr-b $ npr), abs (a $ cpr-b $ cpr)) ## Абсолютная разница в пропорциях

## Конец (не запускается)
```

plot.rimle Методы сюжета для объектов RIMLE

Описание

Постройте надежные результаты кластеризации на основе модели: диаграмму разброса с информацией о кластеризации и подбором кластеров.

Применение

```
## Метод S3 для обода класса
plot (x, what = c ("соответствие", "кластеризация"),
      данные = NULL, поля = NULL, кластер = NULL, ...)
```

Аргументы

Икс	Выход из обода
какие	Тип графика. Это может быть одно из следующих значений: «подходит» (по умолчанию), «кластеризация». Смотрите подробности.
данные	Вектор данных, матрица или data.frame (или какое-то их преобразование), используемый для получение обода объекта. Это актуально, только если what = "clustering".
поля	Вектор целых чисел, обозначающий переменные (количество столбцов данных), которые должны быть используется для парного графика, если what = "кластеризация". Когда поля = NULL, устанавливается значение 1: ncol (данные) (по умолчанию).
кластер	Целое число, обозначающее кластер, для которого возвращается подходящий график. Это только актуально, если что = "подходит".
...	дальнейшие аргументы, переданные другим методам или от них.

Значение

If what = "fit" График PP (график вероятности-вероятности) взвешенного эмпирического распределения функция расстояний Махаланобиса наблюдений от центров скоплений относительно цели получить распространение. Целевое распределение - это распределение хи-квадрат со степенями свободы dom равно ncol (данные). Веса задаются неправильными апостериорными вероятностями. Если

cluster = NULL PP-графики создаются для всех кластеров, в противном случае кластер выбирает один PP сюжет временами.

If what = "clustering" Парная диаграмма рассеяния с членством в кластерах. Баллы, присвоенные Компонент шум / выбросы обозначен знаком +.

Рекомендации

Коретто, П. и К. Хенниг (2016). Надежная неправильная максимальная вероятность: настройка, вычисление и сравнение с другими методами робастной гауссовой кластеризации. Журнал американской статистики Ассоциация, Vol. 111 (516), стр. 1648–1659. DOI: 10.1080 / 01621459.2015.1100996

П. Коретто и К. Хенниг (2017). Последовательность, отказоустойчивость и алгоритмы для надежной неправильная кластеризация максимального правдоподобия. Журнал исследований в области машинного обучения, Vol. 18 (142), С. 1-39. https://jmlr.org/papers/v18/16-382.html

Авторы)

Пьетро Коретто <pcoretto@unisa.it> https://pietro-coretto.github.io

Смотрите также

отримле

Примеры

```
## Загрузить данные о швейцарских банкнотах
данные (банкнота)
x <- банкнота [, - 1]

## Выполните кластеризацию римле с аргументами по умолчанию
set.seed (1)
a <- rimle (данные = x, G = 2)
печать (а)

## Кластеризация графиков
plot (a, data = x, what = "кластеризация")

## Постройте кластеризацию по выбранным полям
plot (a, data = x, what = "clustering", margins = 4: 6)

## Постройте кластеризацию по первым двум основным компонентам
z <- масштаб (x)% *% собственные (кор (x), симметричный = ИСТИНА) $ векторов
colnames (z) <- paste ("ПК", 1: ncol (z), sep = "")
plot (a, data = z, what = "кластеризация", поля = 1: 2)

## Подобрать график для всех кластеров
сюжет (а, what = "fit")

## Подобрать график для кластера 1
plot (a, what = "fit", cluster = 1)
```

26 год ободок

Описание

rimle ищет G кластеры приблизительно гауссовой формы с / без шума / выбросов. В

настройка метода, контролирующая уровень шума, является фиксированной и предоставляется пользователем или будет угадывается функцией довольно быстро и грязно (otrimle выполняет более сложные выбор на основе данных).

Применение

rimle (данные, G, начальный = NULL, logicd = NULL, npr.max = 0,5, erc = 20, iter.max = 500, tol = 1e-6)

Аргументы

данные	Числовой вектор, матрица или фрейм данных наблюдений. Строки соответствуют наблюдения и столбцы соответствуют переменным. Категориальные переменные и NA значения не допускаются.
грамм	Целое число, определяющее количество кластеров.
исходный	Целочисленный вектор, определяющий начальное назначение кластера, где 0 обозначает шум / выбросы. Если NULL (по умолчанию) инициализация выполняется с помощью InitClust.
логичный	Число log (icd), где 0 <= icd <Inf - значение неправильной константы. плотность (icd). Это настройка RIMLE для управления размером шума. Если logicd = NULL (по умолчанию), значение icd определяется на основе данных. Чистый Подгонка модели гауссовой смеси получается с помощью logicd = -Inf.
npr.max	Число в [0,1), определяющее максимальную долю шума / выбросов. Этот определяет ограничение доли шума. Если npr.max = 0 решение без шума компонент вычисляется (соответствует логикед = -Inf.
эрк	Число> = 1, указывающее максимально допустимое соотношение между внутри кластера собственные значения ковариационной матрицы. Это определяет ограничение на собственное отношение. erc = 1 обеспечивает сферические кластеры с равными ковариационными матрицами. Большой erc позволяет для больших межкластерных расхождений ковариации. Чтобы облегчить При настройке erc предлагается масштабировать столбцы данных (см. масштаб) в любое время единицы измерения различных переменных совершенно несовместимы.
iter.max	Целочисленное значение, определяющее максимальное количество итераций, разрешенных в ECM-алгоритм (см. Подробности).
тол	Критерий остановки для лежащего в основе ECM-алгоритма. Итерация ECM останавливается если два последовательных неправильных значения логарифма правдоподобия находятся в пределах допуска.

Подробности

Функция rimle вычисляет решение RIMLE с использованием алгоритма ECM, предложенного в Coretto. и Хенниг (2017).

Могут быть наборы данных, для которых функция не предоставляет решения на основе аргументов по умолчанию. ments. Это соответствует code = 0 и flag = 1 или flag = 2 на выходе (см. Раздел «Значение» ниже). Обычно это происходит, когда возникают некоторые (или все) из следующих обстоятельств: (i) log (icd) слишком большой; (ii) erc слишком велик; (iii) npr.max слишком велик; (iv) выбор начального разбиения. В этих случаях предлагается найти подходящий интервал значений icd с помощью функции otrimle . Детали Раздел otrimle предлагает несколько действий, которые необходимо предпринять, когда возникает ошибка code = 0.

Объект pi, возвращаемый функцией rimle (см. Value), соответствует вектору параметров pi
в базовой псевдомодели (1), определенной в Coretto and Hennig (2017). С logicd = -Inf
функция Римла аппроксимирует MLE для модели простой гауссовой смеси с собственным соотношением
ковариационная регуляризация, в этом случае первый элемент вектора pi устанавливается в ноль, потому что
шумовая составляющая не учитывается. В общем, для выборки iid из моделей конечной смеси
В контексте эти параметры pi определяют ожидаемые пропорции кластеров. Из-за доли шума
ограничение в RIMLE, бывают ситуации, когда это соединение может не произойти на практике. В
последнее, вероятно, произойдет, когда как logicd, так и npr.max велики. Таким образом, расчетная ожидаемая
пропорции кластеров указываются в объекте экспропорции на выходе обода, и это
вычисляется на основе неправильных апостериорных вероятностей, указанных в тау. Увидеть Коретто и Хеннига
(2017) для более подробного обсуждения этого вопроса.

Более ранняя приближенная версия алгоритма была первоначально предложена Коретто и Хеннигом.
(2016). Программное обеспечение для исходной версии алгоритма можно найти в дополнительном материале.
риалы Коретто и Хеннига (2016).

Значение

Объект S3 класса rimle. Компоненты вывода следующие:

код	Целочисленный индикатор сходимости. code = 0, если решение не найдено (см. Подробности); code = 1, если EM-алгоритм не сходится в пределах em.iter.max; code = 2 сходимость полностью достигнута.
флаг	Строка символов, содержащая один или несколько флагов, связанных с итерацией EM в оптимальный icd. flag = 1, если не удалось предотвратить численное вырождение несобственных апостериорных вероятностей (значение тау ниже). flag = 2, если принудительное исполнение ограничение доли шума не удалось по числовым причинам. flag = 3, если принудительно Ограничение на собственное отношение не удалось по числовым причинам. flag = 4, если ограничение доли шума было успешно применено хотя бы один раз. flag = 5 если ограничение на собственное отношение было успешно применено хотя бы один раз.
iter	Количество итераций, выполненных в базовом EM-алгоритме.
логичный	Значение журнала (icd).
илоглик	Значение неправдоподобной вероятности.
критерий	Значение критерия OTRIMLE.
Пи	Расчетный вектор параметров pi базовой псевдомодели (см. De- хвосты).
иметь в виду	Матрица размерности ncol (данные) x G, содержащая средние параметры каждого кластер (по столбцам).
cov	Массив размером ncol (данные) x ncol (данные) x G, содержащий ковариационную ма- матрицы каждого кластера.
тау	Матрица размерности nrow (data) x {1 + G}, где tau [i, 1 + j] - оценка (несобственная) апостериорная вероятность того, что i-е наблюдение принадлежит j-му кластеру - тер. tau [i, 1] - это оценочная (несобственная) апостериорная вероятность того, что i-е наблюдение ция относится к шумовой составляющей.
smd	Матрица размерности nrow (данные) x G, где smd [i, j] - квадрат Маха- расстояние lanobis данных [i,] от среднего [, j] согласно cov [,, j].

кластер	Вектор целых чисел, обозначающий назначения кластера для каждого наблюдения. Это 0 для наблюдения, относящиеся к шуму / выбросам.
размер	Вектор целых чисел с размерами (количеством) каждого кластера.
экспропорция	Вектор предполагаемых ожидаемых пропорций кластеров (см. Подробности).

Авторы)

Пьетро Коретто <pcoretto@unisa.it> https://pietro-coretto.github.io

Рекомендации

Коретто, П. и К. Хенниг (2016). Надежная неправильная максимальная вероятность: настройка, вычисление и сравнение с другими методами робастной гауссовой кластеризации. Журнал американской статистики Ассоциация, Vol. 111 (516), стр. 1648–1659. DOI: 10.1080 / 01621459.2015.1100996

П. Коретто и К. Хенниг (2017). Последовательность, отказоустойчивость и алгоритмы для надежной неправильная кластеризация максимального правдоподобия. Журнал исследований в области машинного обучения, Vol. 18 (142), С. 1-39. https://jmlr.org/papers/v18/16-382.html

Смотрите также

plot.rimle , InitClust , отримле,

Примеры

```
## Загрузить данные о швейцарских банкнотах
данные (банкнота)
x <- банкнота [, - 1]

## ---------------------------------------------- ----------------------------
## ПРИМЕР 1:
## Выполните RIMLE с входами по умолчанию
## ---------------------------------------------- ----------------------------
set.seed (1)
a <- rimle (данные = x, G = 2)
печать (а)

## Кластеризация графиков
plot (a, data = x, what = "кластеризация")

## PP график кластерного эмпирического взвешенного квадрата Махаланобиса
## расстояния относительно целевого распределения pchisq (, df = ncol (data))
сюжет (a, what = "fit")
plot (a, what = "fit", cluster = 1)



## ---------------------------------------------- ----------------------------
## ПРИМЕР 2:
## Сравните решения для различных вариантов логики.
## ---------------------------------------------- ----------------------------
set.seed (1)
```



```
## Случай 1: бесшумное решение, которое соответствует чистой модели гауссовой смеси
b1 <- ободок (данные = x, G = 2, logicd = -Inf)
plot (b1, data = x, what = "кластеризация")
сюжет (b1, what = "fit")

## Случай 2: низкий уровень шума
b2 <- обод (данные = x, G = 2, logicd = -100)
plot (b2, data = x, what = "кластеризация")
сюжет (b2, what = "fit")

## Случай 3: средний уровень шума
b3 <- ободок (данные = x, G = 2, logicd = -10)
plot (b3, data = x, what = "кластеризация")
сюжет (b3, what = "fit")

## Случай 3: большой уровень шума
```

```
b3 <- ободок (данные = x, G = 2, logicd = 5)
plot (b3, data = x, what = "кластеризация")
сюжет (b3, what = "fit")
```


Индекс