Статистический анализ данных для биомедицины

Конспект по методам разведочного анализа

Обновлено 13.04.2009 Автор: Administrator

Мои черновые записи по методам разведочного анализа данных.

There is a contrast between exploratory data analysis, where the aim is to allow the data to speak for themselves, and confirmatory analysis (which includes formal estimation and testing), where the form of the analysis should have been largely decided before the data were collected(!!).

Data cleaning – filling in missing values, smoothing noisy data, removing outliers and resolving inconsistencies.

Pipeline

TODO: описать схему анализа любых новых данных

1) гистограммы всех переменных. Основываясь на этом, ищем нормировку.

Если есть два класса, то удобно строить гистограмму одновременно с ROC

2) корреляции между переменными

3) распределение расстояний между всеми объектами

4) …

Трансформации данных

1. z-score : $z = \frac{(x - \bar{x})}{s}$

такая линейная трансформация хорошо работает если распределение симметрично.

Если нет – то логарифм, сигмоид, степенная функция.

2. using range:

$z = \frac{x}{\max(x) - \min(x)}$ или $z = \frac{x - \min(x)}{\max(x) - \min(x)}$

3. Sphering:

covariance matrix : $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$

transformation: $\mathbf{Z}_i = \Lambda^{-1/2} \mathbf{Q}^T (\mathbf{x}_i - \bar{\mathbf{x}})$

Columns of Q are the eigenvectors obtained from S, lambda is a diagonal matrix of corresponding eigenvalues.

Dimensionality reduction – linear methods

Principal Component Analysis

открыл Pearson, потом Hotelling

PCA основан на SVD матрицы ковариаций или корреляций
лучше использовать матрицу корреляций, т.к. можно сравнивать результаты PCA для разных анализов.
Но: принципиальные компоненты полученные по матрицам ковариаций и корреляций различаются!
Сколько компонент оставить:

1) scree plot
2) возьмем сегмент и случайно разделим на *p* частей. Упорядочим по убыванию по длинне.

Ожидаемый размер k-ой длиннейшей части равен

$g_k = \frac{1}{p} \sum_{i=k}^{p} \frac{1}{i}.$

Q: **откуда это получается? ВАЖНЕЙШАЯ ЗАДАЧА**

Если пропорция дисперсии объясняемая k-ым компонентом больше этой величины, то этот компонент оставляем

Не всегда первые компоненты PCA самые интересные.

Дискриминирующие компоненты – те где коэффициенты имеют разные знаки.

Посмотреть на переменные, у которых самые большие по абс значению коэффициенты

TODO: посмотреть
NIPALS Algorithm ("**N**onlinear **I**terative **P**artial **L**east **S**quares") – лучше, чем SVD?
Может применяться, когда есть пропущенные данные!
`nipals{ade4}`

Singular Value Decomposition

...
TODO: написать!

Factor Analysis

Общая идея: p наблюденных переменных могут быть записаны как
линейная комбинация d < p common факторов f (нескореллированных между собой??)

$$x_1 = \lambda_{11} f_1 + \ldots + \lambda_{1d} f_d + \varepsilon_1$$

...

$$x_p = \lambda_{p1} f_1 + \ldots + \lambda_{pd} f_d + \varepsilon_p$$

λ_{ij} $(i = 1, \ldots, p \text{ and } j = 1, \ldots, d)$ - factor loadings
epsilon-i - своя для каждой переменной – specific factors. Некоррелированы друг с другом.
Дисперсия ε_i - specificity of x_i.

$\lambda_{i1}^2 + \ldots + \lambda_{id}^2$ - communality (сумма loadings для i-ой переменной)

матрицу ковариаций (или корреляций) можно представить в виде

$$\mathbf{S} = \Lambda^T \Lambda + \Psi$$

где Ψ это диагональная матрица представляющая $\mathrm{E}[\mathbf{e}\mathbf{e}^T]$.

Похоже на постановку задачи регрессии. Но оценки не являются уникальными.

Хорошо чтобы побольше λ_{ij} были близки к 0 (достигается вращениями – varimax, equimax, orthomax, etc).

Цель - получить переменную с большой loading по одному фактору и маленькими loadings по другим факторам.

Связь PCA и факторного анализа:

• Both factor analysis and PCA try to represent the structure of the data set based on the covariance or correlation matrix.
Factor analysis tries to explain the off-diagonal elements, while PCA explains the variance or diagonal elements of the matrix.

• Factor analysis is typically performed using the correlation matrix.
PCA can be used with either the correlation or the covariance matrix.

• Factor analysis has a model, but PCA does not have an explicit model

• If one changes the number of PCs to keep, then the existing PCs do not change.
This is not true in factor analysis.

• PCA has a unique solution, but factor analysis does not.

• The PC scores are found in an exact manner, but the factor scores are estimates.

Independent Component Analysis

Idea: the variables we are looking at are linear combinations of independent

random non-gaussian variables; and we want to recover those variables.

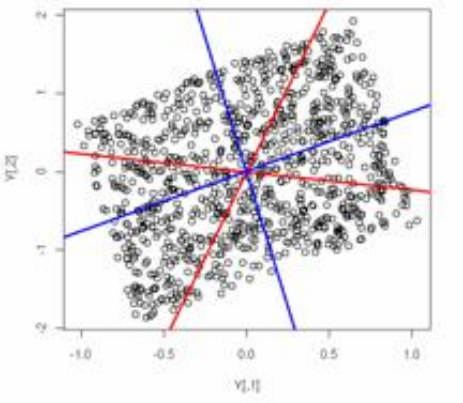


Рисунок 2 Красные линии - результат PCA, синие - ICA

The main idea behind the algorithm is the fact that (from the central limit theorem) a linear combination

of non-gaussian random variables is "more gaussian".

1. Normalize the data(sphere), so that they have a variance equal
   to 1 and that the be uncorrelated.
2. Find (with the usual numeric optimization algorithms)
   the linear transformation that maximizes the
   non-gaussianity.

One can use several measures of non-gaussianity:

- kurtosis (i.e., the fourth moment),

- entropy (integral of -f * log(f), where f is the pdf),

- mutual information, etc.

ICA – решает Blind Source Separation

X=SA,

X – матрица данных

S – independent components

A – mixing matrix

Intrinsic dimensionality

smallest number of variables needed to model the data without loss.

посмотреть testdim {ade4}

Pettis (1979)

Dimensionality reduction – nonlinear methods

Multidimensional scaling

- metric MDS – расстояния в исходном пространстве δ_{rs}
- и расстояния в уменьшенном пространстве d_{rs} связаны монотонной непрерывной функцией.

$$d_{rs} \approx f(\delta_{rs})$$

- nonmetric MDS (еще: ordinal MDS) – сохраняет только ранги расстояний ()

$$\delta_{rs} < \delta_{ab} \Rightarrow f(\delta_{rs}) \leq f(\delta_{ab})$$

Metric MDS

Aims to embed the distance directly in to the mapping domain.

введем objective function

stress = $$\sqrt{\frac{\left(\sum_r \sum_s (f(\delta_{rs}) - d_{rs})^2\right)}{\text{scale factor}}}$$

scale factor = $$\sum_r \sum_s d_{rs}^2$$

Stress минимизируется численной оптимизацией.

Classical MDS

если расстояния евклидовы, то существует решение в замкнутой форме. Задаем функцию f как 1, т.е. $d_{rs} = \delta_{rs}$

Основано на SVD матрицы расстояний, т.е. идейно очень похоже на PCA.

PCA и classical MDS дают одинаковые результаты когда используется евклидово расстояние.

cmdscale()

Sammon mapping

stress= $S_{sam} = \frac{1}{\sum_{i<j} d_{ij}^2} \sum_{i<j} \frac{(d_{ij} - \hat{d}_{ij})^2}{d_{ij}}$

d-hat – это оцененное расстояние в уменьшенном пространстве

sammon()

Metric MDS – SMACOF

ищем мажорирующую функцию, которая бы легко минимизировалась.

SMACOF – Scaling by Majorizing a Complicated Function

Nonmetric MDS

сохраняет rank of distances within the original dataset

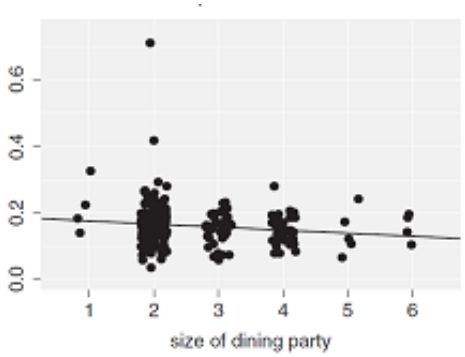
$$S_{kruskal} = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} \hat{d}_{ij}^2}}$$

isoMDS()

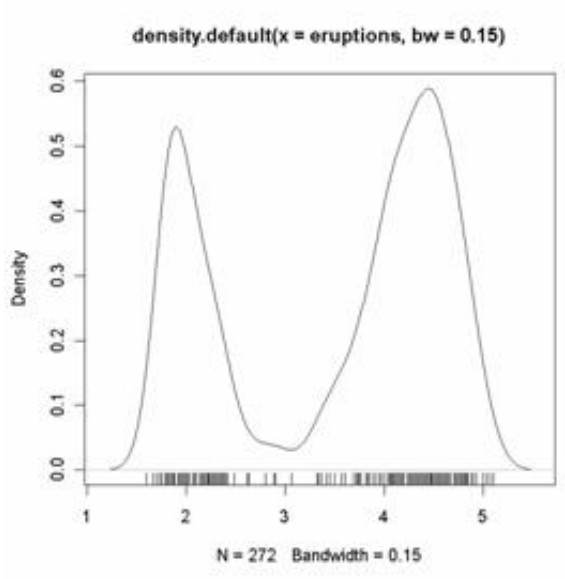
Графика

когда рисуем графики, в которых есть дискретная переменная по одной оси и вещественная по другой,

то всегда лучше трансформировать дискретную в вещественную и чуть-чуть ее jitter



rug



внизу у гистограммы добавляются настоящие данные. Функция rug() в R.

GGobi

В R любой dataframe можно наглядно визуализировать если установлен ggobi

library(rggobi)

ggobi(cbind(dat,cls))

После этого вызвать Tools->Automatic brushing и выбрать cls

Удобно: Interaction->Identify – идентифицировать точки на графике

При этом GGobi понимает свойство rownames у визуализируемого data.frame

g <- ggobi(mtcars). This creates a GGobi object called g. Getting data out isn't much harder: Just

index that GGobi object by position (g[[1]]) or by name (g[["mtcars"]] or g$mtcars). These return GGobiData objects which are linked to the data in GGobi. They act just like regular data frames, except that changes are synchronised with the data in the corresponding GGobi. You can get a static copy of the data using as.data.frame.

т.е. можно редактировать данные ggobi из R

g <- ggobi(iris)

clustering <- hclust(dist(iris[,1:4]), method="average")

glyph_colour(g[1]) <- cuttree(clustering,

A function which saves the contents of a GGobi display to a file on disk, is called

ggobi_display_save_picture. This is what we used to create the images in this document. This

creates an exact (raster) copy of the GGobi display. If you want to create publication quality graphics from GGobi, have a look at the DescribeDisplay plugin and package at http://www.ggobi.org/ describe-display. These create R versions of GGobi plots.

Scagnostics

поиск интересных пар переменных по scatterplots.

Общая идея – для каждой пары переменных оценить двумерную плотность и смотреть на характеристики получающихся контуров. Например площадь, периметр, convexity, number of connected components of those contours (может выявить multimodality!), non-linearity of the principal curves, average nearest-neighbour distance, etc.

Удобно использовать с интерактивной графикой наподобие xgobi

Определение наличия линейных зависимостей в данных

Если определитель матрицы ковариаций равен нулю, то это значит что есть линейные зависимости между переменными. Чтобы их выявить нужно смотеть на главные компоненты с нулевой дисперсией.

Unsorted

В Statistica 6.0 есть замечательный инструмент Brush для анализа многомерных данных. Иногда проще применять его вместо XGobi

дендрограмму можно использовать для выявления выбросов

хорошо использовать для гистограммы логарифмы частот - после этого намного лучше видна тонкая структура

TODO: projection pursuit – целенаправленное проектирование.

Литература

W.Martinez, A.Martinez, Exploratory Data Analysis with MATLAB

< Предыдущая Следующая >

Добавить комментарий

Имя (обязательное)

E-Mail

Осталось: 1000 символов

Подписаться на уведомления о новых комментариях

Обновить

Отправить

Из последнего

Кафедра биоинформатики (РНИМУ им Н.И.Пирогова)

Продолжается набор на годовую программу обучения в Институте биоинформатики!

Московская школа Биоинформатики - набор на новый 2015/2016 учебный год открыт

Летняя школа по биоинформатике, 20-25 июля 2015, Москва

Интенсив по геномной биоинформатике,10-12 апреля 2015, Санкт-Петербург

Онлайн-курсы от Института биоинформатики - начало 15 февраля

EuroQSAR-2014, школа-семинар по методам компьютерного конструирования лекарств.

Московский семинар по биоинформатике - 13 марта 2014

Codelobster PHP Edition - Бесплатный PHP, HTML, CSS, JavaScript редактор

Лекция «Особенности венчурного инвестирования в биотехнологические проекты в России» - 29 ноября 2013

Опрос

Какие материалы вам бы хотелось видеть на сайте?

- Биоинформатика - биология
- Биоинформатика - программирование
- Образовательные материалы по статистике
- Анализ многомерных данных
- Протеомика

Голосовать Итоги