Кластеризация

Корпусные методы исследований языковых процессов

Даша Попова

07.12.2022

Importing data from a csv file

- > data <-- read.csv(“path/filename.csv”, header = TRUE, sep=“;”)
- > data
- > print(is.data.frame(data))
- > print(ncol(data))
- > print(nrow(data))
- > d1 = subset(data, height>155)

Clustering

- deals with data sets with more than two vectors;
- data sets would list the observations in the rows, with the vectors (column variables) specifying the different properties of the observations;
- the goal is to discover structure in such data sets;
- CLUSTERING METHODS: we seek to find structure in the data in terms of grouping of observations;
- these techniques are unsupervised in the sense that we do not prescribe what groupings should be there.

Кластеризация: Метод главных компонент

- Principal Component Analysis (PCA)
- The PCA tries to reduce the number of dimensions required for locating the approximate positions of the data points.

What are principal components ?

A principal component is a normalized linear combination of the original predictors in a data set. In image above, *PC1* and *PC2* are the principal components. Let's say we have a set of predictors as $X^1, X^2..., X^p$

The principal component can be written as:

$$Z^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + ... + \Phi^{p1}X^p$$

where,

- Z^1 is first principal component
- Φ^{p1} is the loading vector comprising of loadings ($\Phi^1, \Phi^2..$) of first principal component. The loadings are constrained to a sum of square equals to 1. This is because large magnitude of loadings may lead to large variance. It also defines the direction of the principal component (Z^1) along which data varies the most. It results in a line in *p* dimensional space which is closest to the *n* observations. Closeness is measured using average squared euclidean distance.
- $X^1..X^p$ are normalized predictors. Normalized predictors have mean equals to zero and standard deviation equals to one.

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

First principal component is a linear combination of original predictor variables which captures the maximum variance in the data set. It determines the direction of highest variability in the data. Larger the variability captured in first component, larger the information captured by component. No other component can have variability higher than first principal component.

The first principal component results in a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line.

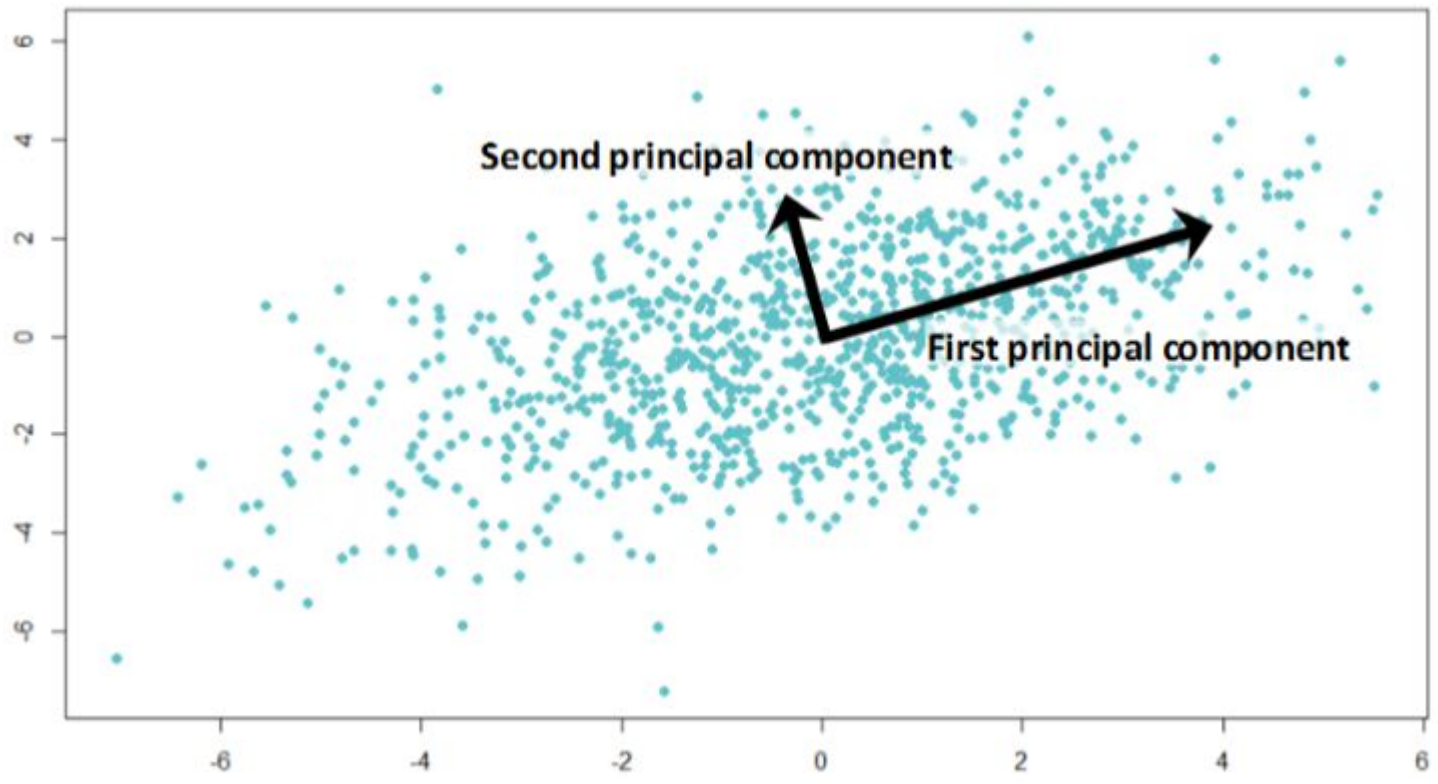
https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

Second principal component (Z^2) is also a linear combination of original predictors which captures the remaining variance in the data set and is uncorrelated with Z^1. In other words, the correlation between first and second component should is zero. It can be represented as:

$$Z^2 = \phi^{12}X^1 + \phi^{22}X^2 + \phi^{32}X^3 + ... + \phi^{p2}X^p$$

If the two components are uncorrelated, their directions should be orthogonal (image below). This image is based on a simulated data with 2 predictors. Notice the direction of the components, as expected they are orthogonal. This suggests the correlation b/w these components in zero.

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/



https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

PCA in R

- > USArrests

- > prcomp(USArrests) # inappropriate
- > prcomp(USArrests, scale = TRUE)

- > prcomp(~ Murder + Assault + Rape, data = USArrests, scale = TRUE)

- > plot(prcomp(USArrests))
- > summary(prcomp(USArrests, scale = TRUE))

- > p = summary(prcomp(USArrests, scale = TRUE))
- > props = round((p$sdev^2/sum(p$sdev^2)),3)
- > props[1:6]
- > barplot(props,col=as.numeric(props>0.05), xlab = "principal components", ylab = "proportion of variance explained")
- > abline(h=0.05, col="red")

- > biplot(prcomp(USArrests, scale = TRUE))

Задание

Проанализируйте данные из файла pca.csv методом главных компонент. Данные взяты из книги *Политический атлас современности: Опыт многомерного статистического анализа политических систем современных государств*. — М.: Изд-во «МГИМО-Университет», 2007. — 272 с. https://www.hse.ru/data/2009/12/15/1230161701/politatlas.pdf

StateshipIndex соответствует индексу государственности стран (стр. 161--163). ThreatIndex соответствует индексу внутренних и внешних угроз (стр. 164--166). InfluenceIndex соответствует индексу потенциала международного влияния (стр. 168--169). DemocracyIndex соответствует индексу институциональных основ демократии (стр. 174--175). Значения округлены.

Задание:

- код, который вводит переменную pcadata, в которой лежат все нужные данные
- код, позволяющий оценить важность каждой из компонент. Сколько компонент можно оставить для описания вариативности данных? Приведите два критерия, которые помогают ответить на этот вопрос.
- приведите столбчатую диаграмму вклада каждой из компонент в описание вариативности данных.
- приведите код, позволяющий оценить вклад каждой переменной в первую и вторую компоненты. Какие из переменных вносят наибольший вклад в первую компоненту?
- приведите график вклада переменных в первую и вторую компоненты.

Можно пользоваться информацией в соответствующих главах:

- http://www.sfs.uni-tuebingen.de/~hbaayen/publications/baayenCUPstats.pdf

Задание

1. код, который вводит переменную pcadata, в которой лежат все нужные данные
- data = read.csv("pca.csv", header=TRUE, sep =";")
2. код, позволяющий оценить важность каждой из компонент. Сколько компонент можно оставить для описания вариативности данных? Приведите два критерия, которые помогают ответить на этот вопрос.
- p = summary(prcomp(data[2:5], scale = TRUE))
- p
3. приведите столбчатую диаграмму вклада каждой из компонент в описание вариативности данных.
- props = round((p$sdev^2/sum(p$sdev^2)),3)
- barplot(props,col=as.numeric(props>0.05), xlab = "principal components", ylab = "proportion of variance explained")
- abline(h=0.05, col="red")
4. приведите код, позволяющий оценить вклад каждой переменной в первую и вторую компоненты. Какие из переменных вносят наибольший вклад в первую компоненту?
- prcomp(data[2:5], scale = TRUE)
5. приведите график вклада переменных в первую и вторую компоненты.
- biplot(prcomp(data[2:5], scale = TRUE))

Важно!

- Исходные данные должны быть сопоставимы
- Иначе, вместо самой информативной может быть выбрана самая шумная.

Рисуем дерево критериев

Категориальные данные?

Да:

критерий хи-квадрат

chisq.test

нулевая гипотеза: переменные независимы

p-value > 0.05 – принимаем нулевую гипотезу

p-value < 0.05 – отвергаем нулевую гипотезу, принимаем альтернативную: переменные зависимы

Рисуем дерево критериев

Категориальные данные?

Нет:

Нормально распределены?

Тест Шапиро-Уилка(Вилка)

shapiro.test()

Нулевая гипотеза: распределены нормально

p-value > 0.05 – принимаем нулевую гипотезу

p-value < 0.05 – отвергаем нулевую гипотезу, принимаем альтернативную: распределены не нормально

Рисуем дерево критериев

Категориальные данные?

Нет:

Нормально распределены?

Нет:

Критерий Вилкоксона

wilcox.test(x, y)

Нулевая гипотеза H_0: медиана разницы в популяции равна нулю/рейтинги статистически не различаются

p-value > 0.05 – принимаем нулевую гипотезу

p-value < 0.05 – отвергаем нулевую гипотезу, принимаем альтернативную: медиана разницы в популяции не равна нулю

Рисуем дерево критериев

Категориальные данные?

Нет:

Нормально распределены?

Да:

Критерий Стьюдента

t.test()

Нулевая гипотеза H_0: различий нет

p-value > 0.05 – принимаем нулевую гипотезу

p-value < 0.05 – отвергаем нулевую гипотезу, принимаем альтернативную: сравниваемые выборки/выборка и величина различаются

Научное знание

- Системность

А.А. Зализняк и С.П. Капица: мин. 5 – 7 и 19 – 24.30: https://www.youtube.com/watch?v=2OmVPytZbGg

- Воспроизводимость
- Верифицируемость – возможность подтвердить утверждение
- Фальсифицируемость – принципиальная возможность опровержения утверждения, опровергаемость, критерий Поппера, который предложил этот критерий в 1935г.

Визуализация: адекватность графика задаче

https://digital.infografika.agency/dataviz-test/