

1. Ненадгледано учење (unsupervised learning)

Претходно шта сте научили: предвиђају се вредности за један или више излаза, или прецизнијим речником, предвиђају се вредности зависне променљиве $Y = (Y_1, Y_2, \dots, Y_m)$ за дати скуп улаза или предиктора $X^T = (X_1, X_2, \dots, X_p)$. Нека је $x_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ улаз за i -ту обсервацију у тренинг скупу, и нека је y_i излазна вредност коју предвиђамо.

Нове предикције се заснивају на тренинг примерима $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ претходно решених случајева где су придружене вредности свих парова унапред познате. Овај процес креирања модела се назива **надгледано учење** или „учење са учитељом“.

Метафора „студент“ представља одговор \hat{y}_i за сваки улаз x_i у тренинг скупу, док „учитељ“ даје тачан одговор и грешку која је придружена сваком студентовом одговору. Ово се обично карактерише неком функцијом губитка $L(y, \hat{y})$, као на пример $L(y, \hat{y}) = (y - \hat{y})^2$.

Данас ће бити нешто више речи о **Ненадгледаном учењу** или „учењу без учитеља“. У овом случају имамо скуп обсервација (x_1, x_2, \dots, x_N) случајног p -вектора X који има густину $Pr(X)$. Циљ ненадгледаног учења је да се директно закључи о својствима функције густине без помоћи супервизора или учитеља који све време пружа тачне одговоре. *Веома често димензије ових вектора су веће него што су то вектори у случају надгледаног учења, а својства која треба открити су доста веома компликована.* Ови фактори су нешто ублажени чињеницом да X представља све променљиве које се посматрају; није потребно доносити закључке како се својства $Pr(X)$ мењају под условом да се мењају вредности другог скупа променљивих.

У мање-димензионим проблемима ($p \leq 3$), постоје различите ефективне непараметарске методе за директну процену густине $Pr(X)$ на основу свих X вредности. Имајући у виду „проклетство димензионалности“ (енг. *curse of dimensionality*), ове методе не функционишу у вишедимензионалном простору. Доста често се мора пристати на процену прилично сировог глобалног модела као што је *Gaussian Mixture* или различите дескриптивне статистике које карактеришу $Pr(X)$. Ове дескриптивне статистике покушавају да окарактеришу X вредности или колекцију таквих вредности, где је $Pr(X)$ релативно комплексна.

Неке примере ненадгледаног учења наводимо у наставку:

- **Principal components, multidimensional scaling, self-organizing maps, и principal curves**, на пример, покушавају да идентификују мање-димензионе фолдове (*manifolds*) унутар X -простора који представљају податке велике густине. Ово пружа информације о асоцијацијама међу променљивама и даје одговор да ли оне могу или не могу бити посматране као функције мањег скупа „латентних“ променљивих.
- **Кластер анализа** покушава да пронађе више конвексних региона унутар X -простора који садржи модове (*modes*) од $Pr(X)$.
- **Правила удруживања** (енг. *association rules*) покушавају да конструишу једноставне описе (коњунктивна правила) који описују регионе велике густине, у посебном случају високо-димензионих података са бинарним вредностима.

Код *надгледаног учења* постоје јасне мере успеха или неуспеха које се могу користити за процену адекватности у одређеним ситуацијама и за упоређивање ефикасности различитих метода у различитим ситуацијама. Недостатак успеха се директно мери очекиваним губитком за $Pr(X, Y)$. Све ово може бити процењено на различите начине укључујући свакако и унакрсну-валидацију.

У контексту *ненадгледаног учења*, не постоји директна мера успеха. Јако тешко је утврдити валидност закључака извученог из резултата већине алгоритама учења без надзора. Доста често се мора посегнути за хеуристичким аргументима, не само за објашњење мотивисаности зашто се неки алгоритам користи, као што је то често случај код надгледаног учења, већ и за оправдање у погледу квалитета резултата.

„Ова непријатна ситуација је довела до великог ширења предложених метода, јер је ефикасност ствар мишљења и не може се директно проверити.“ – Trevor Hastie, Robert Tibshirani, Jerome Friedman.

1.1. Кластеризација података

Кластер анализа, познатија и као **сегментација података**, има различите циљеве. Сви они се односе на груписање или сегментирање колекције објеката у подскупове или „кластере“, тако да су они унутар сваког кластера међусобно ближе повезани од објеката додељених различитим кластерима. Објекат се може описати низом мерења или односом са другим објектима. Поред тога, циљ је понекад распоредити кластере у природну хијерархију. То подразумева сукцесивно груписање самих кластера тако да су на сваком нивоу хијерархије кластери у истој групи више слични једни другима него они у различитим групама.

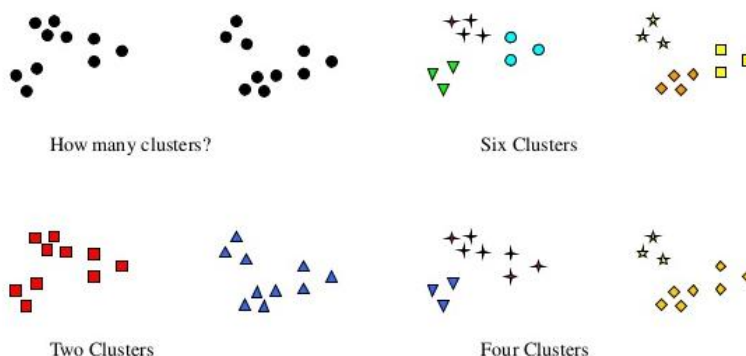


1. *Journal of the American Medical Association*, 2000; 283: 2689-2696.

На пример, у случају обраде огромног броја података, целокупно података могу бити

Поїєм кластерован з цієї єдиноманітної дефініції. Як це приклад показаний на злини 2

очекивати је да је некада потребно извршити грубље кластеровање – у мањи број кластера, а некада финије – у већи број кластера. Алгоритми кластеровања обично омогућавају подешавање нивоа грануларности, односно броја кластера који се у подацима проналази.



Слика 2. Различита кластеровања над различитим подацима.

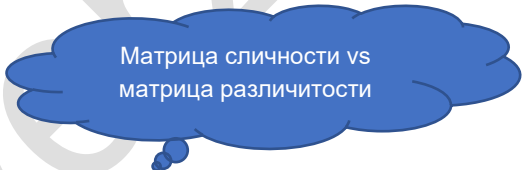
Појам кластеровања није једнозначно дефинисан не само у односу на број кластера који се у подацима могу наћи, већ и у односу на идеју шта једну групу тачака чини кластером. У односу на то, постоји више неформалних дефиниција кластеровања. **Глобуларни** или **центрични кластери** су групе тачака које попуњавају унутрашњост лопте или евентуално елипсоида. Добро раздвојени кластери су групе тачака које су ближе другим тачкама своје групе него било којој тачки из неке друге групе. **Густински кластери** су кластери чије су тачке раздвојене од тачака других кластера регионима мање густине. **Хијерархијски кластери** су или појединачне тачке или кластери чије су тачке такође организоване у структуру хијерархијских кластера.

За све циљеве кластер анализе најважнији је појам **степенa сличности (или различитости)** између појединих објеката који се кластерују. Метода кластеровања покушава да групише објекте на основу дефиниције сличности која му је дата. Ово може произаћи само из разматрања објеката са којима се ради. Ситуација је донекле слична спецификацији функције губитка или трошкова у проблемима предвиђања (надгледано учење). Тамо трошкови (енг. *cost*) повезани са нетачним предвиђањем зависе од разматрања обсервација изван података.

1.1.1. Матрице сличности (*Proximity matrices*)

Подаци су понекад представљени директно у смислу близине (сличности или афинитета) између парова објеката. То могу бити било њихове сличности или разлике (разлика или недостатак афинитета). На пример, у експериментима из друштвених наука, од учесника се тражи да просуде по томе колико се одређени објекти међусобно разликују. Тада се разлике могу израчунати рачунањем просека тако прикупљених пресуда. **Узмите било коју слику и покушајте да обележите регионе које видите. Ако покуша још неко да уради исто, видећете да постоји извесна разлика у обележеним регионима. Шта је онда истина (енг. ground-truth)?**

Овај тип података може да се представи $N \times N$ матрицом D , где је N број објеката, а сваки елемент $d_{ii'}$ представља близину између објеката i и i' . Ова матрица представља улаз у алгоритам за кластеризацију.



Матрица сличности vs
матрица различитости

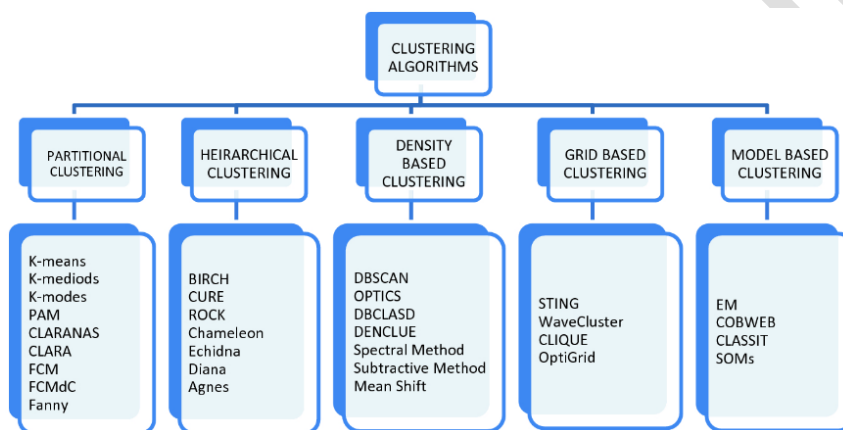
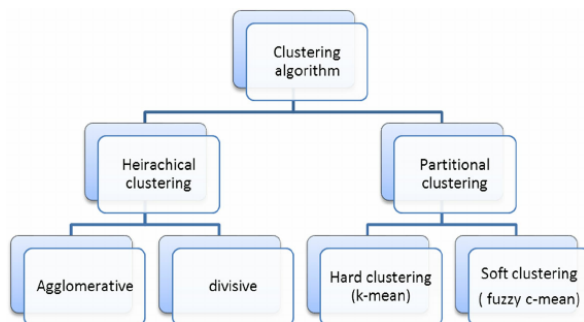
Већина алгоритама претпоставља да се на улазу добија *матрица различитости* са ненегативним елементима и нула елементима на дијагонали: $d_{ii} = 0, i = 1, 2, \dots, N$. **Ако су изворни подаци прикупљени као сличности, може се користити одговарајућа монотono опадајућа функција за њихово претварање у различитост (погледајте пример за Фејсбук где је узета функција $f(x) = \frac{1}{x}$). Такође, већина алгоритама подразумева коришћење симетричне матрице различитости, па ако оригинална матрица D није симетрична, тада се она мора заменити са $(D + D^T)/2$.**

Субјективно процењене различитости ретко су када удаљеност у строгом смислу (дистанца), јер неједнакост троугла $d_{ii'} \leq d_{ik} + d_{i'k}, \forall k \in \{1, 2, \dots, N\}$ не важи. Стога се неки алгоритми који претпостављају **растојања** не могу користити са таквим подацима.

1.1.2. Различитост заснована на атрибутима

Овај део остаје за следећи термин...

1.1.3. Алгоритми за кластеризацију



Overview

- **Partitioning methods:** Given n objects, these methods construct k partitions of the data, by assigning objects to groups, with each partition representing a cluster. Generally, each cluster must contain at least one object; and each object may belong to one and only one cluster, although this can be relaxed
 - Partitioning algorithms do not consider all partitions and can only find local optima
- **Hierarchical methods:** Such methods are said to provide multi-resolution clustering, as a range of clusters at different levels of similarity are returned by the algorithm
 - Create a hierarchical decomposition of the objects by either merging or splitting clusters sequentially
- **Model-based methods:** Formulate a model and fit it to the data by estimating suitable parameters. The models are typically statistical mixture models or neural networks
- **Density-based methods:** Clusters are defined as dense regions in the data space, i.e. a larger than expected number of points in a given subspace.
- **Grid-based methods:** Characterized by the practice of dividing the data space into a finite number of cells to form a grid. All clustering operations are then performed on the cells of this grid.

Petnica, jul 15

4

Many different method and algorithms:

- a. For numeric and/or symbolic data
- b. Deterministic vs. probabilistic
 - i. In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
 - ii. Weights must sum to 1
 - iii. Probabilistic clustering has similar characteristics
- c. Exclusive vs. overlapping
 - i. In non-exclusive clustering, points may belong to multiple clusters.
 - ii. Can represent multiple classes or 'border' point
- d. Hierarchical vs. flat
- e. Top-down vs. bottom-up

Алгоритми партиционисања**K-means**

Алгоритам **k-средина** проналази k кластера у подацима које представља помоћу k центроида тих кластера, од којих се свака добија упросечавањем елемената датог кластера. Ова претпоставка чини алгоритам применљивим само на податке који се могу упросечавати, попут вектора. Под одређеним условима, постоје уопштења алгоритма и на другачије врсте података.

Алгоритам:

Полазних k центроида се бира насумично (мада, ако корисник зна нешто о структури својих података, могу бити и унапред дати), а потом се понављају следећи кораци:

1. распоредити све инстанце у нове кластере тако што се свака инстанца придружи најближој центроиду
2. израчунати нове центроиде као просек инстанци које су им придружене.

Ови кораци се извршавају све док се центроиде мењају. Када су центроиде исте у две узастопне итерације, алгоритам се зауставља.

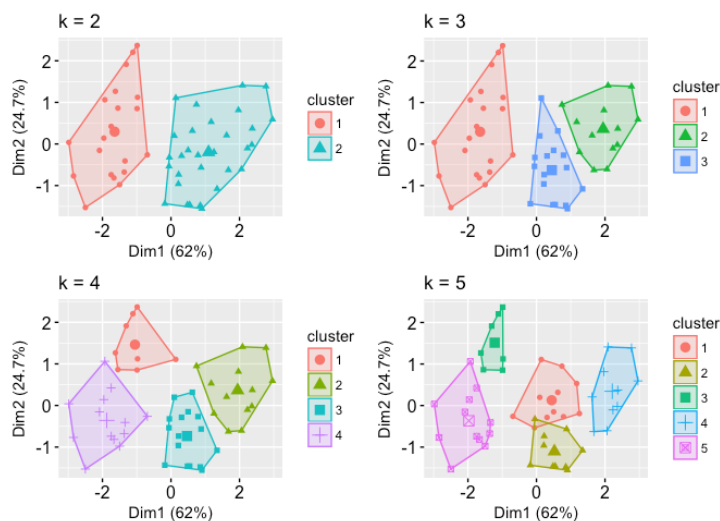
Може се показати да овај алгоритам минимизује функцију

$$\sum_{i=1}^k \sum_{x \in C_i} d(x, c_i)^2$$

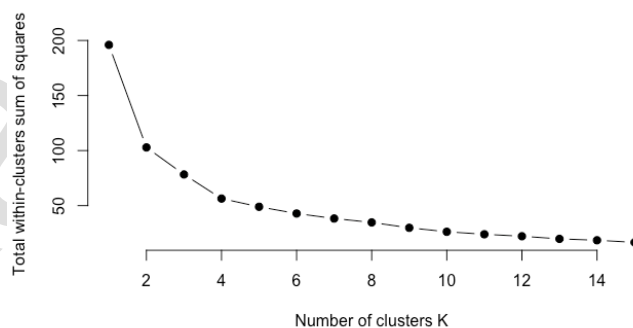
по C_i , где је d еуклидско растојање (али је могуће користити и неко друго). Уопштење математичког записа можете видети у књизи „*The Elements of Statistical Learning*“. На основу овога се може нешто закључити и о његовом понашању. Захваљујући томе што је заснован на минимизацији еуклидског растојања, алгоритам тежи проналажењу кластера у облику лопте. Како је растојање квадрирано, алгоритам је осетљив на податке који значајно одударају од осталих. У том случају ће веће растојање утицати на укупну грешку непропорционално у односу на остала растојања и таква тачка ће непропорционално утицати на локацију центроиде. Такође, ако густина тачака не варира драстично и растојања међу кластерима нису велика, алгоритам преферира кластере са сличним бројем тачака у њима, пошто би у том случају бројан кластер морао садржати и тачке далеко од центроиде које би значајно повећавале суму квадрата растојања.

Чињеница да алгоритам k средина минимизује наведену суму наводи на њену даљу анализу. Битно је питање да ли она има један глобални минимум, односно да ли је најбоље кластеровање у односу на дату суму квадрата растојања јединствено. [Одговор на прво питање је негативан.](#) Могуће је да постоји већи број кластеровања једнаког квалитета. Један пример у којем би то било и интуитивно је када су тачке униформно распоређене унутар круга и потребно их је поделити на два кластера. Ротирање добијених центроида у односу на центар круга даје подједнако добро кластеровање. Другим речима, у случају таквог скупа података, постоји пуно глобалних, и самим тим подједнако добрих, минимума. Таква ситуација није забрињавајућа.

Ипак, испоставља се да могу постојати и локални минимуми слабијег квалитета од глобалног и да алгоритам може наћи такав минимум, што није добро. Овај проблем се ублажава тако што се кластеровање покреће већи број пута са различитим иницијалним тачкама и за резултат се узима кластеровање најмање вредности суме квадрата растојања.



Алгоритам k средина омогућава флексибилност при проналажењу кластера кроз могућност подешавања броја k . Ипак, у пракси често није јасно како изабрати број k и поменута флексибилност често води недоумици. Једно хеуристичко правило је „правило лакта“ које сугерише да се за различите вредности броја k изврши кластеровање, да се нацрта график зависности суме квадрата растојања у зависности од k и да се за изабере кластеровање које одговара броју k који лежи на тачки нагле промене брзине опадања графика или на његовом „лакту“. Оваква ситуација је приказана на слици 2. Интуитивно образложење је да су након „лакта“ кластери већ хомогени и додавање нових центроида не доприноси значајно смањењу суме квадрата растојања.



Слика 2. Правило лакта.

Предности и мане алгорита

(Dis)advantages

- Advantages:
 - **scalable**: the time complexity is $O(n)$ ($O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$), it can work with any L_p
 - summaries for the clusters are available directly from the algorithm
 - **robust** to the order of the data
- Disadvantages:
 - having to provide the number of clusters k ; common practice is to **use a hierarchical clustering** method to suggest a suitable k
 - the method being sensitive to outliers (unable to handle noisy data)
 - tending to find **spherical clusters** of equal size
 - having to **find initial centroids** to start the algorithm. Hartigan and Wong (1979) suggest using actual objects as initial cluster centres (most often selected **randomly**)
 - having to convert objects to the distance-space (increase the computational cost)
 - **applicable only when mean is defined**

Литература:

- T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning
- Делови текста су преузети из књиге: Младен Николић, Анђелка Зечевић. Машинско учење. Скрипта. Природно-математички факултет у Београду, 2019.