دوره آموزشی داده کاوی کاربردی

محمدامین ملاحسینی اردکانی | mollahoseini.ir

خوش آمدید

mollahoseini.ir | محمدامین ملاحسینی اردکانی

قُلْ هَلْ يَسْتَوِي الَّذِينَ يَعْلَمُونَ وَالَّذِينَ لَا يَعْلَمُونَ

بگو: آیا کسانی که می‌دانند با کسانی که نمی‌دانند برابرند؟

سورهٔ زمر آیه ۹

محمدامین ملاحسینی اردکانی

www.mollahoseini.ir
m.amin@mollahoseini.ir

بینا
گروه تحلیل اطلاعات مالی

هم بنیان گذار گروه تحلیل اطلاعات مالی بینا

گسترش فناوری‌های نوین
شرکت سهامی خاص
Hi-Tech Solutions Co.

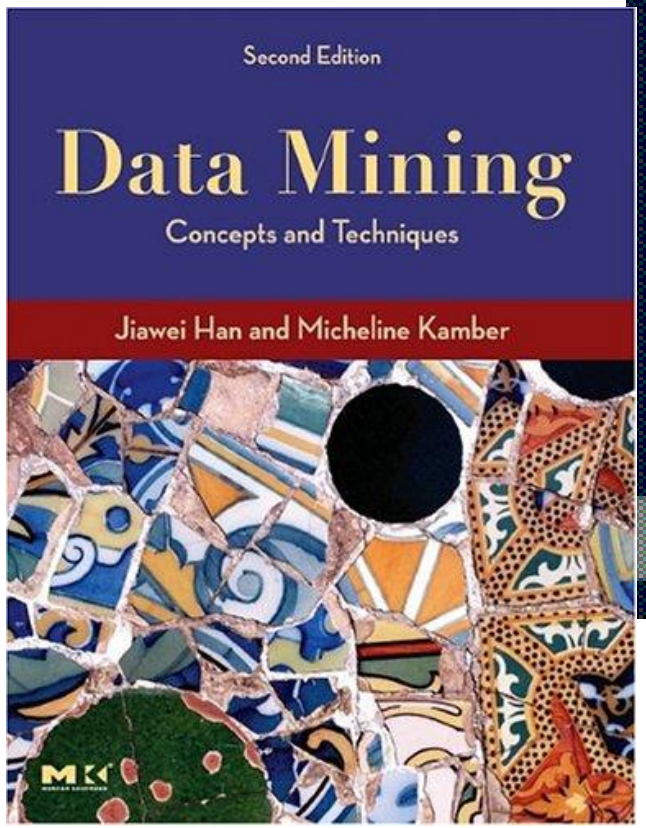
بانک کشاورزی

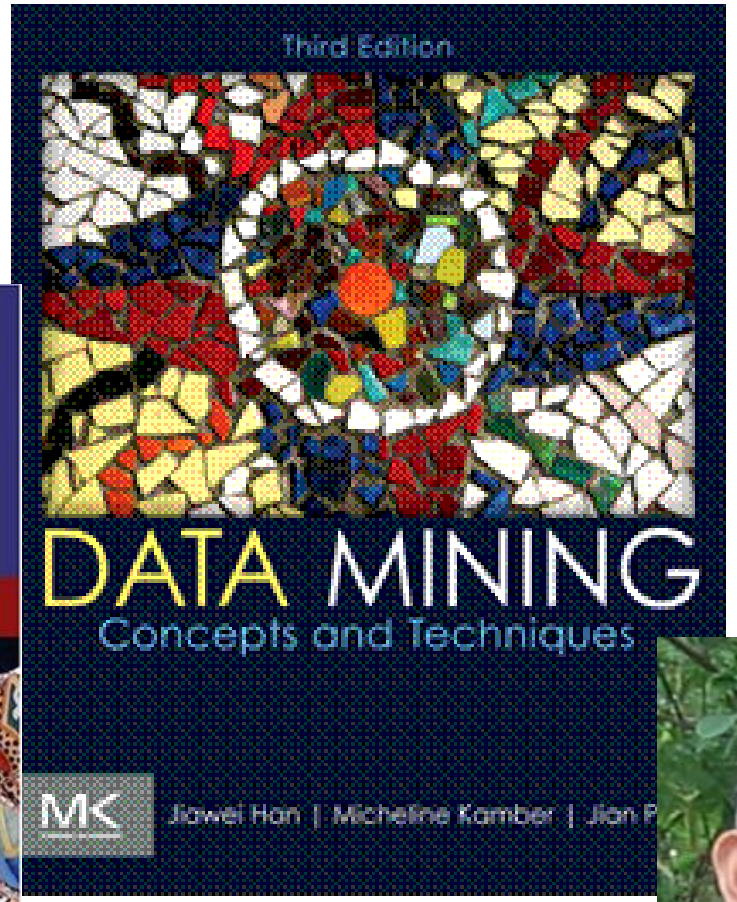
کارشناس کلان داده در شرکت گسترش فناوری‌های نوین بانک کشاورزی

کانور
KANOOR

هم بنیان گذار گروه آموزشی کانور

منابع دوره


Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber


Third Edition
DATA MINING
Concepts and Techniques
MK
Jiawei Han | Micheline Kamber | Jian P

سرفصل‌ها

دوره آموزشی داده کاوی کاربردی

مفاهیم و کاربردها

تعریف داده کاوی . کاربردهای داده کاوی در کسب و کار . مزایای داده کاوی . فرآیند داده کاوی . عومل موثر . انواع داده‌ها . تسک‌های اساسی . چالش‌ها

آماده سازی و پیش پردازش

پاک سازی داده‌ها . تجمیع داده‌ها . تبدیل داده ها . کاهش داده‌ها . گسسته سازی داده‌ها

تکنیک‌های داده کاوی

کشف الگو . طبقه بندی . خوشه‌بندی . پیش بینی

اجرای پروژه داده کاوی

انتخاب مسئله . جمع آوری داده ها . آماده سازی داده ها . انتخاب تکنیک داده کاوی . ارزیابی مدل

- تعریف داده کاوی
- کاربردهای داده کاوی در کسب و کار
- مزایای داده کاوی
- فرآیند داده کاوی
- عوامل موثر در داده کاوی
- انواع داده‌ها
- تسک‌های اساسی در داده کاوی
- چالش‌ها

چرا داده کاوی؟

داده کاوی چیست؟

کاربردهای داده کاوی در کسب و کار

بازارهای مالی

پیش‌بینی قیمت
شناسایی فرصت‌های سرمایه‌گذاری
کشف تقلب
مدیریت ریسک

منابع انسانی

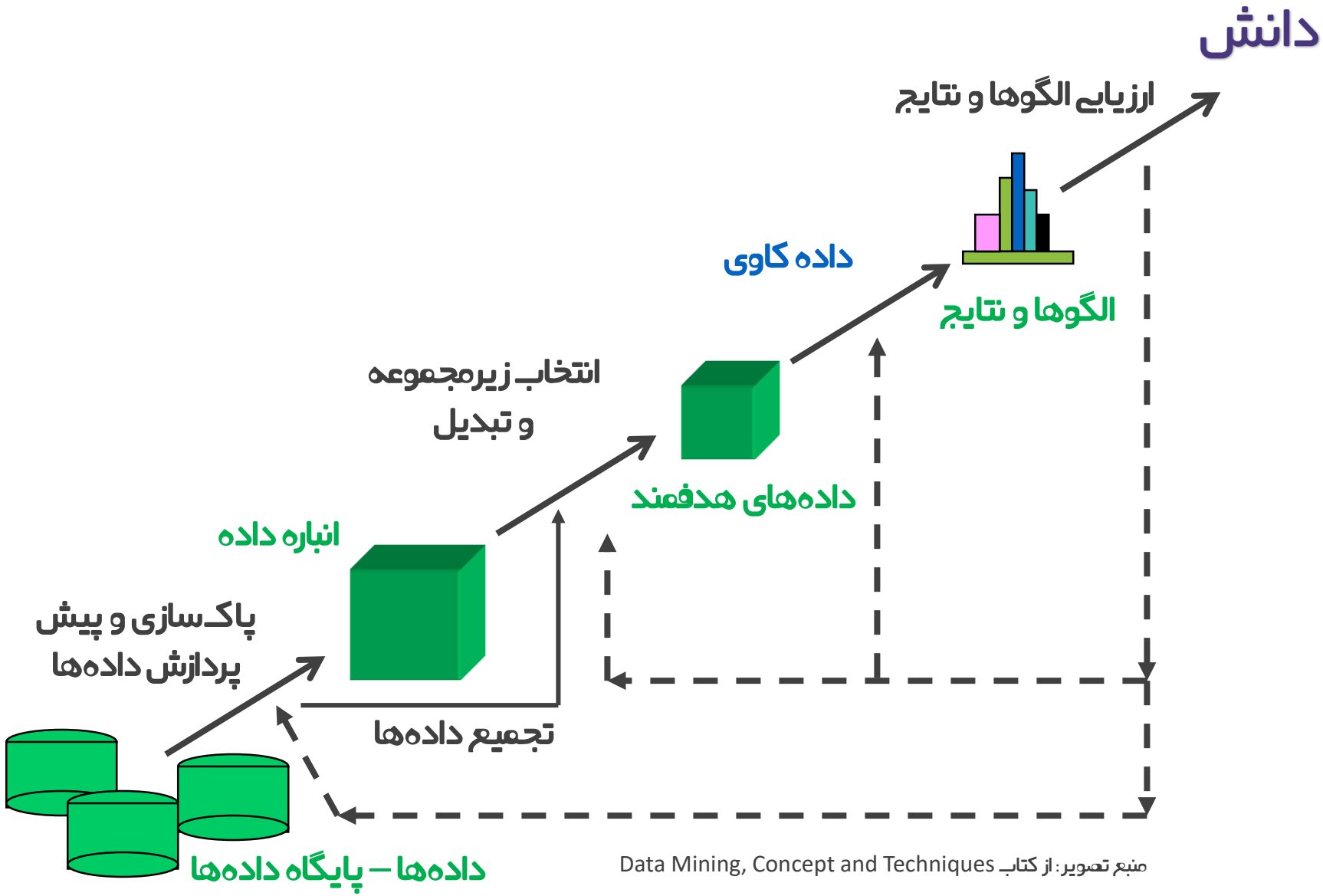
پیش‌بینی عملکرد کارکنان
شناسایی استعدادها
جذب و استخدام کارکنان
ارزیابی عملکرد کارکنان

ت. صنایع غذایی

پیش‌بینی تقاضا
شناسایی فرصت‌های بازاریابی
بهبود کیفیت محصولات غذایی
کاهش هزینه‌های تولید

هتل و هتل‌داری

بررسی رفتار مشتریان
شناسایی فرصت‌های بازاریابی
بهبود تجربه مشتری
کاهش هزینه‌های عملیاتی



منبع تصویر: از کتاب Data Mining, Concept and Techniques

مزایای داده کاوی

بهبود تصمیم گیری
افزایش بهره وری
کاهش هزینه‌ها
افزایش رضایت

فرآیند داده کاوی

عوامل موثر در داده کاوی

چه داده‌ای؟ تکنولوژی؟
نیازمندی؟ کاربرد؟

انواع داده‌ها

داده‌های رابطه‌ای
داده‌های تراکنشی
داده‌های جریانی
داده‌های سری زمانی
داده‌های زمانی

داده‌های ترتیبی
داده‌های گرافی
داده‌های جغرافیایی
داده‌های زمانی مکانی

تسک‌های اساسی در داده کاوی

- ✓ توصیف داده‌ها
- ✓ پیدا کردن پترن‌ها، همبستگی‌ها و وابستگی‌ها
- ✓ رده‌بندی
- ✓ خوشه‌بندی
- ✓ داده‌های پَرت

چالش‌ها

- حجم بالا (توزیع و موازی سازی – مقیاس پذیری – افزایشی)
- واقعی هستند
- ارزیابی نتایج
- دانش خارجی
- حریم خصوصی

- ✓ تعریف داده کاوی
- ✓ کاربردهای داده کاوی در کسب و کار
- ✓ مزایای داده کاوی
- ✓ فرآیند داده کاوی
- ✓ عوامل موثر در داده کاوی
- ✓ انواع داده‌ها
- ✓ تسک‌های اساسی در داده کاوی
- ✓ چالش‌ها

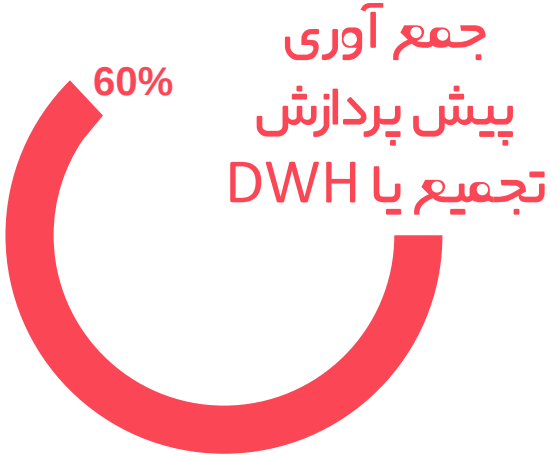
آماده سازی و پیش پردازش

- پاک‌سازی داده‌ها
- تجمیع و یکپارچه سازی داده‌ها
- تبدیل داده‌ها
- کاهش داده‌ها
- گسسته سازی داده‌ها

چرا نیاز به پیش پردازش داریم؟

منبع تصویر: ابزارهای هوش مصنوعی

- کامل نباشند
- نویزی باشند
- متناقض باشند
- تکراری باشند

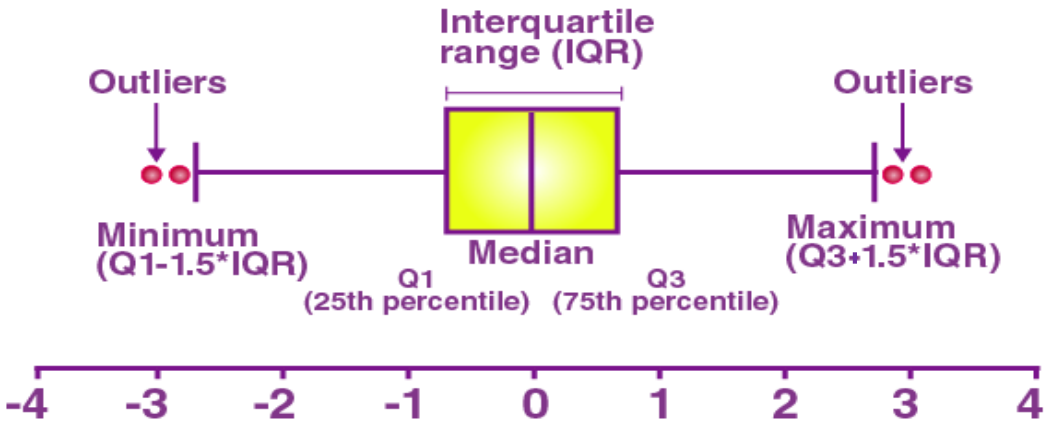


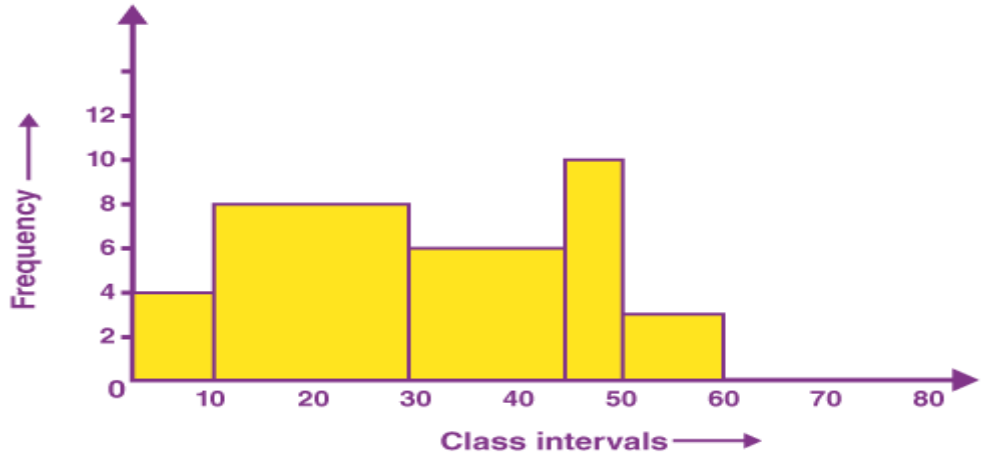
عمل‌های اصلی در پیش پردازش

۱- بررسی و توصیف
۲- پاک‌سازی
۳- تجمیع و یکپارچه‌سازی
۴- تبدیل
۵- کاهش
۶- گسسته سازی

بررسی و توصیف داده‌ها

- ✓ کمینه، بیشینه، میانگین، مد، میانه و ...
- ✓ نمودار جعبه یا Box plot
- ✓ نمودارهای مختلف مانند هیستوگرام (Histogram)





منبع تصاویر: https://byjus.com

پاک‌سازی داده‌ها

منبع تصویر: https://medium.com

با داده‌های ناکامل چیکار کنیم؟

با داده‌های نویزی چطور؟

- Binning – Equal width
- Binning – Equal depth
- Regression
- Clustering

تجمیع داده‌ها

- یکپارچه‌سازی اسکیما
- مقادیر مختلف از یک موجودیت
- تناقض‌ها
- افزونگی داده‌ها
- همبستگی و وابستگی (Pearson، Chi-squared و غیره)

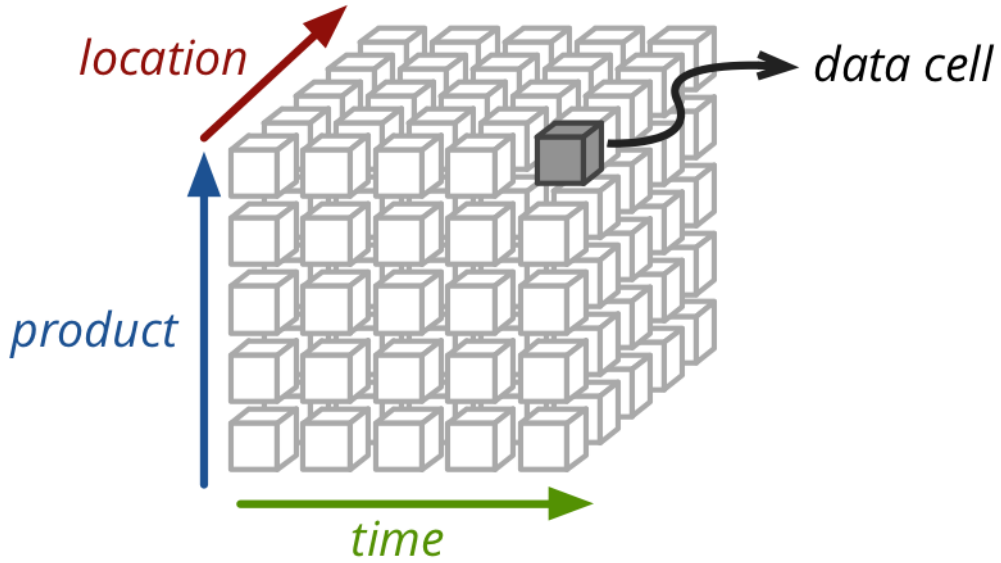
تبدیل داده‌ها

- تجمیع
- ایجاد سلسله مراتب
- نرمال کردن
- تبدیل ویژگی

کاهش داده‌ها

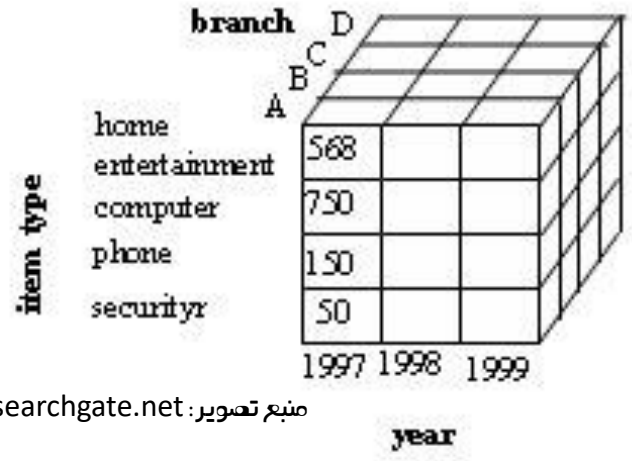
- معکب داده‌ای
- انتخاب زیر مجموعه از ویژگی‌ها
- فشرده سازی
- نمونه برداری

منبع تصویر: http://cubes.databrewery.org



گسسته سازی داده‌ها

- به کمک خوشه بندی
- با استفاده از بی نظمی



منبع تصویر: http://researchgate.net

- ✓ پاک‌سازی داده‌ها
- ✓ تجمیع و یکپارچه سازی داده‌ها
- ✓ تبدیل داده‌ها
- ✓ کاهش داده‌ها
- ✓ گسسته سازی داده‌ها

تکنیک‌های داده کاوی

- کشف الگو
- رده بندی و پیش بینی
- خوشه بندی

کشف الگو یعنی ؟

- کدام محصولات باهم خریداری شده اند؟
- ترتیب به چه صورتی بوده تا به آیتم بعدی تخفیف ارائه نماییم؟

itemset و k-itemset

مجموعه آیتم های مکرر و قوانین انجمنی

- مایع ظرفشویی، دستکش
- لپ تاپ، نرم افزار و ویندوز، پرینتر

هدف: یافتن قوانین A → B

- شرط min support
- شرط min confidence

support: هم شامل A و هم شامل B

confidence: اگر شامل B باشد حتما A را هم داشته باشد

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{support(A,B)}{P(A)}$$

مثال

min support = min confidence = **50%**

شماره تراکنش	آیتم‌های خریداری شده
۱۰	نان، مغزیجات، پنیر
۲۰	نان، قهوه، پنیر
۳۰	نان، پنیر، تخم مرغ
۴۰	مغزیجات، تخم مرغ، شیر
۵۰	مغزیجات، قهوه، پنیر، تخم مرغ، شیر

itemset های مکرر:

نان:۳ مغزیجات: ۳
پنیر:۴ تخم مرغ: ۳
{نان ، پنیر}: ۳

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{support(A,B)}{P(A)}$$

A → B

قوانین استخراج شده:

۱) پنیر → نان
۲) نان → پنیر

✓ Support = 60% , confidence = 100%

✓ Support = 60% , confidence = 75%

فرآیند کشف قوانین و الگوها

الگوریتم Apriori

- اگر itemset ایی مکرر باشد، باید تمام زیر مجموعه هایش هم مکرر باشد

۱- ابتدا 1-itemset ها رو محاسبه می‌کنیم

۲- itemset های کاندید با طول k+1 را از روی itemset های با طول k را می سازیم

۳- برای itemset های کاندید مکرر بودن را بررسی می‌کنیم

۴- به صورت تکراری مراحل ۲ و ۳ را تا زمان پایان itemset های مکرر انجام می‌دهیم

مثال

min support = **2**

مجموعه داده‌ها

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

؟

C_3

Itemset
{B, C, E}

3rd scan

Itemset	sup
{B, C, E}	2

L_3

مشکلات الگوریتم Apriori

- حجم بالا و تعداد اسکن پایگاه داده
- تعداد زیاد کاندید
- محاسبه support

برخی از راهکارهای ارائه شده

- ✓ Partition: رفع مشکل حجم بالا و تعداد اسکن پایگاه داده
- ✓ DIC: رفع مشکل اسکن زیاد پایگاه داده
- ✓ DHP: رفع مشکل تعداد زیاد کاندید و تلاش به محاسبه آسان support

روش Partition

۱. پارتیشن بندی کل داده‌ها
۲. محاسبه itemset های مکرر هر پارتیشن
۳. یافتن itemset هایی که حداقل در یکی از پارتیشن ها مکرر باشد
۴. تجمیع itemset ها و اسکن پایگاه داده

مثال

min support = **75%**

مجموعه داده‌ها

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

(۱)

P1) A, C, D
B, C, E

P2) A, B, C, E
B, E

(۲)

Itemset	sup	partition
{C}	2	P1
{B}	2	P2
{E}	2	P2
{B, E}	2	P2

(۳)

{B, E}

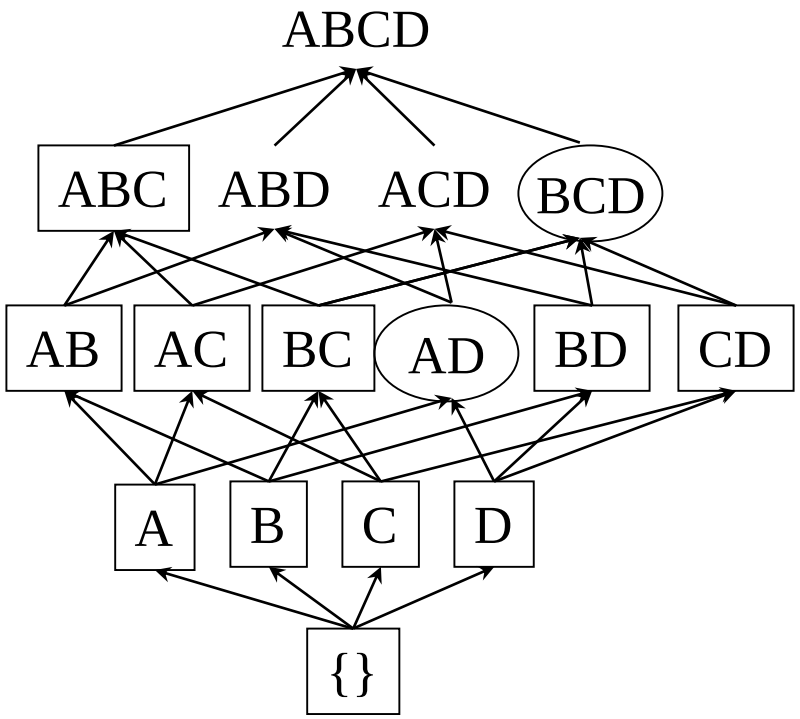
(۴)

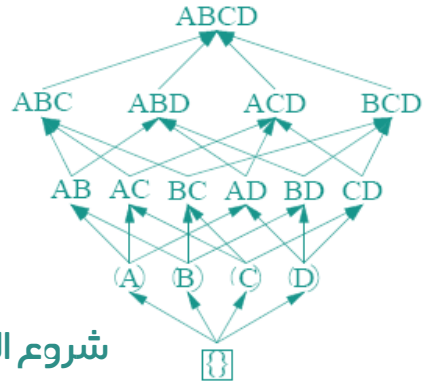
✓

روش DIC

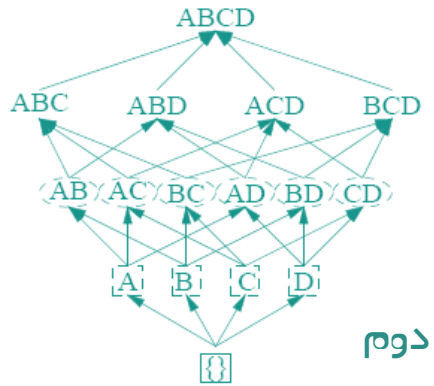
ایده: ساخت شبکه ای از itemset ها

شبکه ای از itemset ها

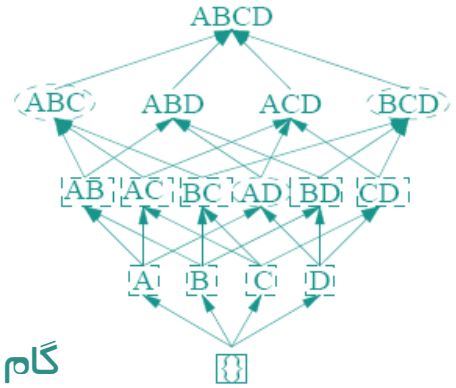




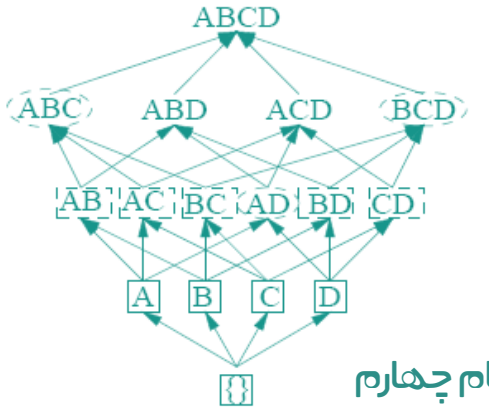
شروع الگوریتم



گام دوم



گام سوم



گام چهارم

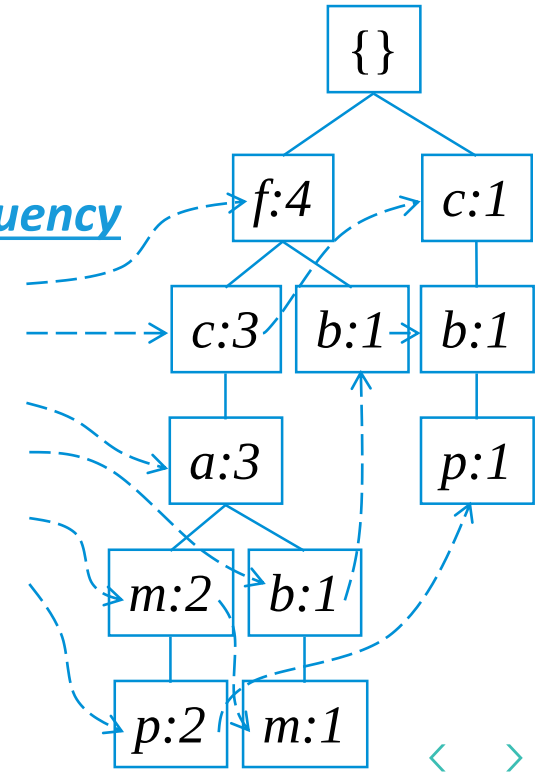
کاوش الگوهای مکرر بدون ایجاد کاندید

۱. پیدا کردن آیتم های مجاز
۲. مرتب سازی آیتم‌ها (ایجاد F-list)
۳. ایجاد ساختار درختی (FP-tree) و درج به صورت ترتیبی

min support = **3**

TID	Items bought	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

Item	frequency
f	4
c	4
a	3
b	3
m	3
p	3



F-list=f-c-a-b-m-p

کاوش الگوهای مکرر بدون ایجاد کاندید

پارتیشن بندی الگوها

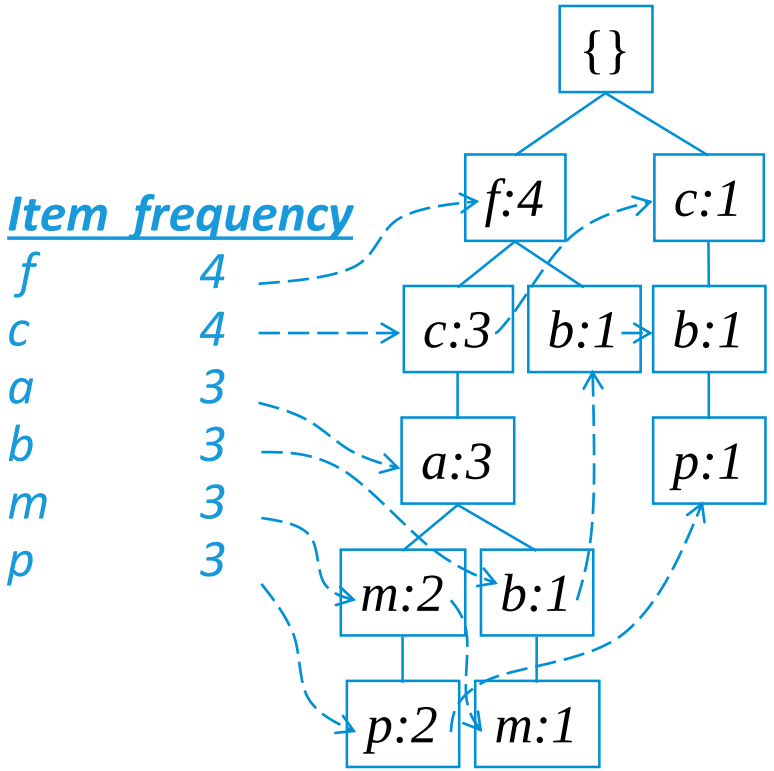
الگوهای شامل p

الگوهای شامل m که شامل p نباشد

الگوهای شامل b که شامل p و m نباشد

.....

الگوی فقط شامل f



پارتیشن بندی الگوها

item	cond. pattern base
c	f:3
a	fc:3
b	fca:1, f:1, c:1
m	fca:2, fcab:1
p	fcam:2, cb:1

کاوش الگوهای مکرر بدون ایجاد کاندید

TID	Items bought	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

پارتیشن بندی الگوها

الگوهای شامل p

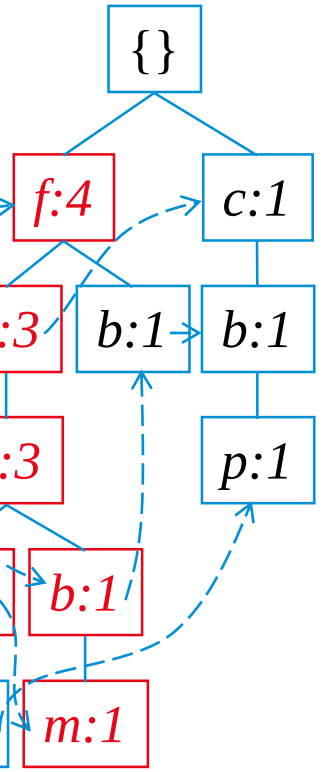
الگوهای شامل m که شامل p نباشد

الگوهای شامل b که شامل p و m نباشد

.....

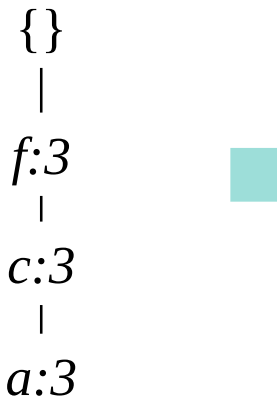
الگوی فقط شامل f

Item	frequency
f	4
c	4
a	3
b	3
m	3
p	3



پارتیشن الگو مبتنی بر وجود m

fca:2, fcab:1



m-conditional FP-tree

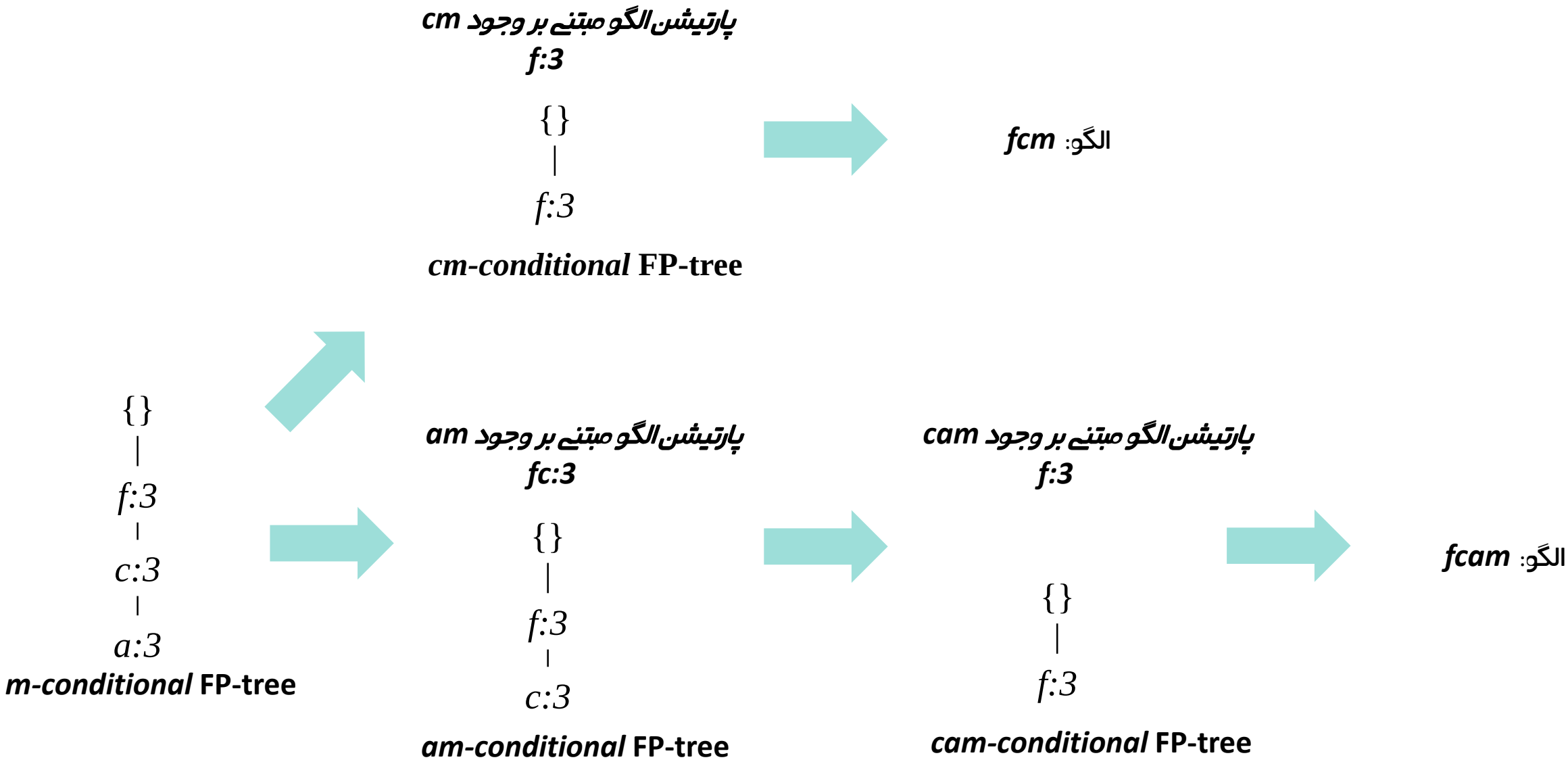
همه الگوهای قابل ایجاد با m

m,

fm, cm, am,

fcm, fam, cam,

fcam


پارتیشن الگو مبتنی بر وجود cm
f:3
{}
f:3
cm-conditional FP-tree
الگو: fcm
{}
f:3
c:3
a:3
m-conditional FP-tree
پارتیشن الگو مبتنی بر وجود am
fc:3
{}
f:3
c:3
am-conditional FP-tree
پارتیشن الگو مبتنی بر وجود cam
f:3
{}
f:3
cam-conditional FP-tree
الگو: fcam

چرا FP-Growth موفق است؟

- ✓ استراتژی D&C
- ✓ نیاز به ایجاد کاندید و تست کاندیدها ندارد
- ✓ پایگاه داده را فشرده نگه می دارد
- ✓ نیاز به اسکن چندباره پایگاه داده ندارد

آیا معیارهای min support و min confidence کافی است؟

min support = **40%** , min confidence = **60%**

	دستگاه بازی	خریدن دستگاه بازی	مجموع (سطری)
دستگاه ویدیو	۴۰۰۰	۳۵۰۰	۷۵۰۰
خریدن دستگاه ویدیو	۲۰۰۰	۵۰۰	۲۵۰۰
مجموع (ستونی)	۶۰۰۰	۴۰۰۰	۱۰۰۰۰

قوانین جهت بررسی:

۱) خرید دستگاه بازی → خرید دستگاه ویدیو
۲) خرید دستگاه ویدیو → خرید دستگاه بازی

قانون یک : support = **4000/10000 = 40%** , confidence = **4000/7500 = 53.3%** ✗

قانون دو : support = **4000/10000 = 40%** , confidence = **4000/6000 = 66.7%** ✓

۷۵درصد به هرحال دستگاه ویدیو رو میخرن و وابسته به خرید دستگاه بازی نیست.

معیار lift: بررسی میزان همبستگی آیتم‌ها

$A \rightarrow B \qquad lift(A, B) = \frac{P(A \cap B)}{P(A) \times P(B)}$

اگر lift = 1 مستقل
اگر lift > 1 همبستگی مثبت
اگر lift < 1 همبستگی منفی

	دستگاه بازی	نخریدن دستگاه بازی	مجموع (سطری)
دستگاه ویدیو	۴۰۰۰	۳۵۰۰	۷۵۰۰
نخریدن دستگاه ویدیو	۲۰۰۰	۵۰۰	۲۵۰۰
مجموع (ستونی)	۶۰۰۰	۴۰۰۰	۱۰۰۰۰

قوانین جهت بررسی:

۲) خرید دستگاه ویدیو → خرید دستگاه بازی
۳) نخریدن دستگاه ویدیو → خرید دستگاه بازی

قانون دو: $lift = \frac{4000/10000}{6000/10000 \times 7500/10000} = 0.89$ ✗

قانون سه: $lift = \frac{2000/10000}{6000/10000 \times 2500/10000} = 1.33$ ✓

سایر معیارها

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1 . . . 1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 . . . 1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 . . . 1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 . . . 1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 . . . 0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 . . . 1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
AV	added value	-0.5 . . . 1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
K	Klosgen's Q	-0.33 . . . 0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$
g	Goodman-kruskal's	0 . . . 1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 . . . 1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i), -\Sigma_i P(B_i)\log P(B_i)\log P(B_i))}$
J	J-Measure	0 . . . 1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
G	Gini index	0 . . . 1	$\max(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A}[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B}[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]-P(A)^2-P(\overline{A})^2)$
s	support	0 . . . 1	$P(A,B)$
c	confidence	0 . . . 1	$max(P(B\|A), P(A\|B))$
L	Laplace	0 . . . 1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 . . . 1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 . . . 1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 . . . 1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 . . . ∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 . . . ∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 . . . ∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 . . . ∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	χ^2	0 . . . ∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

اعمال محدودیت ها

- الگوهای زیاد و نامتمرکز
- اعمال محدودیت‌ها و درخواست های کاربر مثلا مجموع قیمت کمتر از مبلغی باشد
- محدودیت ها و شرایط کاربر باید بررسی شوند

مثال

min support = **2**
sum(price) < **5**

item	price
A	1
B	2
C	3
D	4
E	5

مجموعه داده‌ها

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1^{st} scan

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2^{nd} scan

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

تمرین

- مجموعه آیتم های مکرر را بیابید.
- چند قانون یا الگو متناسب با شرایط زیر پیشنهاد دهید.

min support = **33.33%** , min confidence = **60%**

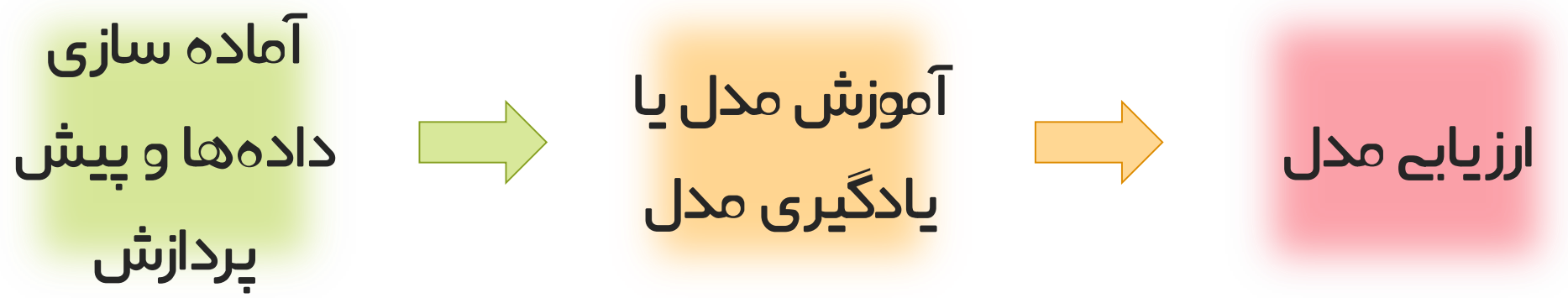
مجموعه داده‌ها

شناسه	اقلام تراکنش
T1	سوسیس، نان ساندویچی، سس گوجه
T2	سوسیس، نان ساندویچی
T3	سوسیس، نوشابه، چیپس
T4	چیپس، نوشابه
T5	چیپس، سس گوجه
T6	سوسیس، نوشابه، چیپس

رده بندی یا Classification

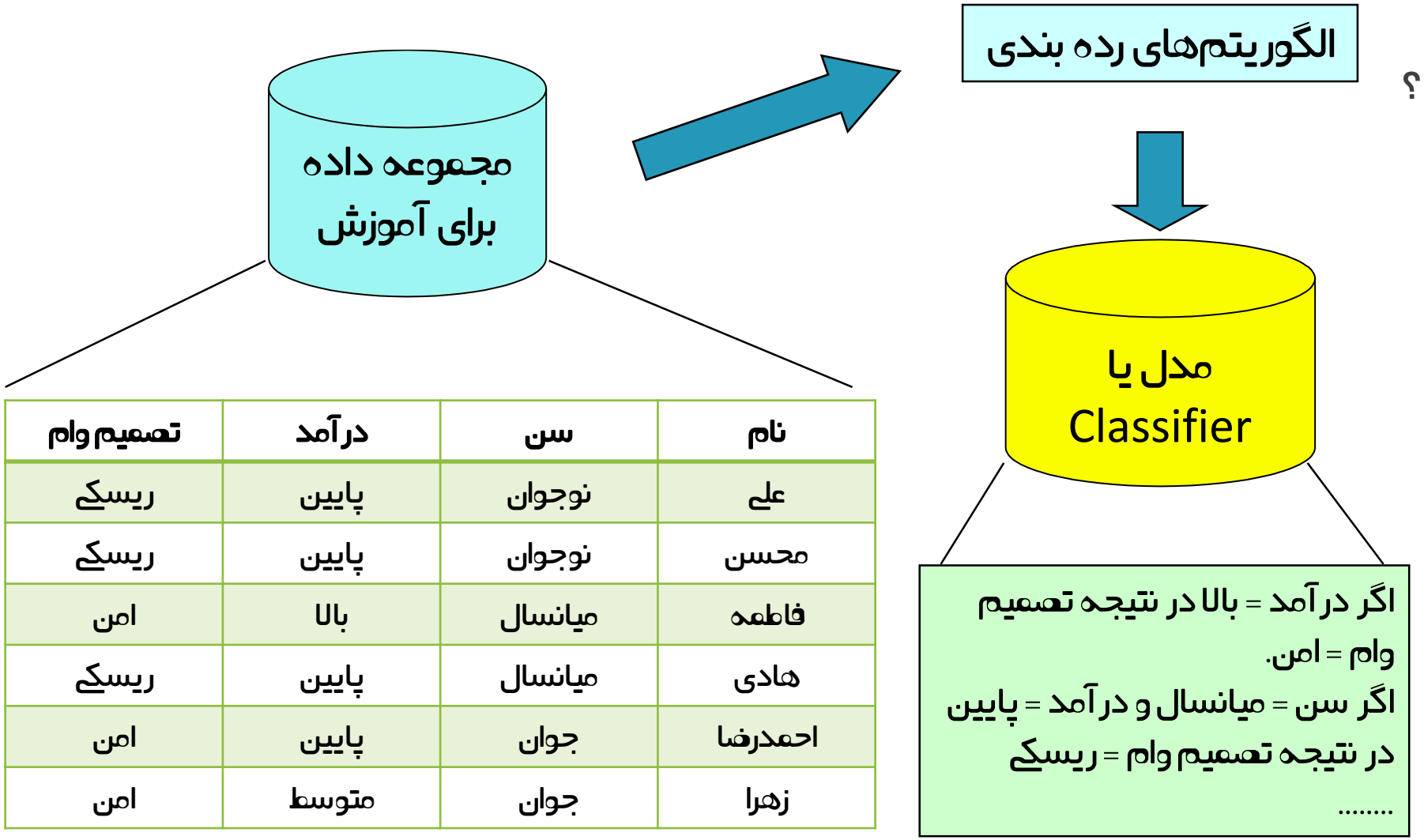
- هدف چیست؟
- بررسی مفاهیم مانند تاپل، کلاس یا برچسب و ...
- کاربردها

فرآیند کلی



مثال

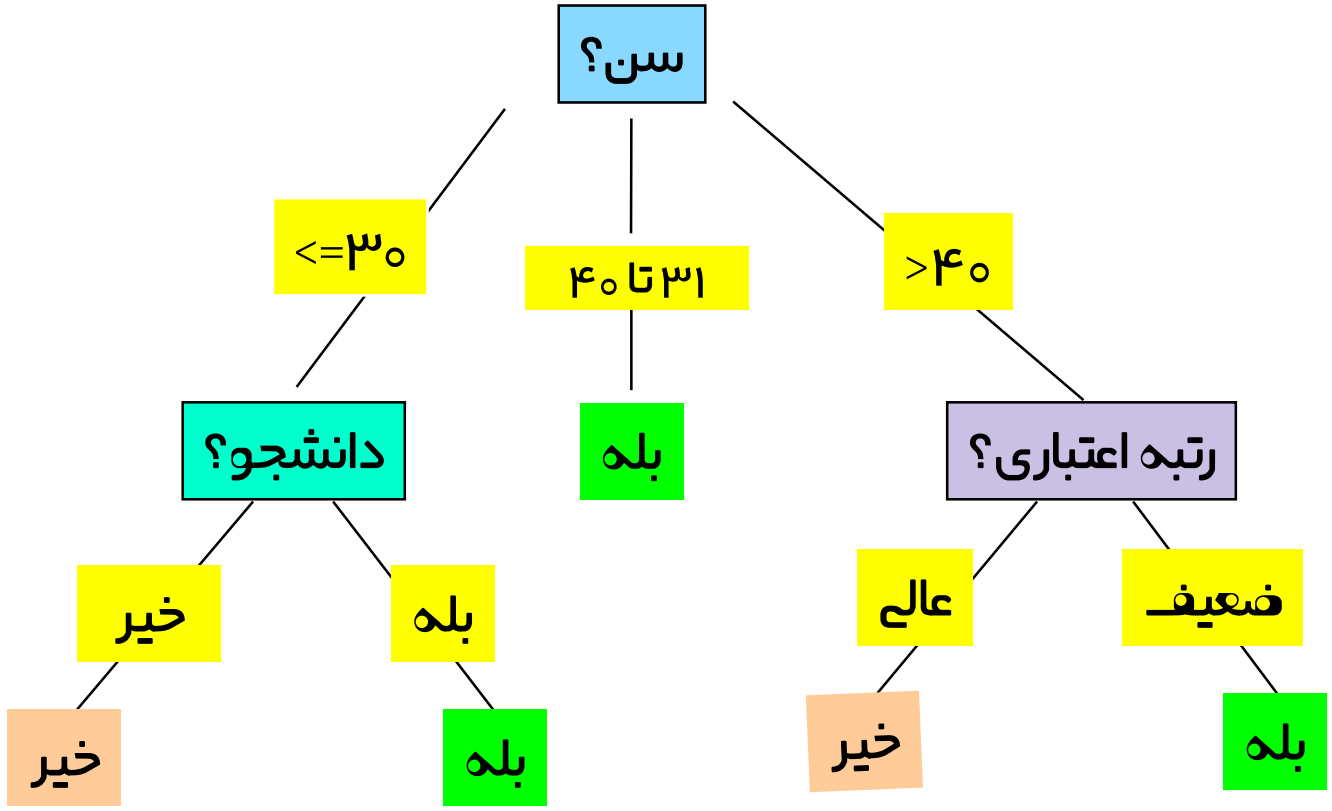
وام بدهیم یا خیر؟



نام	سن	درآمد	تصمیم وام
علی	نوجوان	پایین	ریسکی
محسن	نوجوان	پایین	ریسکی
فاطمه	میانسال	بالا	امن
هادی	میانسال	پایین	ریسکی
احمدرضا	جوان	پایین	امن
زهرا	جوان	متوسط	امن

سن	درآمد	دانشجو	رتبه اعتباری	خرید کامپیوتر
<=۳۰	بالا	خیر	ضعیف	خیر
<=۳۰	بالا	خیر	عالی	خیر
۳۱ تا ۴۰	بالا	خیر	ضعیف	بله
>۴۰	متوسط	خیر	ضعیف	بله
>۴۰	پایین	بله	ضعیف	بله
>۴۰	پایین	بله	عالی	خیر
۳۱ تا ۴۰	پایین	بله	عالی	بله
<=۳۰	متوسط	خیر	ضعیف	خیر
<=۳۰	پایین	بله	ضعیف	بله
>۴۰	متوسط	بله	ضعیف	بله
<=۳۰	متوسط	بله	عالی	بله
۳۱ تا ۴۰	متوسط	خیر	عالی	بله
۳۱ تا ۴۰	بالا	بله	ضعیف	بله
>۴۰	متوسط	خیر	عالی	خیر

درخت تصمیم – ID3



درخت تصمیم – ID3

- چرا ابتدا سن انتخاب شد؟

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

سن	درآمد	دانشجو	رتبه اعتباری	خرید کامپیوتر
<=۳۰	بالا	خیر	ضعیف	خیر
<=۳۰	بالا	خیر	عالی	خیر
۳۱ تا ۴۰	بالا	خیر	ضعیف	بله
>۴۰	متوسط	خیر	ضعیف	بله
>۴۰	پایین	بله	ضعیف	بله
>۴۰	پایین	بله	عالی	خیر
۳۱ تا ۴۰	پایین	بله	عالی	بله
<=۳۰	متوسط	خیر	ضعیف	خیر
<=۳۰	پایین	بله	ضعیف	بله
>۴۰	متوسط	بله	ضعیف	بله
<=۳۰	متوسط	بله	عالی	بله
۳۱ تا ۴۰	متوسط	خیر	عالی	بله
۳۱ تا ۴۰	بالا	بله	ضعیف	بله
>۴۰	متوسط	خیر	عالی	خیر

درخت تصمیم – ID3

چرا ابتدا سن انتخاب شد؟

۱) $$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

۲)

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

۳) $$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

۴) $$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$ ✓

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

مشکل الگوریتم ID3

- در IG ، تلاش در کاهش آنتروپی و در نتیجه کاهش عدم خلوص
- بر اساس کد ملی مشتریان چطور؟

✓ الگوریتم C4.5: نرمال سازی – جریمه فیلد به ازای تنوع در مقادیر

✓ الگوریتم CART: استفاده از معیار Gini و ایجاد درخت باینری

Overfitting و هرس کردن

رده بندی بیزین

- مبتنی بر احتمالات و فرمول بیز
- احتمال تعلق یک نمونه به کلاس –> کلاس با بیشترین احتمال
- مثال نمونه X: سن ۳۰ و درآمد ۴۰ میلیون تومان

اگر بدانیم فرضیه برقرار است،
احتمال مشاهده نمونه X

احتمال خرید یا نخریدن کامپیوتر

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

احتمال مشاهده نمونه

احتمال هر یک از فرضیات به شرط X
مثلا: احتمال خرید کامپیوتر به شرط X
احتمال نخریدن کامپیوتر به شرط X

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})} \rightarrow P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

مثال از رده بند Naïve Bayes

سوال: پیش بینی کنید که اگر یک مشتری دانشجو با سن کمتر از ۳۰، درآمد متوسط و دارای رتبه اعتباری ضعیف باشد، کامپیوتر می‌خرد؟

سن	درآمد	دانشجو	رتبه اعتباری	خرید کامپیوتر
<=۳۰	بالا	خیر	ضعیف	خیر
<=۳۰	بالا	خیر	عالی	خیر
۳۱ تا ۴۰	بالا	خیر	ضعیف	بله
>۴۰	متوسط	خیر	ضعیف	بله
>۴۰	پایین	بله	ضعیف	بله
>۴۰	پایین	بله	عالی	خیر
۳۱ تا ۴۰	پایین	بله	عالی	بله
<=۳۰	متوسط	خیر	ضعیف	خیر
<=۳۰	پایین	بله	ضعیف	بله
>۴۰	متوسط	بله	ضعیف	بله
<=۳۰	متوسط	بله	عالی	بله
۳۱ تا ۴۰	متوسط	خیر	عالی	بله
۳۱ تا ۴۰	بالا	بله	ضعیف	بله
>۴۰	متوسط	خیر	عالی	خیر

گام اول) محاسبه $P(C_i)$ ها:

P(خریدن کامپیوتر) = 9/14 = 0.643
P(نخریدن کامپیوتر) = 5/14= 0.357

گام دوم) محاسبه $P(X|C_i)$ برای هر کلاس:

P(سن کمتر از ۳۰ | خریدن کامپیوتر) = 2/9 = 0.222
P(سن کمتر از ۳۰ | نخریدن کامپیوتر) = 3/5 = 0.6
P(درآمد متوسط | خریدن کامپیوتر) = 4/9 = 0.444
P(درآمد متوسط | نخریدن کامپیوتر) = 2/5 = 0.4
P(دانشجو بودن | خریدن کامپیوتر) = 6/9 = 0.667
P(دانشجو بودن | نخریدن کامپیوتر) = 1/5 = 0.2
P(رتبه اعتباری ضعیف | خریدن کامپیوتر) = 6/9 = 0.667
P(رتبه اعتباری ضعیف | نخریدن کامپیوتر) = 2/5 = 0.4

مثال از رده بند Naïve Bayes

سن	درآمد	دانشجو	رتبه اعتباری	خرید کامپیوتر
<=۳۰	بالا	خیر	ضعیف	خیر
<=۳۰	بالا	خیر	عالی	خیر
۳۱ تا ۴۰	بالا	خیر	ضعیف	بله
>۴۰	متوسط	خیر	ضعیف	بله
>۴۰	پایین	بله	ضعیف	بله
>۴۰	پایین	بله	عالی	خیر
۳۱ تا ۴۰	پایین	بله	عالی	بله
<=۳۰	متوسط	خیر	ضعیف	خیر
<=۳۰	پایین	بله	ضعیف	بله
>۴۰	متوسط	بله	ضعیف	بله
<=۳۰	متوسط	بله	عالی	بله
۳۱ تا ۴۰	متوسط	خیر	عالی	بله
۳۱ تا ۴۰	بالا	بله	ضعیف	بله
>۴۰	متوسط	خیر	عالی	خیر

P(خریدن کامپیوتر) = 9/14 = 0.643
P(نخریدن کامپیوتر) = 5/14= 0.357

P(سن کمتر از ۳۰ | خریدن کامپیوتر) = 2/9 = 0.222
P(سن کمتر از ۳۰ | نخریدن کامپیوتر) = 3/5 = 0.6
P(درآمد متوسط | خریدن کامپیوتر) = 4/9 = 0.444
P(درآمد متوسط | نخریدن کامپیوتر) = 2/5 = 0.4
P(دانشجو بودن | خریدن کامپیوتر) = 6/9 = 0.667
P(دانشجو بودن | نخریدن کامپیوتر) = 1/5 = 0.2
P(رتبه اعتباری ضعیف | خریدن کامپیوتر) = 6/9 = 0.667
P(رتبه اعتباری ضعیف | نخریدن کامپیوتر) = 2/5 = 0.4

گام سوم) محاسبه احتمال نهایی:

$P(X|C_i)$: P(X|خریدن کامپیوتر) = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|نخریدن کامپیوتر) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|خریدن کامپیوتر) * P(خریدن کامپیوتر) = 0.028
P(X|نخریدن کامپیوتر) * P(نخریدن کامپیوتر) = 0.007

بنابراین نظر ماشین با این مجموعه داده، خریدن کامپیوتر است.

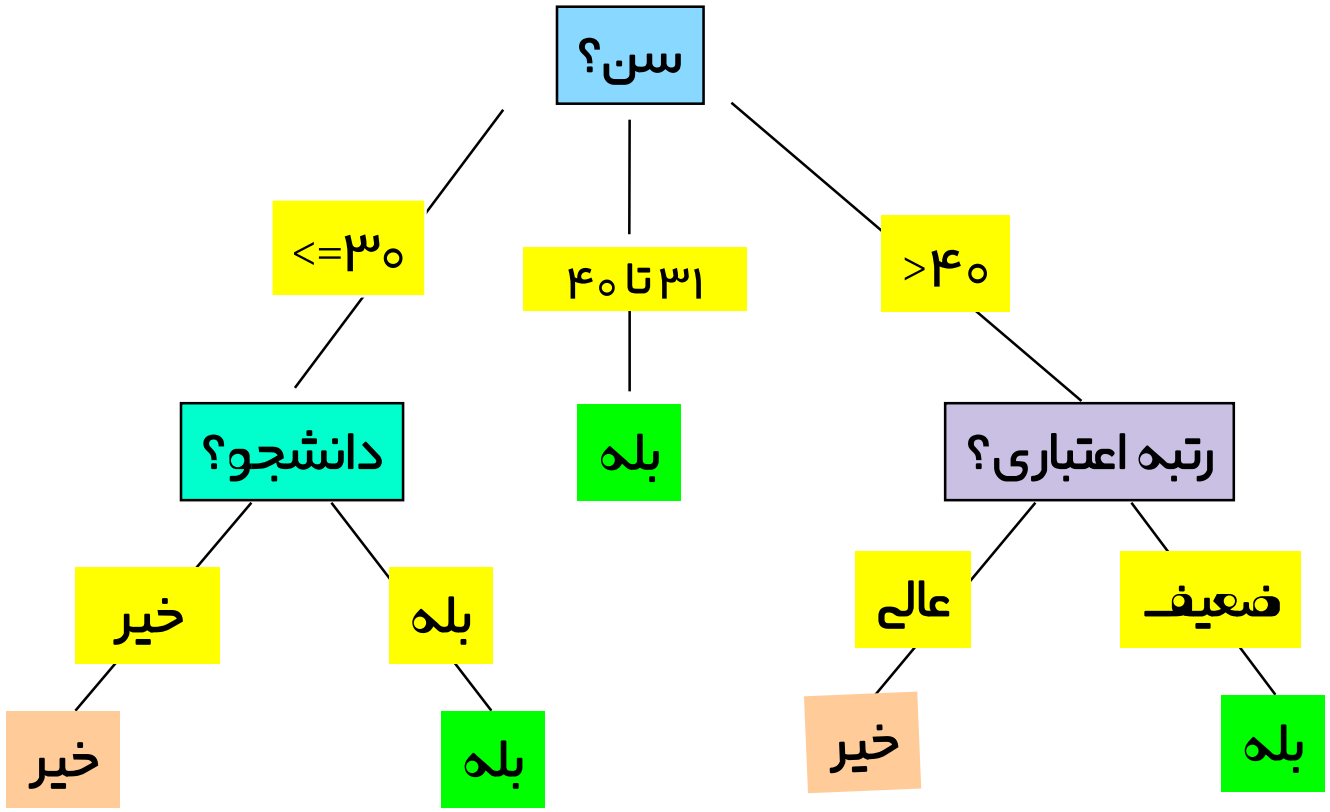
معایب رده بند بیزین

- فرض مستقل بودن: مثال بیمار
- انواع داده‌ها به ویژه داده‌های متنی

رده بندی با قوانین

- قوانین را چگونه به دست آوریم؟
- برای تداخل‌ها چه اقدامی می‌توان انجام داد؟
  - ✓ اعمال محدودیت‌های بیشتر
  - ✓ مرتب سازی کلاس‌ها بر اساس فرکانس
  - ✓ مرتب سازی قوانین با نظر کارشناس یا معیارهای آماری
- عدم تطابق قوانین با نمونه جدید

استخراج قوانین از درخت تصمیم

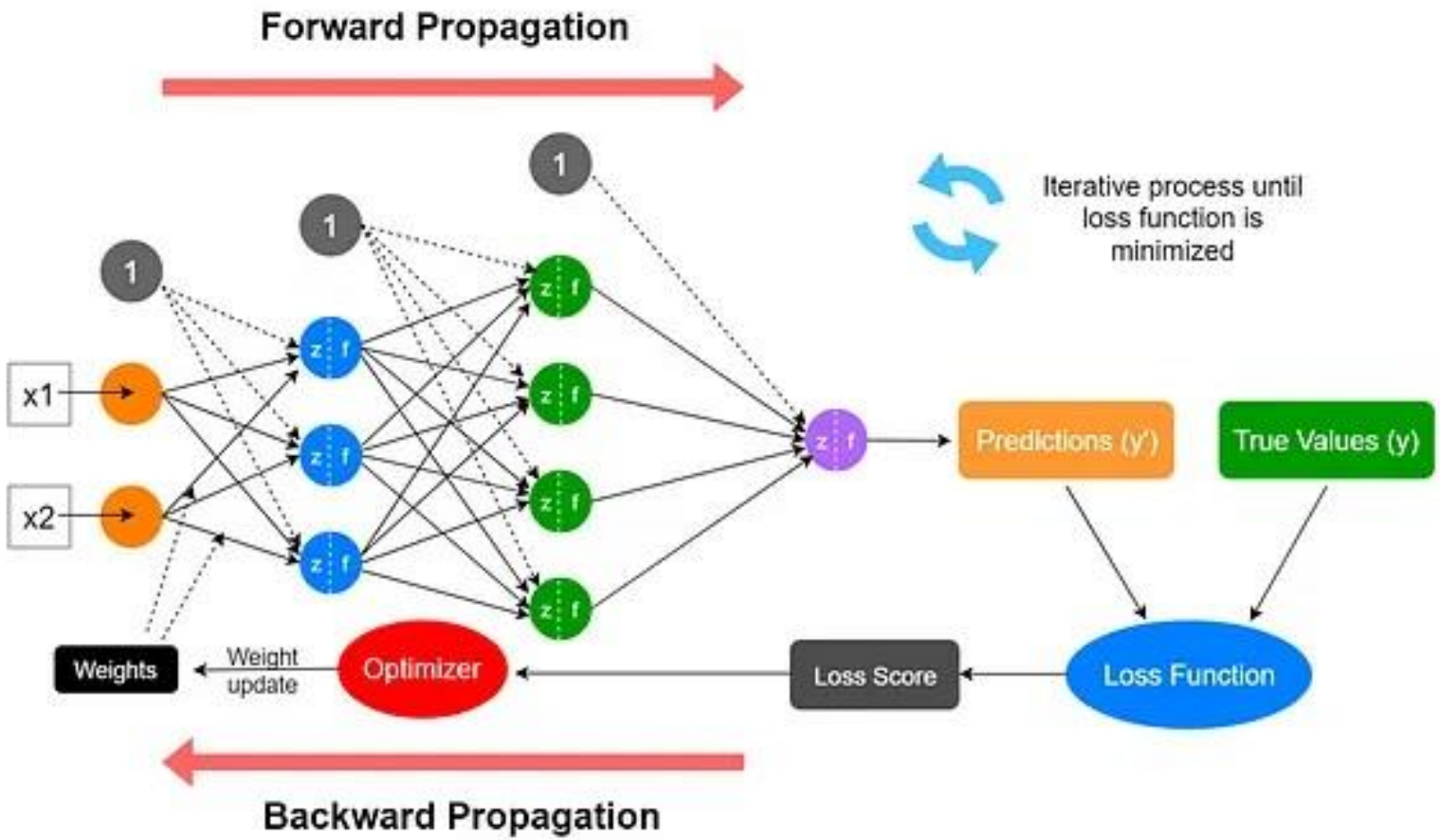


نمونه‌هایی از قوانین:

اگر سن کمتر از ۳۰ و دانشجو نبود، در نتیجه کامپیوتر نمی‌خرد
اگر سن کمتر از ۳۰ و دانشجو بود، در نتیجه کامپیوتر می‌خرد
اگر سن بین ۳۱ تا ۴۰ بود، در نتیجه کامپیوتر می‌خرد

استخراج قوانین از قوانین انجمنی

رده بندی با شبکه‌های عصبی



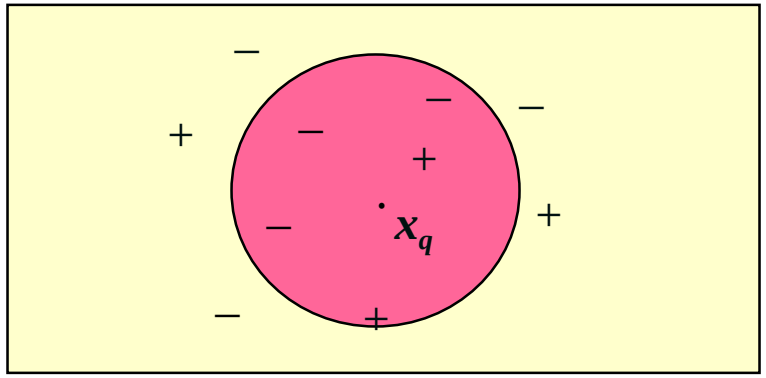
منبع تصویر: https://miro.medium.com

- زمان آموزش بالا
- طراحی مدل
- عدم تفسیر مدل ایجاد شده

- مقاوم در برابر نویز
- استفاده در اکثر کاربردها

رده بندی مبتنی بر نزدیک‌ترین همسایه‌ها یا *k*-Nearest Neighbor

- یک روش lazy است
- نزدیک‌تر بودن – وزن بیشتر
- انتخاب k مناسب
- مقاوم در برابر نویز



رده بندی رگرسیون

- هدف و کاربرد
- ارتباط بین متغیر و پاسخ

ارزیابی

توانایی مدل در پیش بینی درست

- دقت

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

خطاهای مدل در پیش بینی

- نرخ خطا

$$error\ rate = \frac{FP + FN}{TP + TN + FP + FN}$$

- اگر داده‌ها بالانس نباشند؟

ماتریس در هم ریختگی (confusion)

مقدار پیش بینی شده/مقدار واقعی	(P) C_1	(N) $\sim C_1$
(P) C_1	True Positives (TP)	False Negatives (FN)
(N) $\sim C_1$	False Positives (FP)	True Negatives (TN)

TP : کلاس مثبت و به درستی پیش بینی شده
FP : کلاس منفی و به اشتباه پیش بینی شده
TN : کلاس منفی و به درستی پیش بینی شده
FN : کلاس مثبت و به اشتباه پیش بینی شده

مثال: آیا فرد مبتلا به کرونا است؟

	بلی	**خیر**
بلی	TP(71)	FN(4)
خیر	FP(10)	TN(15)

ارزیابی

ماتریس در هم ریختگی (confusion)

مقدار پیش بینی شده/مقدار واقعی	(P) C_1	(N) $\sim C_1$
(P) C_1	True Positives (TP)	False Negatives (FN)
(N) $\sim C_1$	False Positives (FP)	True Negatives (TN)

TP : کلاس مثبت و به درستی پیش بینی شده
FP : کلاس منفی و به اشتباه پیش بینی شده
TN : کلاس منفی و به درستی پیش بینی شده
FN : کلاس مثبت و به اشتباه پیش بینی شده

موارد مثبتی که مدل درست پیش بینی کرده است

❖ حساسیت

$$sensitivity = \frac{TP}{P}$$

موارد منفی که مدل درست پیش بینی کرده است

❖ تشخیص پذیری (خاصیت)

$$specificity = \frac{TN}{N}$$

مثال: آیا فرد مبتلا به کرونا است؟

	بله	خیر
بله	TP(71)	FN(4)
خیر	FP(10)	TN(15)

ارزیابی

درست بودن نتیجه، وقتی مثبت پیش بینی می شود

صحت

$$precision = \frac{TP}{TP + FP}$$

پوشش مدل در پیش بینی

پوشش

$$recall = \frac{TP}{TP + FN}$$

ماتریس در هم ریختگی (confusion)

مقدار پیش بینی شده/مقدار واقعی	(P) C_1	(N) $\sim C_1$
(P) C_1	True Positives (TP)	False Negatives (FN)
(N) $\sim C_1$	False Positives (FP)	True Negatives (TN)

TP : کلاس مثبت و به درستی پیش بینی شده
FP : کلاس منفی و به اشتباه پیش بینی شده
TN : کلاس منفی و به درستی پیش بینی شده
FN : کلاس مثبت و به اشتباه پیش بینی شده

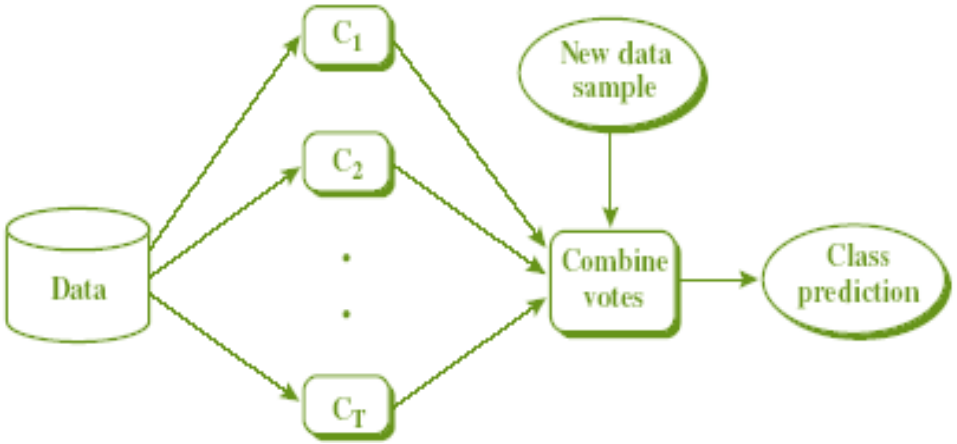
مثال: آیا فرد مبتلا به کرونا است؟

	بله	**خیر**
بله	TP(71)	FN(4)
خیر	FP(10)	TN(15)

نحوه ارزیابی مدل

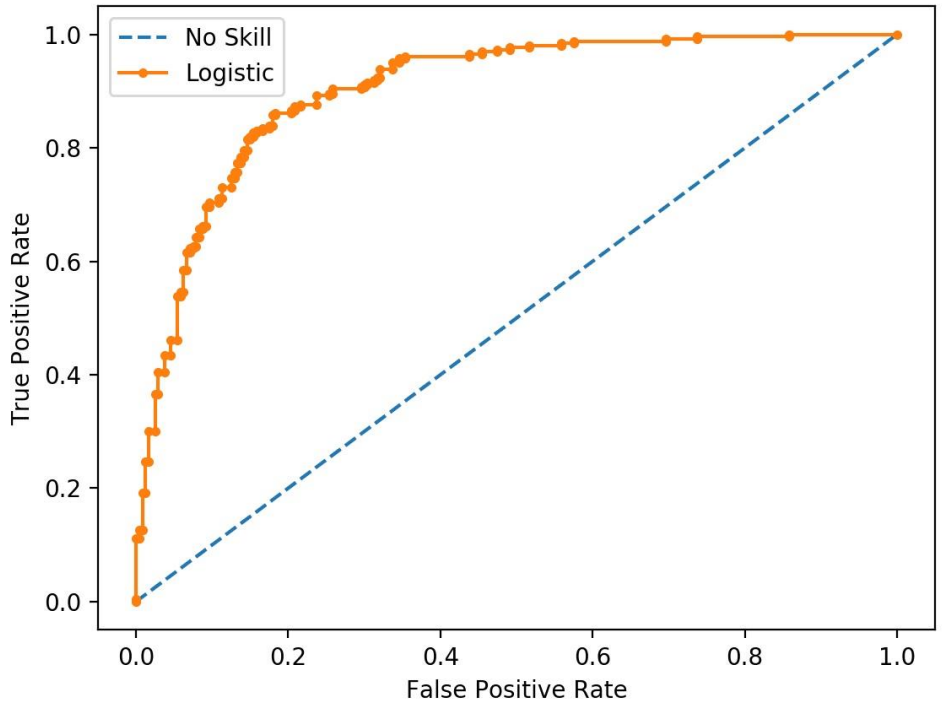
- روش Holdout
  - به نسبت مشخص
  - به صورت تصادفی بدون برگرداندن
- روش Cross-validation
  - K پارتیشن
  - برچسب اختصاص در یک پارتیشن
- روش Bootstrap

روش گروهی یا Ensemble



- نظر چند رده بند
- Bagging
- Boosting

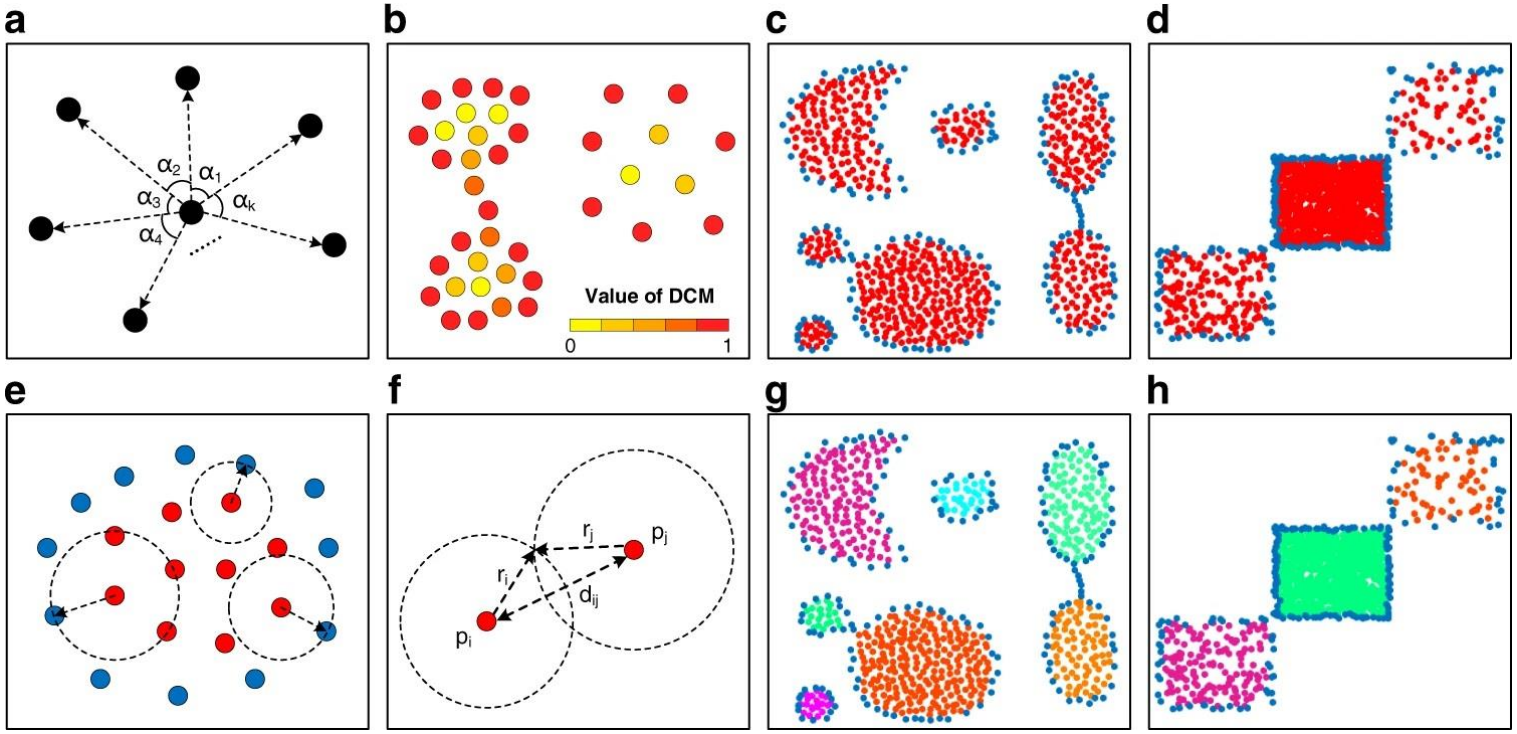
نمودار ROC



- سطح زیر نمودار

خوشه بندی

- هدف
- به عنوان تسک (بدون برچسب)
- به عنوان پیش پردازش
- کاربرد
- شباهت = فاصله
- خوشه بندی خوب



منبع تصویر: مقاله

Clustering by measuring local direction centrality for data with heterogeneous density and weak connectivity

شباهت یا Similarity

- بر اساس نوع فیلد

باینری
اسمی
عددی
ترتیبی

باینری		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

Object ID	Test-1 (categorical)
1	Code-A
2	Code-B
3	Code-C
4	Code-A

اسمی

Object ID	Test-1 (ordinal)
1	Excellent
2	Fair
3	Good
4	Excellent

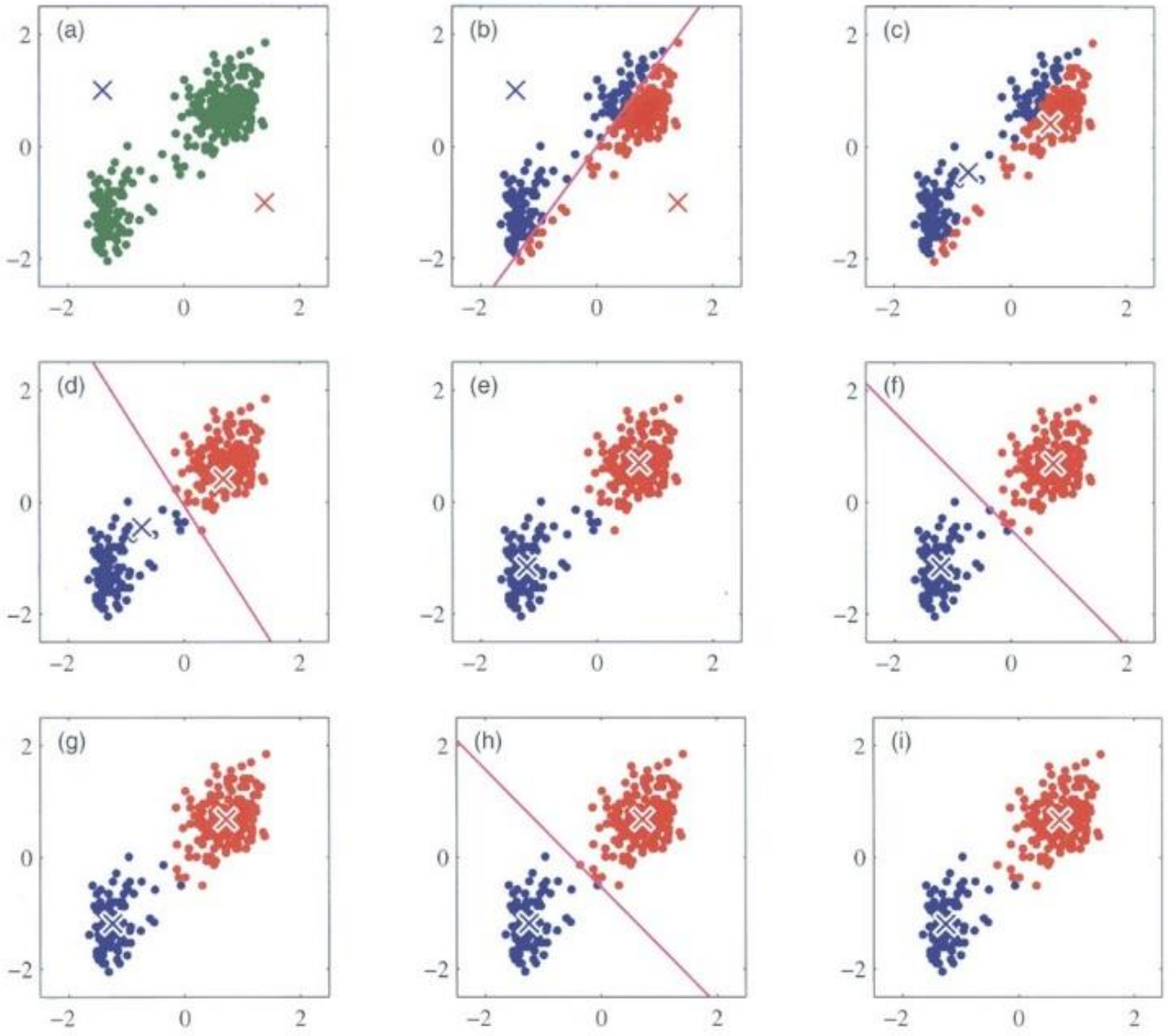
ترتیبی

شباهت خوشه‌ها

- Single Link کمترین فاصله بین عناصر دو خوشه
- Complete Link بیشترین فاصله بین عناصر دو خوشه
- Average متوسط فاصله بین عناصر دو خوشه
- Centroid فاصله مرکزهای خوشه
- Medoid فاصله عناصر مرکزی خوشه‌ها

خوشه بندی – پارتیشنینگ

- k-means
- k-medoids
- CLARA



منبع تصویر: dendroid.sk

الگوریتم k-means

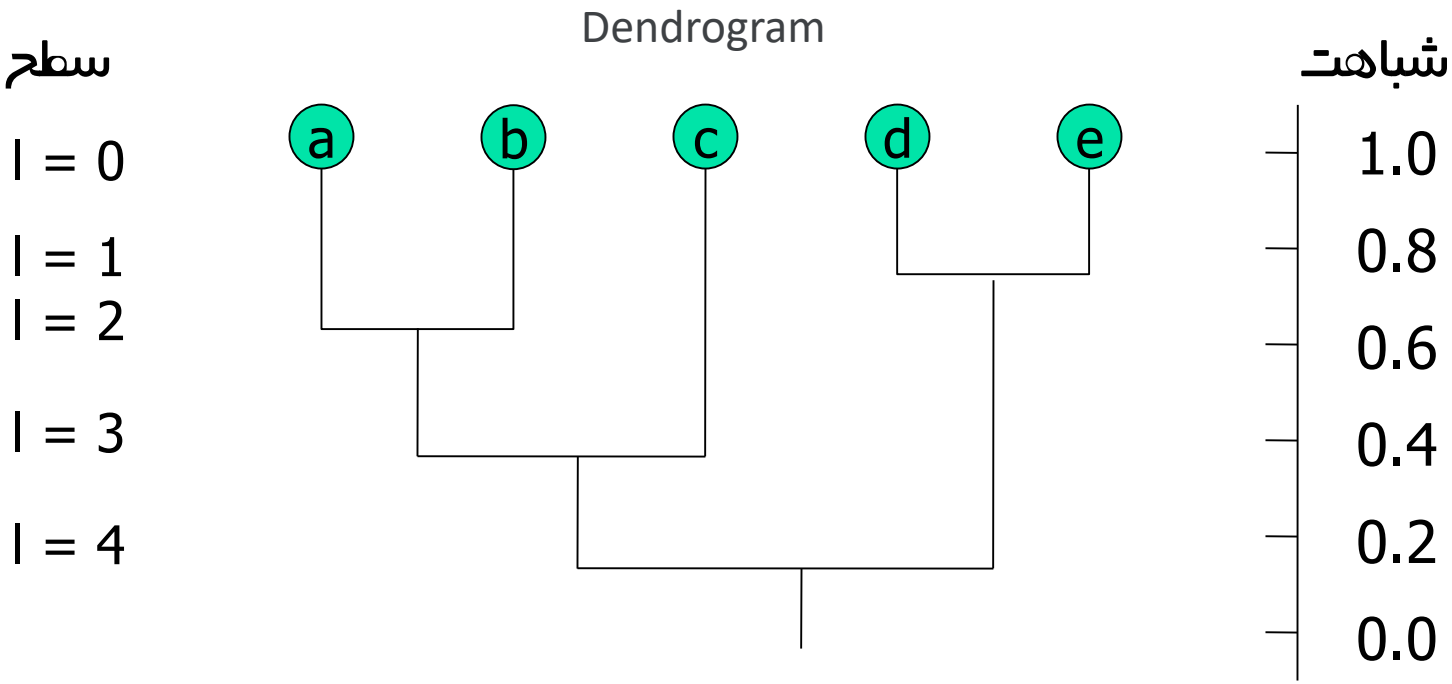
۱. انتخاب k مرکز و خوشه بندی
۲. پیدا کردن مرکز جدید و بررسی جابه‌جایی آبجکت ها
۳. تکرار مرحله دو، تا زمانی که آبجکتی جابه‌جا نشود

معایب الگوریتم k-means

- اگر خصیصه عددی نبود
- مشخص نمودن k یا تعداد خوشه‌ها
- داده‌های نویز و پَرت

خوشه بندی – سلسله مراتبی

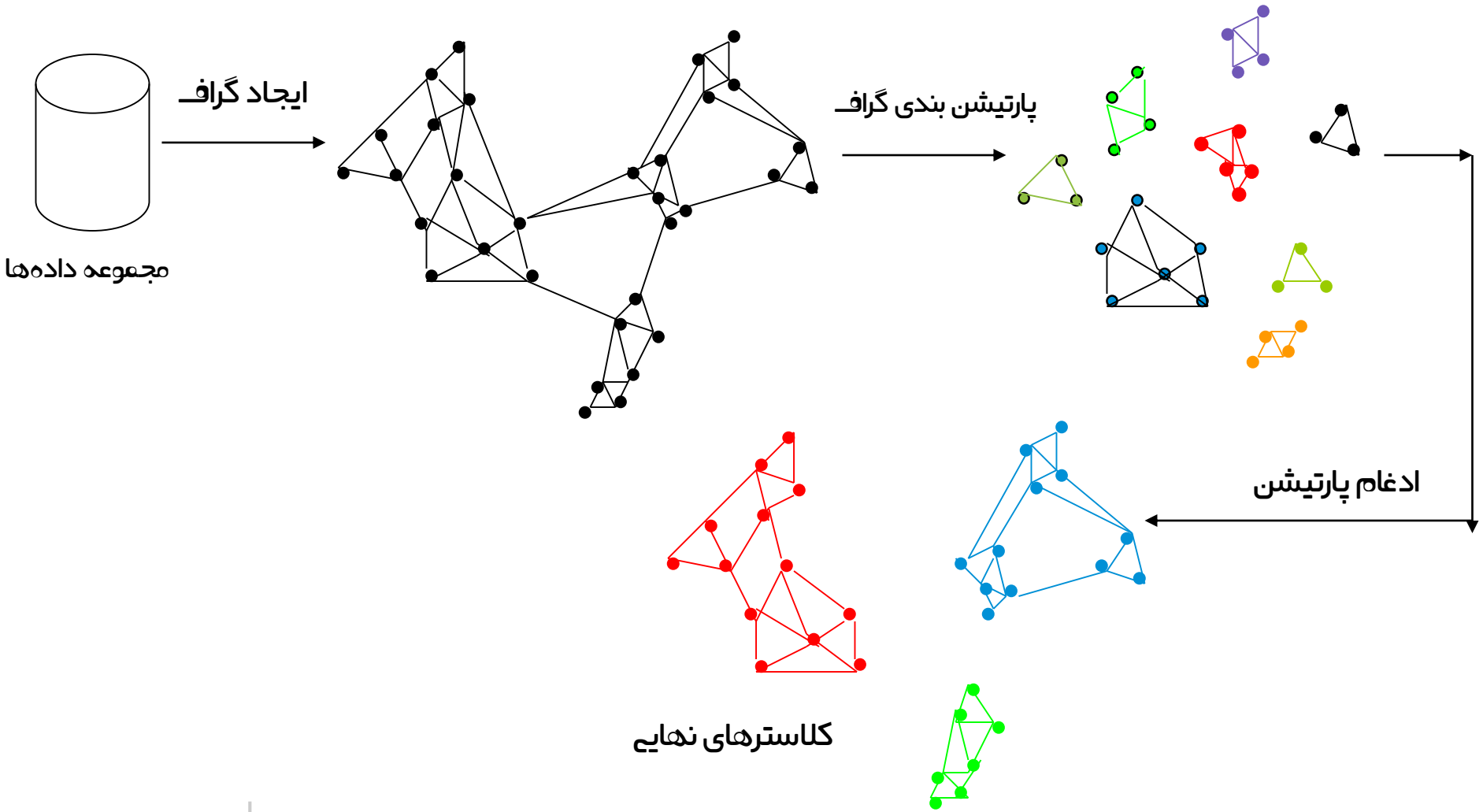
- عدم نیاز به مشخص کردن k
- نیاز به شرط خاتمه
- پایین به بالا – بالا به پایین



CHAMELEON – ROCK – BIRCH – AGNES – DIANA

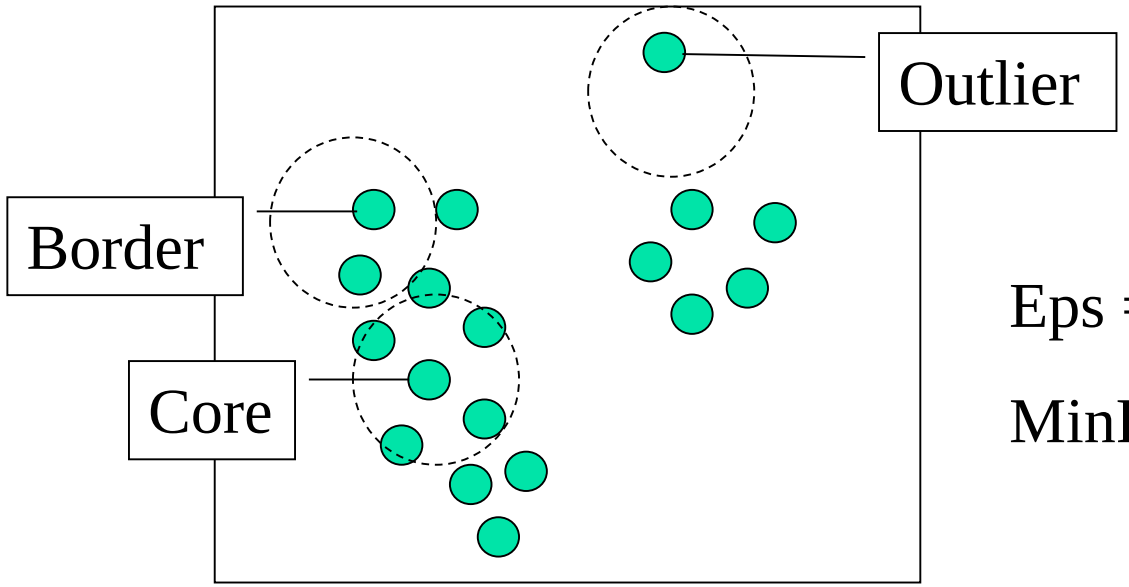
الگوریتم CHAMELEON

- سلسله مراتبی



خوشه بندی – مبتنی بر چگالی

- استفاده از چگالی نقاط و همسایگی
- حساس به پارامترها



Eps = 1cm

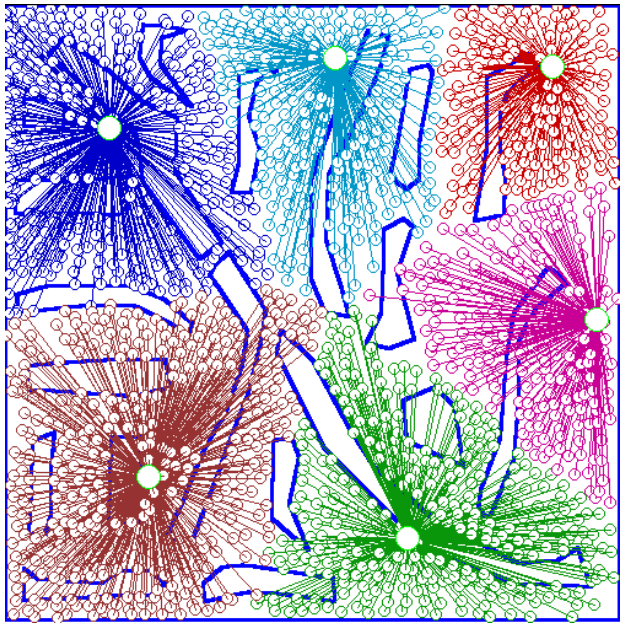
MinPts = 5

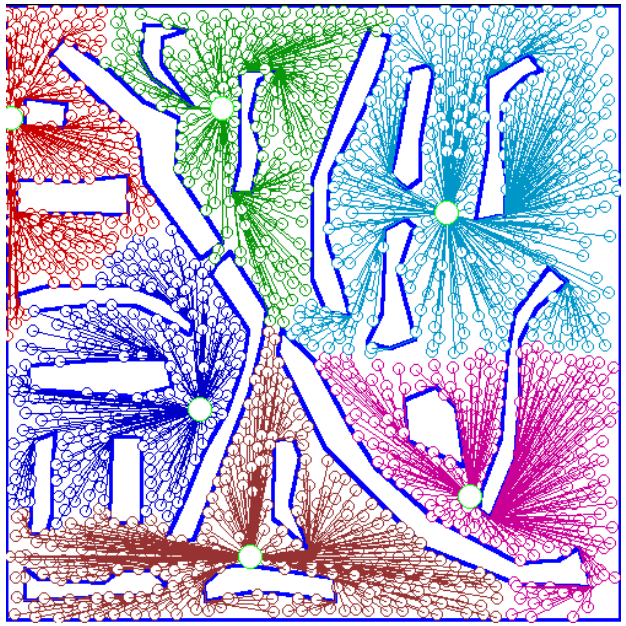
- DBSCAN
- OPTICS
- DENCLUE
- CLIQUE

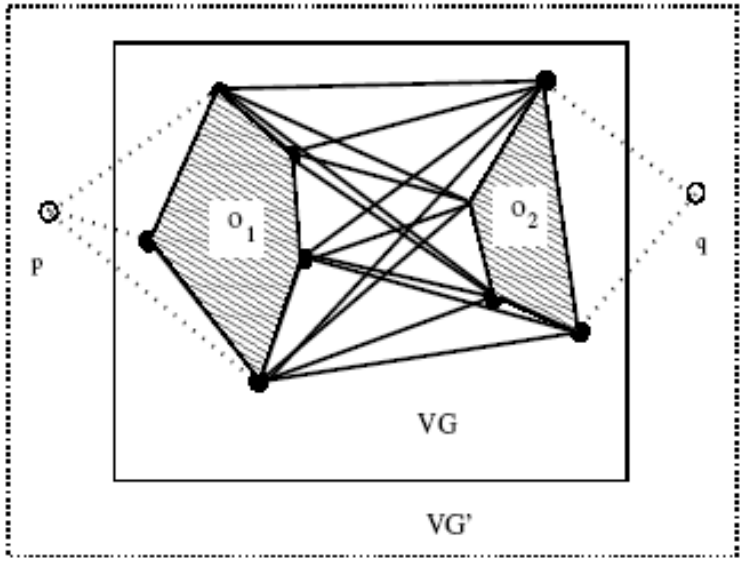
خوشه بندی – متناسب با محدودیت ها

- بر اساس دانش کاربر
- مثال: کمینه کردن فاصله شهروندان از دستگاه های ATM
- سخت و نرم

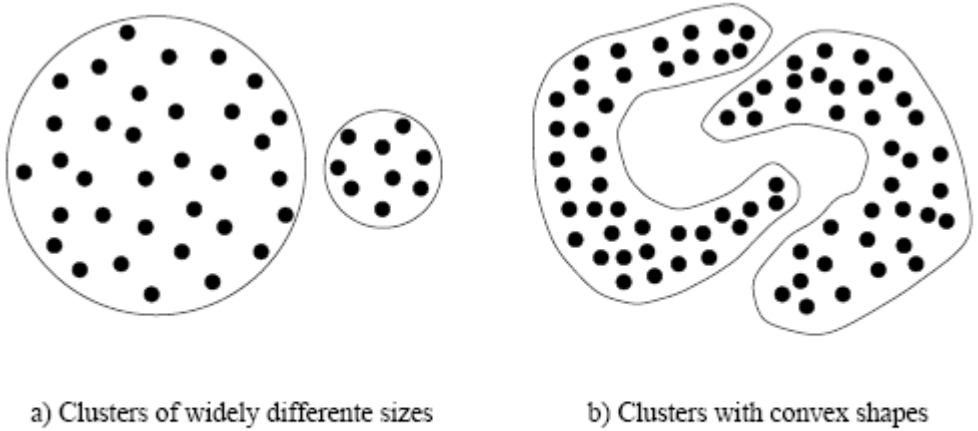
- COP-k-means
- CVQE







روش خوشه بندی برتر؟

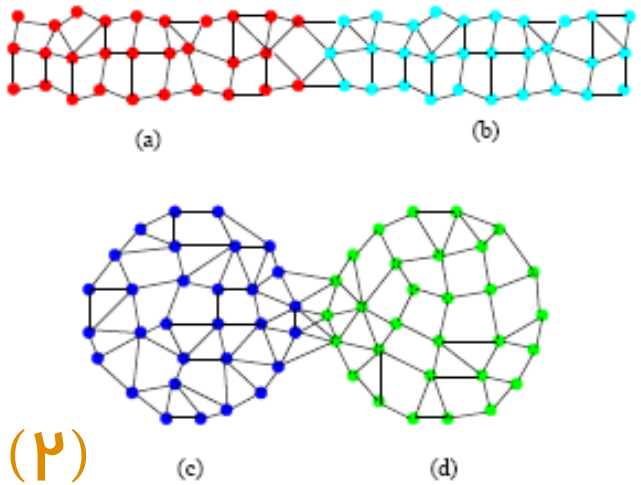


a) Clusters of widely differente sizes

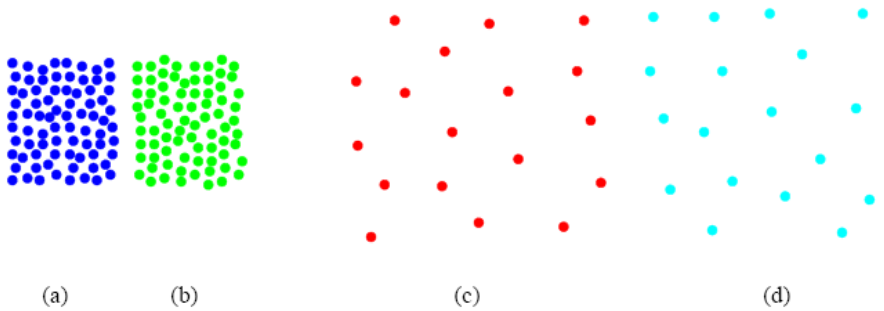
b) Clusters with convex shapes

(۱)

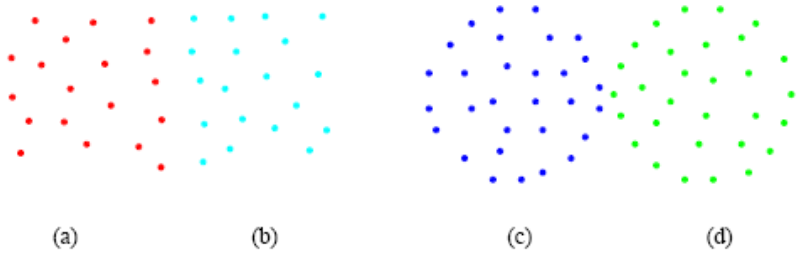
Data sets on which centroid and medoid approaches fail.



(۲)



(۳)



(۴)

ارزیابی خوشه‌ها

- داده ها خوشه پذیر هستند؟
- مجموعه استاندارد
- تست Hopkins

$$H = \frac{\sum_{i=1}^{n} q_i}{\sum_{i=1}^{n} q_i + \sum_{i=1}^{n} w_i}$$

w: فاصله آبجکت با همسایگان نزدیک

q: فاصله نزدیکترین آجکت ها با نقاط انتخاب شده

- اگر H کوچک باشد و نزدیک به ۰/۵ : یکنواخت
- اگر H نزدیک به یک باشد: ساختار خوشه بندی

ارزیابی خوشه‌ها

تعیین تعداد خوشه‌ها

بر اساس دانش قبلی

بر اساس تجربه

روش Elbow

اندازه گیری کیفیت خوشه‌ها

- ✓ همگن بودن
- ✓ کامل بودن
- ✓ Rag bag
- ✓ خوشه‌های کوچک

BCubed precision
BCubed recall

روش‌های بررسی نزدیکی

آیتم ها داخل خوشه‌ها

- ✓ کشف الگو
- ✓ رده بندی و پیش بینی
- ✓ خوشه بندی

اجرای پروژه داده کاوی

- معرفی چند نرم افزار کاربردی
- فرآیند CRIPS
- پروژه و ارزیابی
- پروژه عملی

معرفی نرم افزارهای کاربردی

نرم افزار Weka

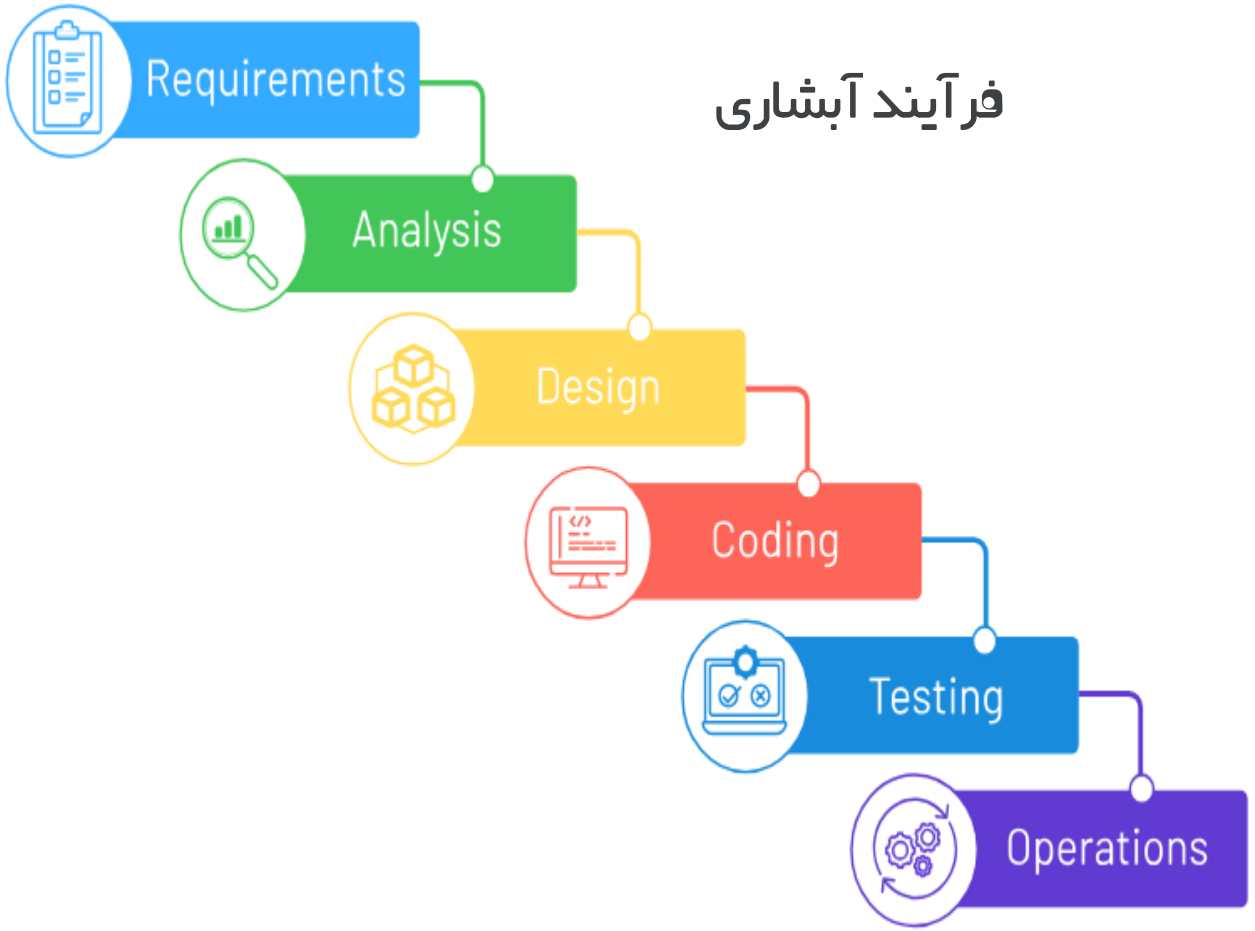
نرم افزار RapidMiner

نرم افزار Orange

نرم افزار KNIME

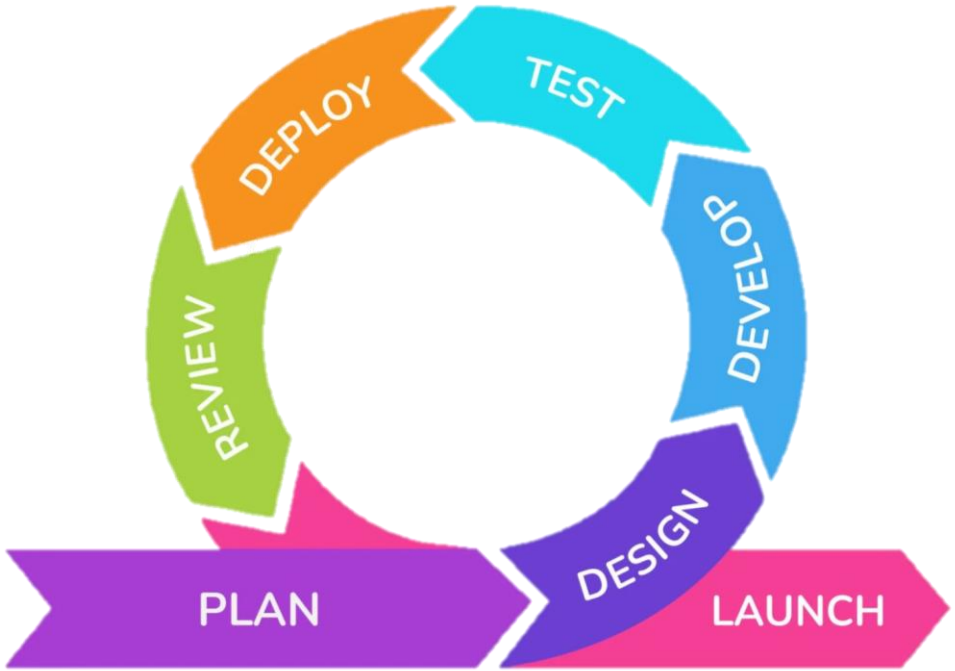
فرآیند کریسپ CRIPS

فرآیند آبشاری



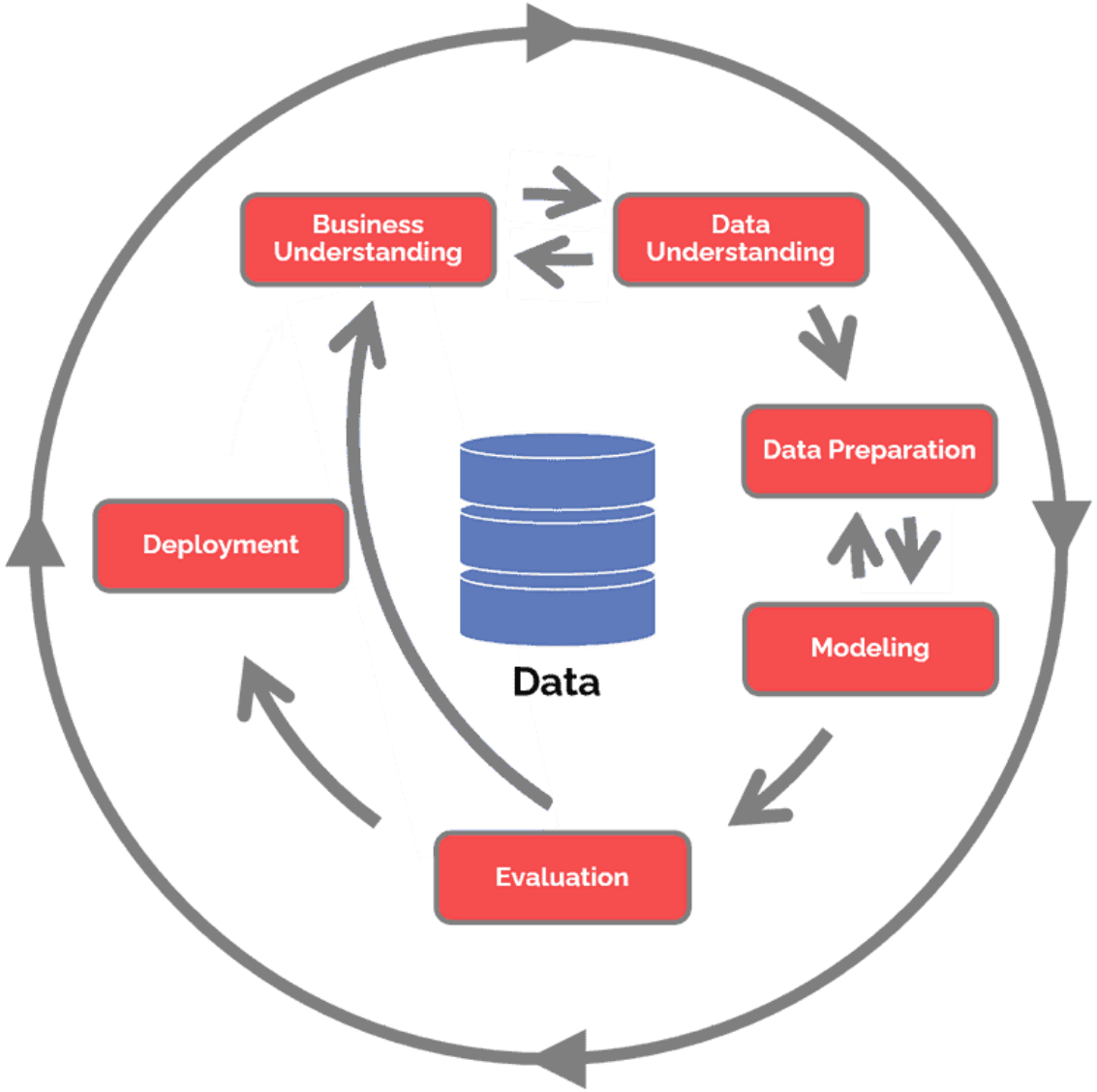
منبع تصویر: https://www.actitime.com

فرآیند چابک



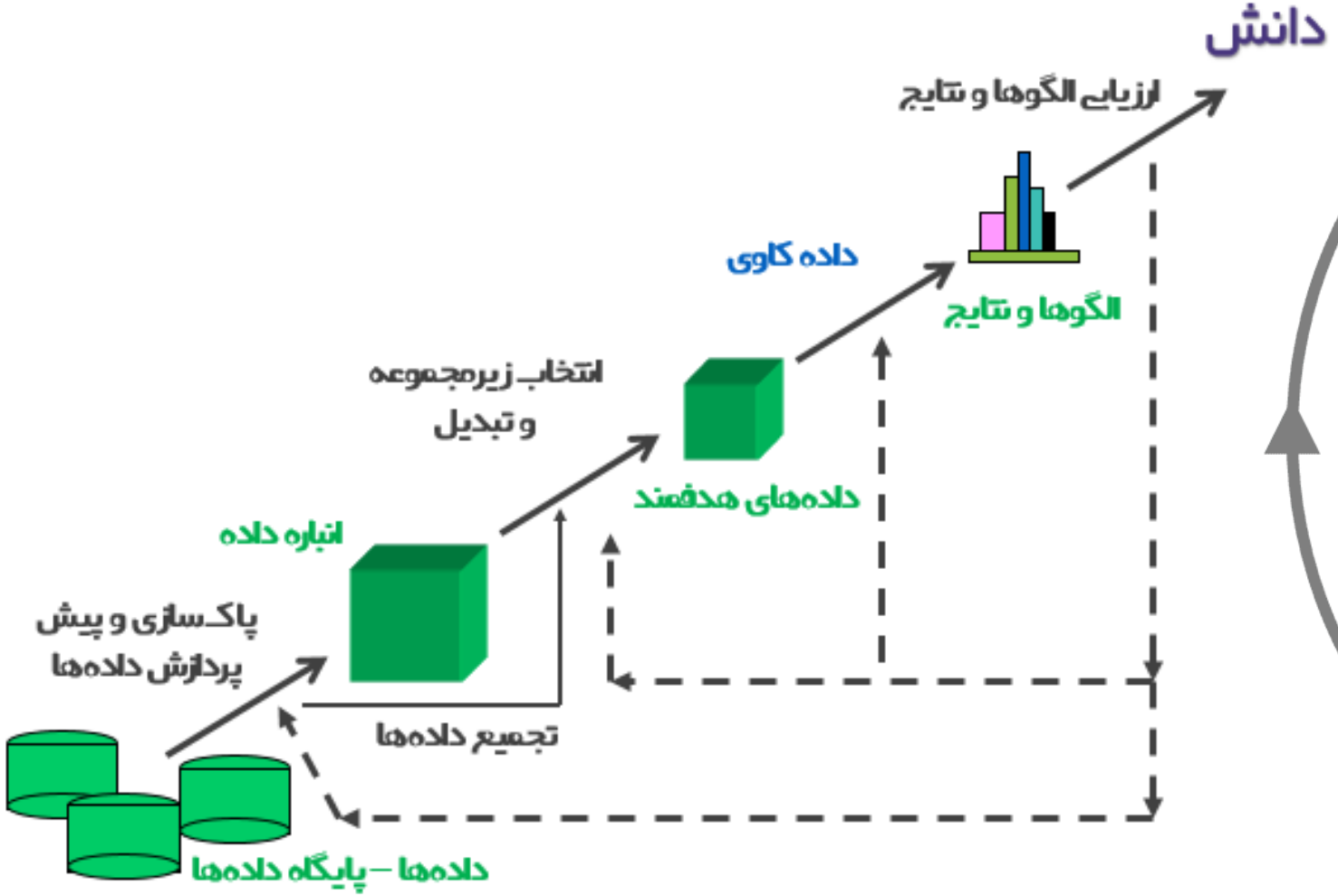
منبع تصویر: https://www.krasamo.com

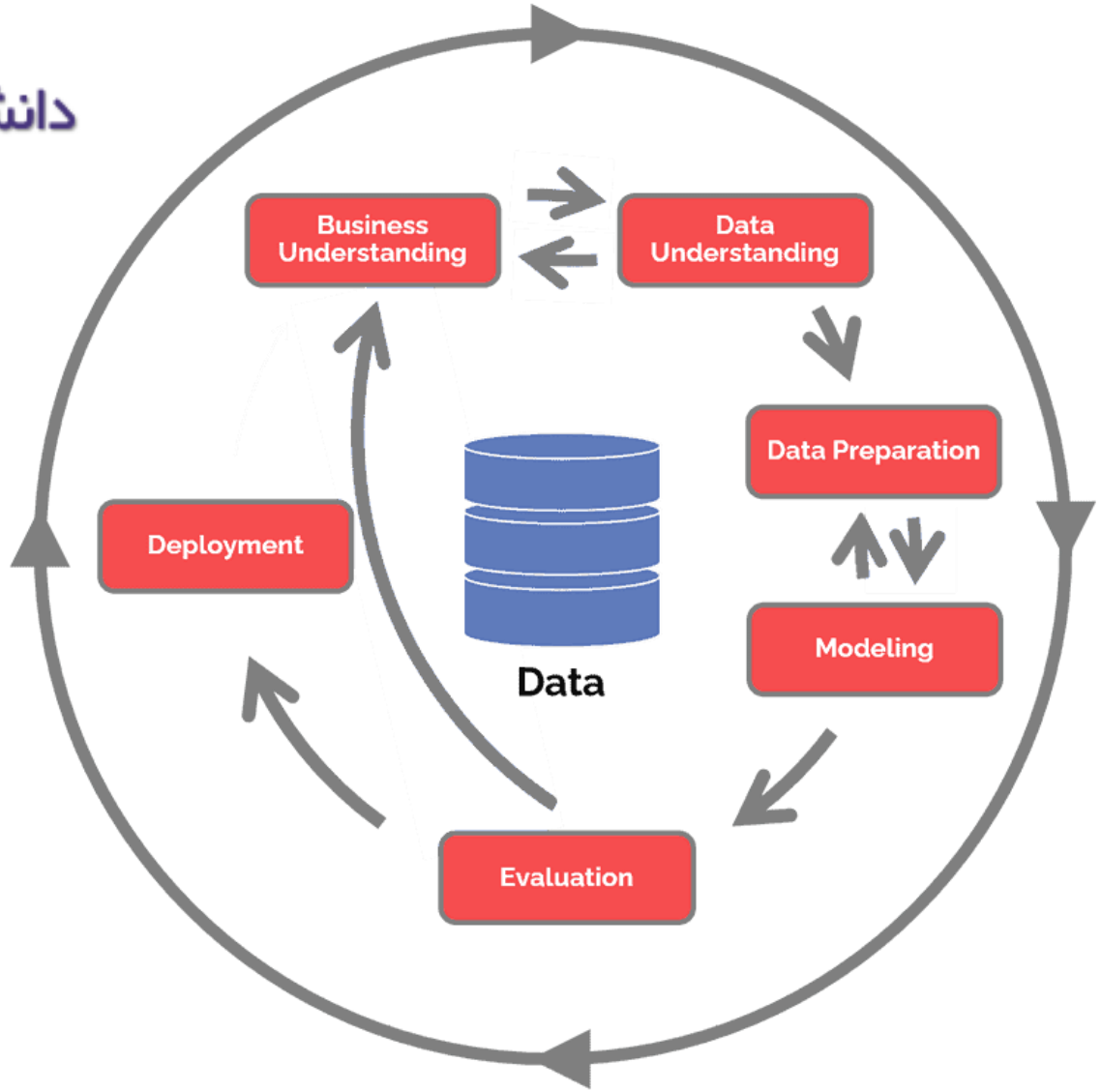
فرآیند کریسپ CRIPS



منبع تصویر: https://www.datascience-pm.com

۱. فهم کسب و کار (Business Understanding)
۲. فهم داده (Data Understanding)
۳. آماده سازی داده (Data Preparation)
۴. مدل سازی (Modeling)
۵. ارزیابی (Evaluation)
۶. پیاده سازی و انتشار (Deployment)


دانش
ارزیابی الگوها و نتایج
داده کاوی
الگوها و نتایج
انتخاب زیرمجموعه
و تبدیل
داده‌های هدفمند
انباره داده
پاک‌سازی و پیش
پردازش داده‌ها
تجمیع داده‌ها
داده‌ها – پایگاه داده‌ها


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

پروژه و ارزیابی

۱. دوبخش (حضور و مشارکت (۲۰درصد) + انجام پروژه (۸۰ درصد)

۲. بر اساس کریسپ، مرحله یک و دو را مستند نمایید و شرح دهید.

۳. داده‌های کثیف را در داده‌های در دسترس شناسایی و راهکاری برای تمیز شدن آن‌ها ارائه نمایید.

۴. پیش پردازش‌های لازم، تجمیع ها و تبدیل های مورد نیاز را همراه با علت شرح دهید.

۵. ده کاربرد از داده‌های در دسترس به همراه نحوه‌ی اجرا یا ایجاد مدل را شرح دهید.

۶. امتیازی: تولید مدل و ارزیابی مدل متناسب به نیازمندی

مهلت ارسال: ۳۰ دی ماه ۱۴۰۲ ارسال به ایمیل: m.amin@mollahoseini.ir

پروژه عملی

- موضوع: بررسی و پردازش داده‌ها با استفاده از نرم افزار رپیدماینر

دانلود و نصب نرم افزار رپیدماینر

https://download.ir/rapidminer-studio-developer/

بررسی کامپوننت ها در نرم افزار


پایان
مشکرم