Subject:
Year. Month. Date. ( )

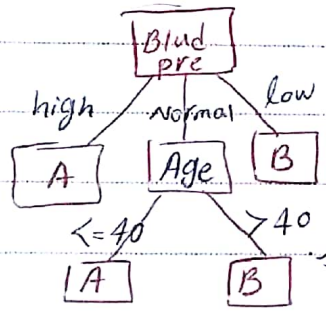
برای Age می‌توانیم Info حساب کنیم. چون دسته‌های آن زیاد است.
روش‌های گسسته سازی را گفتیم روی آن اعمال می‌کنیم
مثلاً دو دسته کوچکتر از 40 و بزرگتر از 40

روش‌های طبقه‌بندی: Classification
1. درخت تصمیم
2. روش بیزین
3. rule-based

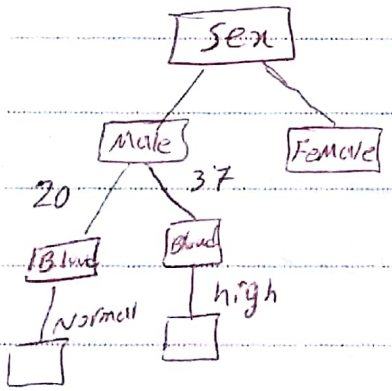
مثال:

ID	Sex	Age	Blood_Pre	Drug
1	Male	20	Normal	A
2	Female	73	Normal	B
3	M	37	High	A
4	M	33	low	B
5	F	48	H	A
6	M	29	N	A
7	F	52	N	B
8	M	42	l	B
9	M	61	N	B
10	F	30	N	A
11	F	26	l	B
12	M	54	H	A
13	M	56	Normal	B

مثال‌ها را تعیین می‌کنیم براین اساس می‌گوید بیمار داروی A بخورد یا داروی B می‌خواهیم مدلی درست کنیم که بگوید بیمار داروی A بخورد یا داروی B و برای بیماران جدید از آن برای پیش‌بینی و دسته‌بندی استفاده کنیم.



مثال یک درخت مینیمال هستیم که تمام این نمونه‌ها در آن بگنجند.
درخت تصمیم را می‌خواهیم مثلاً عمق درخت کم باشد
شاخه‌ی زیاد نباشد ← باید هیوریستیک داشته باشیم
در این درخت جنسیت را در نظر نگرفتیم چون تأثیری در جدا کردن ندارد.
حالا یک درخت تصمیم دیگر:



→ درخت بزرگ

این که چه ویژگی در هر سطح انتخاب بشود حریصانه است.
۳ تا معیار داشتیم (قبلاً گفته شد)

$$Info_{(D)} = -\left[\frac{6}{13}\log_2\frac{6}{13} + \frac{7}{13}\log_2\frac{7}{13}\right]$$

(A) (B)

$$Info = \frac{8}{13}\times -\left[\frac{4}{8}\log_2\frac{4}{8} + \frac{4}{8}\log_2\frac{4}{8}\right] + \frac{5}{13}\times -\left[\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right]$$

جنسیت: تعداد مردها | مردهایی که داروی A می‌خورند | مردهایی که داروی B می‌خورند | تعداد زن‌ها | زن‌هایی که داروی A می‌خورند | زن‌هایی که داروی B می‌خورند

Subject:
Year. Month. Date. ( )

ID	Age	Job	House	Credit	وام میدیم یا نه
1	old	F	T	Excellent	yes
2	old	F	T	Good	yes
3	Middle	F	F	Fair	NO
4	Middle	T	T	Good	yes
5	Young	F	F	Fair	No
6	old	F	F	Fair	NO
7	middle	F	T	Excellent	yes
8	young	T	F	Good	yes
9	Young	T	T	Fair	yes
10	middle	F	F	Good	NO

کل $Gini(D) = 1 - \left(\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2 \right) = 0.48$

$Gini(Job) = \frac{7}{10} \times \left[1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{4}\right)^2 \right] + \frac{3}{10} \times \left[1 - \left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2 \right]$
$= 0/343$

$Gini(House) = \frac{5}{10} \times \left[1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 \right] + \frac{5}{10} \left[1 - \left(\frac{0}{5}\right)^2 + \left(\frac{5}{5}\right)^2 \right]$
$= 0/16$

برای Credit ۳ تا دسته داریم نمی‌توانیم از Gini استفاده کنیم. Gini فقط دوتا دسته را محاسبه می‌کند. سن هم همینطور اما اینفورمیشن گین می‌تواند محاسبه کند.

House ← اولین شرط درخت

حالا برای اینکه بتوانیم از Gini استفاده کنیم باید هر کدام را به ۲ دسته تقسیم کنیم

مثلاً سن ← Middle+young و old — old,middle و young یا

هر بار بیرون دوتا را بگیریم با هم و Gini حساب کنیم ببینیم بهترین دسته‌بندی کدام است برای

credit هم همینطور

Subject:
Year. Month. Date. ( )

اول کار یک Dataset داریم باید عملیات زیر را انجام بدهیم:
Pre Processing ها براساس شاخص های مرکزی پراکندگی کارهای انجام می‌شود
و شامل ۴ مرحله زیر می‌باشد. { ۱ داده‌های ناقص داریم.
۲ داده‌های پرت و نویز

برای داده‌ی ناقص می‌توانیم آن‌ها را حذف کنیم
وقتی داده‌ی ما زیاد باشد بعضی حذف کردن تأثیری در کل دیتا ها ندارد آن را حذف می‌کنیم
یک راه دیگر این است که یک میانگین در نظر بگیریم و به جای داده‌ی ناقص بذاریم.
اینکه کدام راه را انتخاب کنیم. بستگی به مقدار داده‌ی ما دارد.
می‌توانیم میانگین یک دسته یا کلاس را قرار دهیم ← مثلاً از روی معدل و نمرات
می‌فهمیم که داده‌ی ما در این کلاس قرار دارد
اینکه میانگین را قرار دهیم یا چیزهای دیگر ← یک چیز تجربی است

داده‌ی پرت را حذف می‌کنیم که این هم تجربی است. هر جوری بستگی دارد

روش‌های تشخیص داده‌ی پرت: { ۱ توزیع نرمال
۲ IQR چارک
۳ روش تشخیص مبتنی بر فاصله برای جاهایی که کلاستربندی می‌کنیم خوب است.

۳ یک داده داریم و می‌خواهیم ببینیم پرت است یا نه؟
فاصله این دیتا با همین نمونه را حساب می‌کنیم. (با P نمونه)
اگر فاصله آن با همه‌ی داده‌ها بیشتر از یک مقدار threshold(d) باشد می‌گوییم
یک داده پرت است. d و P به صورت تجربی تعیین می‌شوند.

۳۵ داده

۱۰
۱۵
۲۰
۱۳
۷
.
.
.

۱۰ تا P نمونه > threshold(d)

Subject:
Year. Month. Date. ( )

مراحل آماده سازی داده:

مراحل preprocessing:
1. Data Cleaning ← 1. نقص Miss 2. پرت و نویز
2. Data Intigration
3. Data Transformation
4. Data Reduction

داده‌ی پرت:
- معنی نرمال
- IQR
- دسته بندی
- خوشه بندی
- رگرسیون

دسته بندی ⟷ حذف: حذف داده نویز و پرت
کاهش سایز داده

مثال:
{۳, ۲, ۱, ۵, ۴, ۳, ۱, ۷, ۵, ۳}
اول باید داده ها را مرتب کنیم.
{۱, ۱, ۲, ۳, ۳, ۴, ۵, ۵, ۷}
بعد از آن دسته ها را انتخاب کنیم
{۱, ۱, ۲} {۳, ۳, ۳} {۴, ۵, ۵, ۷}
بعد برای هر دسته یک شاخص می گذاریم یعنی به جای هر دسته شاخص را می گذاریم.
شاخص ← میانگین {۱, ۳ و ۱, ۳ و ۱, ۳} {۳, ۳, ۳} {۵, ۲۵ و ۵, ۲۵, ۵, ۲۵, ۵, ۲۵}
مد {۱, ۱, ۱} {۳, ۳, ۳} {۵, ۵, ۵, ۵}
مقادیر کناری دسته {۱, ۱, ۲} {۳, ۳, ۳} {۴, ۴, ۴, ۷}

مهم این است که چند تا دسته درست کنیم و طول هر دسته چقدر باشد؟

k ← تعداد دسته ها
n ← تعداد کل نمونه ها

$k = 1 + 3.3 \log n$

Min Error → | میانگین یا میانه دسته − مقادیر یک دسته |
میانه یا مد

به جا قدر مطلق ممکن است مجذور هم گفته شود.

Subject:
Year. Month. Date. ( )

مثال: {۶, ۸, ۱, ۲, ۹, ۲, ۲, ۸, ۱, ۵}

دسته بندی به شکل الف

الف - {۱, ۱, ۲} {۲, ۲, ۵} {۶, ۸, ۸, ۹}

{۱, ۱, ۱} {۲, ۲, ۲} {۸, ۸, ۸, ۸}

ب {۱, ۲, ۸}

Error = {|۱-۱| + |۱-۱| + |۲-۱| + |۲-۲| + |۵-۲|

+ |۶-۸| + |۸-۸| + |۸-۸| + |۹-۸| = ۷

دسته بندی به شکل ب

ب = {۱, ۱, ۲, ۲, ۲} {۵, ۶} {۸, ۸, ۹}

ب 2 5 8

Error = |1-2| + |1-2| + |2-2| + |2-2| + |2-2| + |5-5|

+ |6-5| + |8-8| + |8-9| = 4

Error کمتر شد پس دسته بندی به شکل ب بهتر است.

خوشه بندی: داده ای مشابه به هم را در یک دسته بندی قرار میدهیم و متغیر خروجی یا lable یا کلاس نداریم.

کاربرد ← مثلاً مشتری های شبیه به هم را در یک دسته قرار میدهیم.

مشتری جدید را در یکی از این دسته ها قرار می دهیم (شباهت بین آن ها پیدا می کنیم)

و پیش بینی کنیم مشتری جدید چه چیزهایی می خرد.

کاربرد دیگر ← پردازش تصویر، تشخیص الگو، تشخیص مشتری ها

اشیا شبیه به هم را در یک دسته قرار می دهیم.

مثلاً میوه های شبیه به سیب را در یک دسته قرار میدهد

اگر داده ها برچسب بود و به هم کدام از دسته ها شبیه بود به آن می فهمیم داده برچسب است

پس برای تشخیص داده های برچسب کاربرد دارد

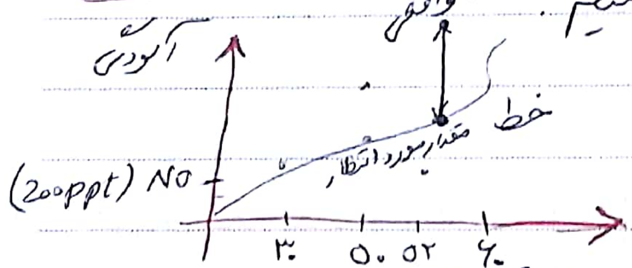
Subject:
Year. Month. Date. ( )

رگرسیون: برای اینکه مقدار یک متغیر را براساس متغیر دیگر پیش‌بینی کنیم.
مثلاً رابطه بین دما و آلودگی.

رگرسیون برای دو متغیری استفاده می‌شود که این دو متغیر با هم ارتباط داشته باشند
یعنی ارتباط مثبت یا ارتباط منفی اما ارتباط صفر نباشد (صفر یعنی مستقل)
پس حتماً بین این دو متغیر ضریب همبستگی داریم. (Corrolation - Covariance)

رگرسیون ←
- خطی
- غیرخطی → یک جوری تبدیل به خطی می‌کنند.
- چندگانه → رابطه بین یک متغیر و چندین متغیر
  پیش‌بینی مقدار یک متغیر براساس چندین متغیر

آلودگی	دما



می‌خواهیم رابطه بین آلودگی و دما را بررسی کنیم.
یک خط پیدا می‌کنیم. $y = \alpha x + \beta$
بعضی از داده‌ها روی خط نمی‌افتند
که به فاصله این داده‌ها تا خط Error
می‌گوییم. خطی را انتخاب می‌کنیم که Error نداشته باشد.

$y = \alpha_1 x_1 + \alpha_2 x_2 + \beta$ → رگرسیون چندگانه

$\alpha = \bar{y} + \beta \bar{x}$

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Year. Month. Date. ( )

یکپارچه سازی: Integeration

چندین دیتابیس داریم که می‌خواهیم آن‌ها را یکپارچه کنیم. مثلاً ویژگی قد در یک دیتابیس برحسب foot و در یک دیتابیس برحسب متر است. ← ناسازگاری

یا داده‌های تکراری داشته باشیم. پس در یکپارچه سازی هدف حذف افزونگی و ناسازگاری است ← تبدیل می‌شود به انبار داده.

حالا فرض کنیم یک انبار داده داریم. مثلاً داده‌های مابین ۵۰ تا ۵۵ هستند اما بعضی متدهای داده کاوی نیاز دارند که داده‌ها مابین [۰ و ۱] باشد ← نرمالایز

تغییر شکل داده: Transformation نرمالایز کردن یا استاندارد سازی

روش‌های استاندارد سازی:

① حرکت نقطه اعشار مثلاً data: 800 تا 3.76-

قدر مطلق بزرگترین data = 800 ≅ 1000 → 10^3

$\frac{\text{اعداد}}{\text{توان } 10} \rightarrow \frac{\text{همه اعداد}}{10^3}$

② $\frac{\text{عدد} - Min}{Max - Min}(NewMax - NewMin) + NewMin$

می‌خواهیم اعداد در بازه [Min, MAX] قرار بگیرند.
NewMin ← Min ، New MAX ← MAX

③ $\frac{\text{عدد} - \text{میانگین}}{\text{انحراف معیار}}$

معمولاً قبل از اعمال شبکه عصبی نرمال سازی انجام می‌دهیم.

data Reduction ← داده‌ها خیلی زیادند. یک دیتابیس خیلی بزرگ

حالا می‌خواهیم یک سری داده‌ها را حذف کنیم.

یا مثلاً داده‌های ما ویژگی‌های خیلی زیادی دارند. مثلاً دیتابیس از دانشجویان داریم آدرس - شماره دانشجویی - اسم - معدل و ... وقتی می‌خواهیم سطح علمی دانشجویان را حساب کنیم ویژگی‌های آدرس و اسم ...... تأثیر ندارند پس آن‌ها را حذف می‌کنیم.

حالا مثال دیگر می‌خواهیم تأثیر بینایی را بررسی کنیم. رنگ چشم - قهوه‌ای - قهوه‌ای روشن - قهوه‌ای تیره
قهوه‌ای خیلی تیره ← می‌توانیم دو دسته در نظر بگیریم ← قهوه‌ای روشن و قهوه‌ای تیره

۱- کاهش تعداد ویژگی‌ها
۲- کاهش تعداد نمونه‌ها ←
۳- کاهش مقادیر یک ویژگی

۱- انتخاب تصادفی که می‌تواند با جایگذاری یا بدون جایگذاری باشد
۲- نظم ← شلغ از ۱۰ تای اول یکی را انتخاب کن از ۱۰ تای دوم یکی
۳- براساس کلاس‌های مختلف انتخاب کنیم.
شلغ می‌خواهیم از دخترها و پسرها به یک تعداد انتخاب کنیم
← جنسیت در بررسی‌ها اثر می‌گذارد.

مثل تبدیل رنگ چشم به ۳ دسته

کاهش مقادیر یک ویژگی: فرض کنید سن از ۱۳۰-۰ سال باشد
می‌خواهیم دسته‌بندی کنیم. دو تا روش داریم
گسسته سازی کنیم ← روش‌ها گسسته سازی اند. χ^2 کای دو ← آنتروپی

ID	ویژگی سن	class نوع
1	1	A
2	3	B
3	7	A
4	8	A
5	9	A
6	11	B
7	23	B
8	37	A
9	39	B
10	45	A
11	46	A
12	58	A

این ویژگی‌ها خاصیت نوع کلاس می‌تواند دختر یا پسر باشد.
خودمان شروع به دسته‌بندی می‌کنیم.
تعداد این دسته‌ها زیاد است.
بعضی از دسته‌ها را می‌توان ادغام کرد.
می‌خواهیم ببینیم چطوری می‌توان این دسته‌ها را باهم ادغام کرد.

[0,2) [2,5) [5,7.5) [7.5,8.5) [8.5,10)
[10,17) [17,30), ...

اسم کلاس‌ها را می‌نویسیم
اسم دسته‌ها را می‌نویسیم
نگاه می‌کنیم
از این کلاس در این دسته چند نفر هستند.

	A	B	
[7.5,8.5)	$A_{11}=1$	$A_{12}=0$	→ جمع = $R_1=1$
[8.5,10)	$A_{21}=1$	$A_{22}=0$	→ جمع = $R_2=1$
	جمع $C_1=2$	$C_2=0$	$N=2$

بعد این‌ها را سطری و ستونی باهم جمع می‌کنیم.

Subject:
Year. Month. Date. ( )

فرمول کای دو:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} (A_{ij} - E_{ij})^2 / E_{ij}$$

- i: دو تا دسته داریم
- j: تعداد نمونه‌ها در تا. اصلی دسته و ل - امین کلاس
- A_{ij}: فراوانی مورد انتظار

می‌خواهیم ببینیم دو تا دسته را با هم ادغام کنیم یا نه؟

N: تعداد کل نمونه‌ها

$$E_{ij} = R_i \times C_i / N$$

درجه آزادی = تعداد کلاس‌ها − 1 ⟵ درجه آزادی = 1 ← در این مثال

یک جدول کای دو داریم که بر اساس درجه آزادی مقدار کای دو در آن نوشته شده

$$\chi^2 = (1-1)^2/1 + (0-0/1)^2/0/1 + (1-1)^2/1 + (0-0/1)^2/0/1$$

$$= 0/2 < 2.706$$ ← مقدار χ^2 بر اساس جدول

چون مقدار χ^2 ما کمتر از مقدار χ^2 که باید باشد شد پس این دو دسته را ادغام می‌کنیم. اینقدر این کار را برای دسته‌ی مختلف تکرار می‌کنیم تا به این نتیجه برسیم هیچ دو دسته‌ای با هم ادغام نمی‌شوند.

(7.5, 10)

برای این مثال ←

$E_{11} = 1$ $\quad E_{12} = 0 \Rightarrow 0/1$ (صفر نباید باشد)

$E_{21} = 1$ $\quad E_{22} = 0 \Rightarrow 0/1$

گسسته‌سازی متنی بر آنتروپی: مجموعه D را به دو مجموعه D_1 و D_2 می‌شکنیم حالا می‌خواهیم ببینیم این تقسیم‌بندی خوب است یا نه؟

D_1 | D_2

SP نقطه انفصال

$$Info_A(D) = \frac{|D_1|}{|D|} \times Entropy(D_1) + \frac{|D_2|}{|D|} \times Entropy(D_2)$$

Subject:
Year. Month. Date. ( )

احتمال کلاس k-ام در D

$$Entropy = -\sum_{i=1}^{k} P_i \log_2 P_i$$

مثلاً بازه سنی از ۱ تا 59 است
تا آن را به دو دسته تقسیم کنیم
از ۱ تا 30 و از 31 تا 59
30 نقطه انفصال می‌خواهیم ببینیم آیا 30 نقطه خوبی است؟

خب ← حذف ستون‌ها (حذف بعضی از ویژگی‌ها) خوب است و پرکاربرد
کاهش حجم نمونه هم بله به کار میده اما کاربردی تر از هم حذف ویژگی‌ها است
اما روش دسته بندی چندان مهم نیست ← کاربردی نیست (کاهش مقادیر یک ویژگی)

حذف بعضی از ویژگی‌ها ← واضح است که بعضی از ویژگی‌ها در تعیین خروجی یا پیش بینی‌ها اثری ندارند.

ویژگی X	ویژگی Y	نوع کلاس
3	7	A
2	9	B
6	6	A
5	5	A
8	7	B
4	9	A

می‌خواهیم بر اساس ویژگی X و ویژگی Y نوع کلاس را مشخص کنیم
می‌خواهیم ببینیم برای تعیین کلاس آیا هر دو ویژگی را نیاز داریم یا با یک ویژگی می‌توانیم نوع کلاس را تشخیص دهیم.

$$\text{فرمول} = \frac{|\text{میانگین } A - \text{میانگین } B|}{\sqrt{\frac{\text{واریانس } A}{N_1} + \frac{\text{واریانس } B}{N_2}}}$$

N_1 و N_2 تعداد هر کدام

برای ردیف x این فرمول را جدا و برای ردیف y این فرمول را جدا حساب می‌کنیم.

x هایی که در کلاس A هستند: $X_A = \{3, 6, 5, 4\}$ $Mean(X_A) = 4.5$ $Var(X_A) = 1.25$

x هایی که در کلاس B هستند: $X_B = \{2, 8\} \rightarrow Mean(X_B) = 5$ $Var(X_B) = 9$

$$\text{فرمول} = \frac{4.5 - 5}{\sqrt{\frac{1.25}{4} + \frac{9}{2}} = 2.19} = ۰/۲۲۷۹ < ۰/۵$$

فاصله منهتن : $d(O_1,O_2) = |4-7| + |5-8| + |4-3| + |8-1| = 4$

$d(O_2,O_3) = 12$ $\quad$ $d(O_1,O_3) = 8$

$$\sqrt[h]{\sum_{i=1}^{k} (x_{ik} - x_{jk})^h}$$

مینوکوفسکی : برای h ها مختلف
ای h را بیشتر از 2 می‌گیرند معمولاً.

فاصله همسایگی متقابل : Mutual Neighbor Distance (داده‌کاوی)
فرض کنید چندنمونه داریم که دارای دو ویژگی x و y هستند

	X	Y
A	4	2
B	6	4
C	8	8
D	3	2
E	3	1
F	2	2

فرض کنید می‌خواهیم فاصله بین A و C را بدست آوریم.

$MND = NN(A,C) + NN(C,A) = 2+5 = 7$

$NN(A,C) = 2$ اول شباهت ~~A~~ C با
$d(A,C) = 7.21$ می‌نویسیم بعد می‌شماریم A در رتبه
$d(B,C) = 4.47$ همینجا C شباهت دارد. اولین
$d(D,C) = 7.81$ شباهت را B دارد و بعد A.
$d(E,C) = 8.60$
$d(F,C) = 8.48$

$NN(C,A) = 5$ حالا شباهت A با C را بدست میاریم.
$d(A,B) = 2.83$ اول همه‌ها A می‌نویسیم یعنی A با همه
$d(A,C) = 7.21$ حالا C با A رتبه 5ام را دارد
$d(A,D) = 1.00$
$d(A,E) = 1.41$
$d(A,F) = 2.00$

Year. Month. Date. ( )

باینری - دوحالته

معیارهای تشابه برای داده‌های باینری:

	قد	وزن	lips	Hair	Hand	Gender
O_1	155 0	65 0	thin 0	curly 1	R 1	F 0
O_2	172 1	75 1	thick 1	straight 0	R 1	M 1
O_3	162 0	68 1	thin 0	curly 1	R 1	M 1
O_4	168 1	62 0	thick 1	curly 1	R 1	F 0
O_5	175 1	80 1	thick 1	curly 1	L 0	M 1

دوتا از ستون‌ها که نا باینری هستند آن‌ها را باینری می‌کنیم.

قد: $\geq 165 \Rightarrow 1$, $< 165 \Rightarrow 0$

وزن: $> 70 \Rightarrow 1$, $< 70 \Rightarrow 0$

حالا می‌خواهیم تشابه نمونه‌های ۳ و ۴ را پیدا کنیم.

O_4 \ O_3	1	0	
1	a=2	b=2	4
0	c=1	d=1	2
	3	3	6

$$\text{تشابه} = \frac{a+d}{a+b+c+d}$$

تاکید مضاعف بر تعداد حالت‌های یکسان $= \dfrac{2(a+d)}{b+c+2(a+d)}$ یعنی شباهت‌ها را وزن 2 بدهیم.

تاکید بر حالت‌های عدم همسان $= \dfrac{a+d}{2(b+c)+a+d}$

رابطه Dissimilarity و Similarity:

$$Sim = \frac{1}{1+Dis} \qquad D = \sqrt{2(1-Sim)} \qquad S = 1 - Dis$$

بعد از پیدا کردن تشابه، خوشه‌بندی می‌کنیم.

Subject:
Year. Month. Date. ( )

این مقدار که بیشتر آمد آن را با یک مقدار ترشولد مقایسه می‌کنیم که این مقدار آستانه را خودمان مشخص می‌کنیم که اینجا گفتیم ۰/۵ باشد.
حالا که این مقدار کمتر از ۰/۵ باشد آن ویژگی را حذف می‌کنیم. یعنی مقدار x در تعیین نوع کلاس نقشی ندارد و با ویژگی y نوع کلاس تعیین می‌شود.

ویژگی

$$Y_A = \{7, 6, 5, 9\} \Rightarrow Mean(Y_A) = 6.75 \quad Var(Y_B) = 2.20$$

$$Y_B = \{9, 7\} \Rightarrow Mean(Y_B) = 8 \quad Var(Y_B) = 1$$

$$\frac{6.75 - 8}{\sqrt{\frac{2.20}{4} + \frac{1}{2}}} = 1.2198 > 0.5 = 1.0422$$

حذف بعضی از ستون‌ها ↔ بررسی تمام ویژگی‌ها صورت مجزا
روش آنتروپی

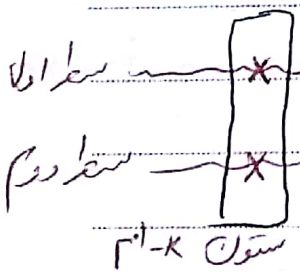
روش آنتروپی: می‌خواهیم ببینیم دو تا سطر از دیتاست چقدر با هم تفاوت دارند.

$$S_{ij} = e^{\alpha D_{ij}} \qquad \alpha = 0.5$$

α به صورت تجربی تعیین می‌شود اما ۰/۵ بهترین حالت است.

برای داده‌های عددی

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk}) / (max_k - min_k)^2}$$



مقدار این ستون − مقدار این ستون / max ستون − min ستون

برای داده‌هایی که عددی نیستند $S_{ij} = \left(\sum_{k=1}^{n} \left(|x_{ik} = x_{jk}| \right) / n \right)$

Year. Month. Date. ( )

$$Enthropy = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[S_{ij} \times \log S_{ij} + (1 - S_{ij}) \times \log (1 - S_{ij}) \right]$$

با استفاده از روش‌های AI هوش مصنوعی این نوع آنتروپی را تابع هدف قرار می‌دهیم هرچه آنتروپی کمتر باشد بهتر است.

Clustering: می‌خواهیم داده‌ها را شبیه با دیگر دسته قرار دهیم

متغیر خروجی یا label نداریم.

باید تشابه و عدم تشابه را حساب کنیم.

فرض کنید نمونه‌ای 1 ... n داریم.

ماتریس تشابه اینا تفاوت‌ها را در آن می‌نویسیم!

هر نمونه با خودش تفاوت ندارد ← عدم تشابه = 0

یک طرف قطر را حساب می‌کنیم

داده‌ای با بیزری و اسمی - عددی

نمونه‌ها / ویژگی‌ها	Att1	Att2	Att3	Att4
O_1	4	5	1	8
O_2	7	8	3	1
O_3	3	4	5	3

متغیرها ویژگی‌ها از نوع عددی

روش اقلیدسی = فاصله بین هر دو نمونه

$$\sum_{i=1}^{k} \sqrt{(x_{iK} - x_{jK})^2}$$

$$d(O_1, O_2) = \sqrt{(4-7)^2 + (5-8)^2 + (4-3)^2 + (8-1)^2} = 8.25$$

$$d(O_2, O_3) = 6.32$$

هرچقدر این عدد کمتر باشد یعنی فاصله کمتر است یعنی تشابه بیشتر است

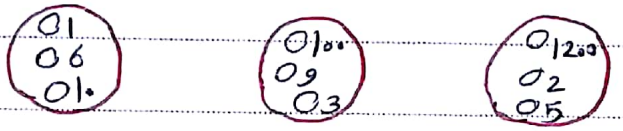
$$d(O_1, O_3) = 5.29$$

Subject:
Year. Month. Date. ( )

3 K-Mean

می‌خواهیم نمونه‌ها را در K تا دسته بندازیم. و n تا رکورد داریم.
از این n تا K تا را به طور تصادفی انتخاب می‌کنیم و هرکدام را در یک دسته می‌ذاریم.
O_2 را می‌خواهیم در یک دسته بذاریم.
شباهت O_2 را با هرکدام حساب می‌کنیم
و به هرکدام شبیه‌تر بود در همان دسته قرار میدهیم.
تا آخر نمونه‌ها همه‌ی نمونه‌ها مقایسه می‌شوند و دسته‌بندی می‌شوند.
وقتی همه نمونه‌ها دسته‌بندی شدند. بعد نظری میانگین هر دسته را می‌ذاریم نماینده آن دسته.

روش K-Mode هم داریم.



می‌خواهیم دسته‌ها را ببریم به سمت همین کردن K.
وقتی نماینده‌ها عوض کردیم دوباره همین کار را از اول انجام می‌دهیم.
آخر زمانی این کار را ادامه میدهیم که نمونه‌ها در دسته‌ها جابه‌جا نشوند.

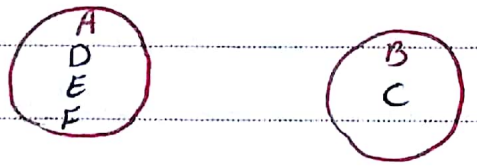
مثال: $K=2$

	X	Y
A	4	2
B	6	4
C	8	8
D	3	2
E	3	1
F	2	2

اول ماتریس شباهت را می‌سازیم:

	A	B	C	D	E	F
A	0					
B	2.80	0				
C	7	4.47	0			
D	1	3.60	7.80	0		
E	1.40	4.24	8.66	10.0	0	
F	2	4.47	8.48	1	1.40	0

مرحله اول:



میانگین ← میانگین دسته اول را می‌گیریم روی x
میانگین دسته اول را می‌گیریم روی y

$$mean_{1x} = \frac{4+3+3+2}{4} = 3$$

$$mean_{1y} = \frac{2+2+1+2}{4} = 1.75$$

$$Mean_{2x} = \frac{6+8}{2} = 7$$

$$Mean_{2y} = \frac{4+8}{2} = 6$$

Subject:
Year. Month. Date. ( )

x y	x y
(3, 1.75)	(7, 6)
A	B
D	C
E	
C	

دسته‌ی جدید:

فاصله اقلیدسی:

d(A, Mean1) = 1.03   d(A, mean2) = 5

d(B, Mean1) = 3.75   d(B, Mean2) = 2.23

در این افراز K می‌گیرند یعنی به K تا دسته تقسیم کنیم

چون میانگین تحت تاثیر داده‌ی پرت است از میانه و مد استفاده می‌کنیم

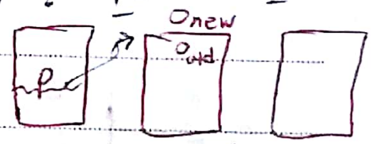
K-medoids اول K تا دسته درست کن K تا نمونه را به صورت تصادفی در این دسته‌ها قرار می‌دهد و n-K تای دیگر بین این K دسته تقسیم می‌کند. تفاوت با K-Mean در انتخاب نماینده دسته. برای انتخاب نماینده جدید یک نمونه از یکی از دسته‌ها به صورت تصادفی انتخاب می‌کند و به عنوان نماینده جدید برای دسته دیگر قرار می‌دهد.

برای هر دسته یک نماینده جدید و یک نماینده قدیم داریم. برای اینکه ببینیم نماینده جدید بهتر شده یا نه یک متد داریم به نام متد خطا.

$$\text{متد خطا} = \sum_{i=1}^{K} \sum_{p \in C_i} (p - O_i)$$

(خوشه i-ام) (O_i: نماینده هر دسته)

(هر دسته یک خوشه است)



دو تا نکته: برای تعیین نماینده جدید O_{new} به جای نماینده کنونی O_{old} یک خوشه

عملیات زیر باید برای هر نمونه غیر نماینده مثل p انجام شود

اگر p متعلق به خوشه‌ای باشد که O_{old} نماینده آن خوشه باشد، چنانچه با جایگذاری O_{new} به جای O_{old} فاصله p به نماینده جدید نزدیکتر شود پس p را در خوشه به نمایندگی O_{new} می‌گذاریم. در غیر این صورت p را در خوشه i قرار دهید به نحوی که فاصله نماینده خوشه i با p حداقل باشد و $old \neq i$ است.

Subject:
Year. Month. Date. ( )

اگر P متعلق به خوشه‌ای باشد که O_i نماینده آن است چنانچه با جایگزینی O_{new} به جای O_{old} همچنان فاصله P با نماینده خوشه خود یعنی O_i حداقل است هیچ تغییری صورت نمی‌گیرد در غیر این صورت P در خوشه‌ای قرار می‌گیرد که نماینده آن O_{new} است.

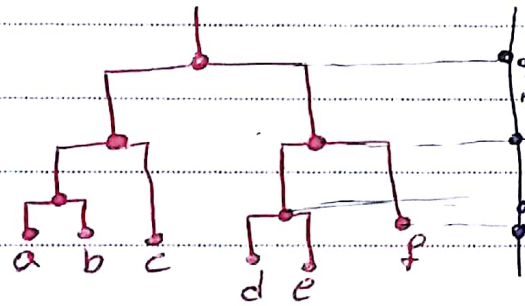
n: تعداد عضو K: تعداد دسته‌ها t: تعداد تکرار الگوریتم

مرتبه زمانی الگوریتم: $O(nkt) \cong O(n)$

سلسله مراتبی: در روش بالا به پایین یک دسته ضعیف بزرگ داریم و بعد از آن دسته جدید را می‌شکنیم.

یا از پایین به بالا اول تعداد زیادی دسته داریم و دسته‌هایی که می‌توان ادغام کرد را ادغام می‌کنیم.

یک محور عمودی کنار این نمودار می‌کشیم که هر عدد نشان دهنده دسته‌ای می‌شود.



با دو مفهوم تقسیم کردن و ادغام کردن روبرو هستیم.

معمولاً از روش پایین به بالا استفاده می‌کنند. علاوه بر تشابه بین دو عضو تشابه بین دو خوشه را بررسی می‌کنیم.

مثال:

	x	y
O_1	1	2
O_2	5	2
O_3	6	1
O_4	1	1
O_5	4	1

ماتریس تشابه:

	O_1	O_2	O_3	O_4	O_5
O_1	0				
O_2	4	0			
O_3	5.1	1.4	0		
O_4	1.0	4.1	5.0	0	
O_5	3.1	1.4	2.0	3.0	0

$\{O_1\}$ $\{O_2\}$ $\{O_3\}$ $\{O_4\}$ $\{O_5\}$

انتخاب دو عضو با بیشترین تشابه:

O_1 و O_4 بیشتر از همه با هم شبیه‌اند.

Subject:
Year. Month. Date. ( )

دوباره ماتریس تشابه را بروزرسانی می‌کنیم.

	$\{O_1,O_4\}$	O_2	O_3	O_5
$\{O_1,O_4\}$	0			
O_2	4	0		
O_3	5	1.4	0	
O_5	3	1.4	2	0

بیشترین تشابه را کدام دارد؟

حال می‌توانیم بگوییم که O_1 و O_4 را با هم بگیریم

هر بار دو تا از آن‌ها را با هم بگیر.

$$d(C_i, C_j) = \min\{d(O_1, O_2)\}$$

$$= d(C_K, C_i \cup C_j) = \min\{d(C_K, C_i), d(C_K, C_j)\}$$

$$d(O_2, \{O_1,O_4\}) = \min\{d(O_1,O_2), d(O_4,O_2)\} = \min\{4, 4.1\} = 4$$

	$\{O_1,O_4\}$	$\{O_2,O_3\}$	$\{O_5\}$
$\{O_1,O_4\}$	0		
$\{O_2,O_3\}$	4	0	
$\{O_5\}$	3	1.4	0

یعنی O_2 و O_3 و O_5 باید ادغام شوند.

بیشترین تشابه.

	$\{O_1,O_4\}$	$\{O_2,O_3,O_5\}$
$\{O_1,O_4\}$	0	
$\{O_2,O_3,O_5\}$	3	0