ترم ۱ سال ۱۳۹۵

داده‌کاوی

میثم مدنی
دانشگاه شهید بهشتی
فصل سوم
www.madani.pro
me@madani.pro

چشم‌انداز

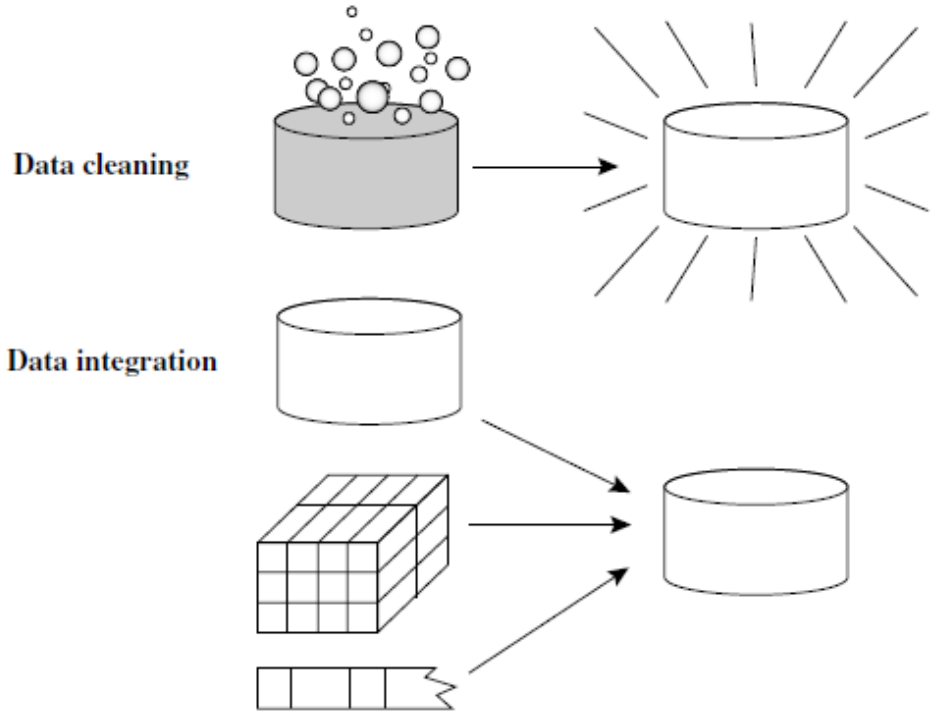
۱ پیش پردازش داده

پیش‌پردازش داده

- داده‌ها در دنیای امروز با توجه به عظمت و سرعت دریافت: نویزدار، متناقض و نامفهوم هستند.
- کیفیت پایین اطلاعات موجب سخت شدن فرایند کشف دانش خواهد شد.
- روش‌های متعددی وجود دارند که سعی در بهبود کیفیت داده‌ها را دارند. به این روش‌ها پیش‌پردازش داده می‌گویند.

۱ **پاکسازی**: نویز را از بین می‌برد و تناقض را در حد امکان به حداقل می‌رساند.
۲ **جمع‌آوری داده‌ها**: به نحوی منسجم، داده‌ها را از منابع مختلف یکجا کرده و آن را مانند یک انبار مناسب ذخیره می‌کند.
۳ **کاهش داده:** اندازه و تعداد داده‌ها را کاهش می‌دهد برای نمونه با حذف یا یکی کردن خصیصه‌های زاید یا خوشه بندی
۴ **تبدیل داده:** نرمال‌سازی داده

پیش‌پردازش داده



مروری کلی برپیش‌پردازش داده

کیفیت داده

کیفیت داده یک مفهوم مهم در داده‌کاوی است که با پنج شاخص زیر اندازه‌گیری می‌شود:

۱. دقت
۲. کامل بودن
۳. سازگاری
۴. تناسب
۵. معقول بودن
۶. قابل تفسیربودن

مثال AllElectronics

- ممکن است برخی داده‌ها وارد نشده باشند.
- کاربران در وارد کردن داده‌ها ممکن است اشتباهاتی کرده باشند.
- در برخی تراکنش‌ها اشتباهاتی وجود دارد که باعث تناقض می‌شود.

شما با داده‌هایی مواجه هستید که

۱. **کامل نیستند:** مقادیر و خصیصه‌های اصلی مورد نظر وجود ندارند.
۲. **نادقیق و نویزدار هستند:** شامل خطا هستند و مقادیری دارند که از مقدار مورد انتظار انحراف دارند.
۳. **متناقض و ناهماهنگ:** اختلافاتی در مورد واحد فروش و کد کالاها وجود دارد.

به دنیای واقعی خوش آمدید!

برخی نکات

- کمبود برخی داده‌ها به چند علت می‌تواند باشد مثلا توضیحات کالا در برخی موارد وجود ندارد.
  یا فراموشی در وارد کردن داده‌ها!

پاکسازی داده

در مثال AllElectronics بسیاری از بردارها ممکن است مقدار ثبت شده‌ای در آنها وجود نداشته باشد. روش‌های مختلفی وجود دارد که می‌توان با این داده‌های مفقوده رفتار کرد:

۱ **فراموش کنید:** این کار معمولا زمانی است که کل یا قسمت عمده ای از یک خصیصه وارد نشده است. این روش کارا نیست.

۲ **مقادیر مفقوده را دستی خودتان وارد کنید:** این روش زمانبر است و در حالتی که داده‌ها زیاد باشند عملی نیست.

۳ **از یک مقدار ثابت برای تمام مقادیر استفاده کنید:** همه مقادیر را با مقداری شبیه unknown یا $\infty-$ پر کنید. در اینصورت برنامه داده‌کاوی ممکن است مقدار unknown را پر اهمیت جلوه دهد! پس اگرچه ساده است ولی خیلی کارا نیست.

۴ **از یک اندازه مرکزی برای آن خصیصه استفاده شود:**

۵ **از مورد قبل ولی برای کلاس‌ها به صورت متفاوت استفاده شود:** مثلا برای مشتریان کلاس ۱ از میانگین خودشان و مشتریان کلاس ۲ نیز میانگین مربوط به خودشان.

۶ **از مقداری که بیشترین احتمال وقوع را دارد استفاده کنید:** استفاده از رگرسیون، ابزارهای مخصوص، فرمول بیزی یا درخت تصمیم

مقادیر نویزدار

نویز تفاوت یا خطایی تصادفی است که در یک مقدار اندازه گیری شده ممکن است به وجود آید. فرض کنید در مثال فروشگاه می‌خواهیم نویز را به اصطلاح هموار کنیم.
روشهای هموار کردن به شرح زیر است.

۱ - روش جعبه ای: داده‌ها را ابتدا مرتب می‌کنیم و سپس آنها را به بسته‌ها یا جعبه‌های شبه‌مساوی تقسیم می‌کنیم

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

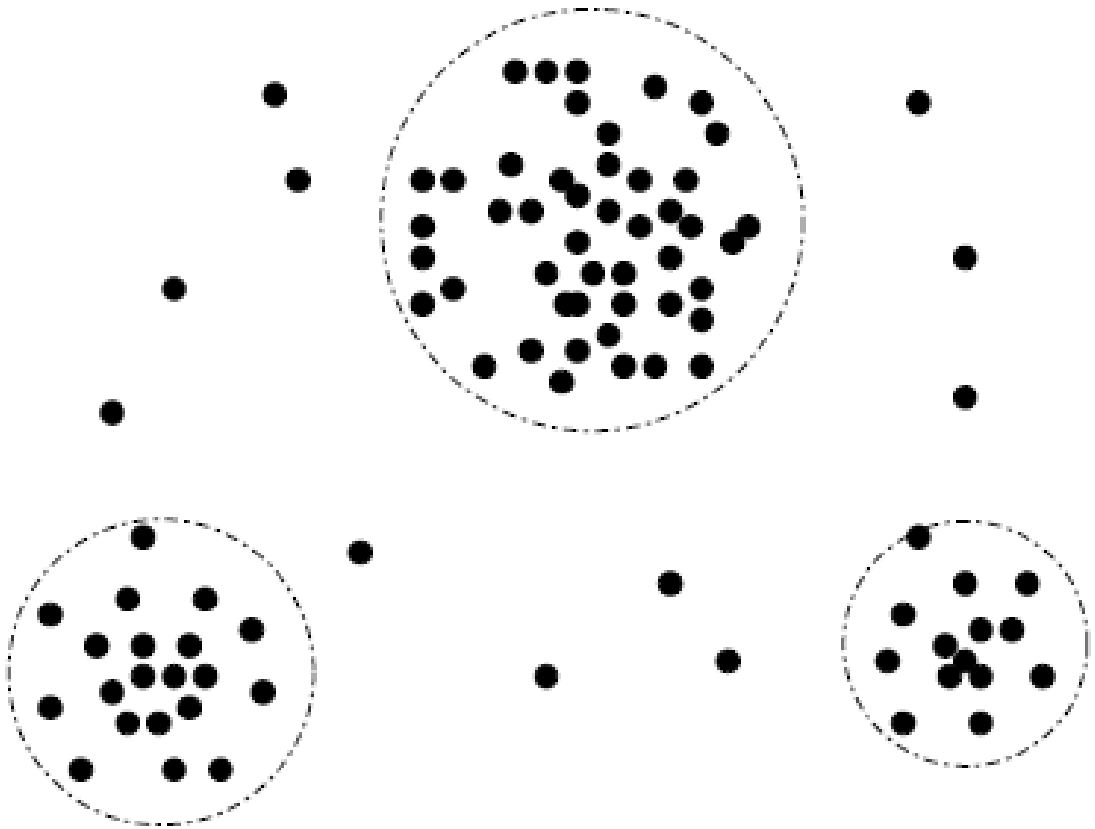
Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

۱. در روش میانگین جعبه‌ها: هر داده در هر جعبه با میانگین آن جعبه جایگزین می‌شود.
۲. در هموار کردن با میانه: هر داده در هر جعبه با میانه آن جعبه جایگزین می‌شود.
۳. در هموار کردن با کران‌ها: هر داده در هر جعبه با نزدیکترین کران آن جعبه جایگزین می‌شود.

مقادیر نویزدار

۲- رگرسیون: در این روش مقادیر داده‌ها را به یک تابع مشخص تقریب می‌زنیم.

۳- تحلیل برون هشته outlier analysis مقادیر پرت را می‌توان با خوشه بندی تشخیص داد، برای مثال مقادیری که خارج خوشه‌ها می‌افتند به عنوان داده پرت محسوب می‌شوند.



پاکسازی به عنوان یک فرآیند

- اولین گام در پاکسازی داده‌ها، حذف اختلافات است!
- اختلافات ممکن است به چند دلیل پیش بیاید:
  1. طراحی یا کارکرد غلط فرآیند ورود داده‌ها
  2. خطای انسانی در وارد کردن داده
  3. خطاهای عمدی
  4. گذشتن تاریخ مصرف! (آدرس)
  5. نمایش متناقض داده‌ها
  6. استفاده ناجور از کدها
  7. خطای دستگاه ذخیره داده
- خطای مربوط به جمع آوری داده‌ها (یک شی چند اسم داشته باشد)

اولین قدم‌ها برای پاکسازی

۱ **استفاده از metadata**
داده برای توصیف داده! مثل عنوان ستون‌ها، مشخصات موجود، چولگی، برد، ارتباط بین داده‌ها، میانگین و ...

۲ استانداردسازی ورود داده‌ها یا خود جدول (مثل تاریخ، مشخصه و ...)

۳ **قواعد یکتا**: بایستی همگی از یک قاعده یکتا پیروی کنند.

۴ **قاعده پیوستگی** مقداری گم شده بین کوچکترین و بزرگترین مقدار وجود ندارد (تاریخ! موجودی، ...)

۵ **قاعده پوچی** بایستی در ورود داده‌ها روندی اتخاذ کرد که عدم وجود خود یک گزینه باشد و قابل تشخیص.
۱ - مثلا مایل به پاسخ نیستم. ۲ - نمی‌دانم ۳ - زمان ندارم ۴ - به دلایلی امکان پاسخ به این سوال را ندارم)محرمانگی(۵ - در مرحله بعد می‌گویم.

۶ برخی ابزارها هستند که این کارها را برای ما انجام می‌دهند.
Data scrubbing tools، Data auditing tools
Data migration tools
ETL (extraction/transformation/loading) tools

۷ بروز رسانی متادیتا

جمع آوری داده

مسئله تشخیص هویت

۱. **مسئله تشخیص هویت:** چطور موجودیت‌های مشابه دنیای واقعی را در منابع مختلف داده تشخیص دهیم؟
۲. مثلا چگونه تشخیص دهیم costId با costnumber در دو منبع مختلف یکسان است؟
۳. یکی از روش‌ها استفاده از ابزارهای اسلاید قبل است (متادیتا، قاعده پوچی و پیوستگی)
۴. **ساختار داده‌ها:** بایستی ساختار داده‌ها از نظر ورودی یکسان باشد. مثلا در یک فروشگاه ممکن است تخفیف به فاکتور اعمال شود در دیگری برای هر کالا.

تحلیل همبستگی و افزونگی

۱ **تحلیل همبستگی:**
با تحلیل همبستگی برای دو خصیصه مختلف، می‌توان دریافت که دو خصیصه چقدر به هم وابسته هستند.

۲ برای خصیصه‌های اسمی‌از محک χ^2 استفاده می‌کنیم.

۳ برای خصیصه‌های عددی از ضریب همبستگی و کوواریانس استفاده می‌کنیم.

محک همبستگی برای خصیصه‌های اسمی

۱ فرض کنید

$$S(A) = \{a_1, a_2, ..., a_c\}$$
$$S(B) = \{b_1, b_2, ..., b_r\}$$
$$e_{ij} = \frac{(A = a_i) \times (B = b_j)}{n}$$
$$o_{ij} = |((A_i, B_j)|(A_i, B_j) = (a_i, b_j))|$$

۲
$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

۳ درجه آزادی در این توزیع برابر $(r-1) \times (c-1)$ است.

۴ محک χ^2 فرض می‌گیرد که دو خصیصه مستقل هستند. اگر فرض غلط باشد می‌گوییم دو خصیصه همبسته هستند.

	male	*female*	*Total*
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are *gender* and *preferred_reading* correlated?

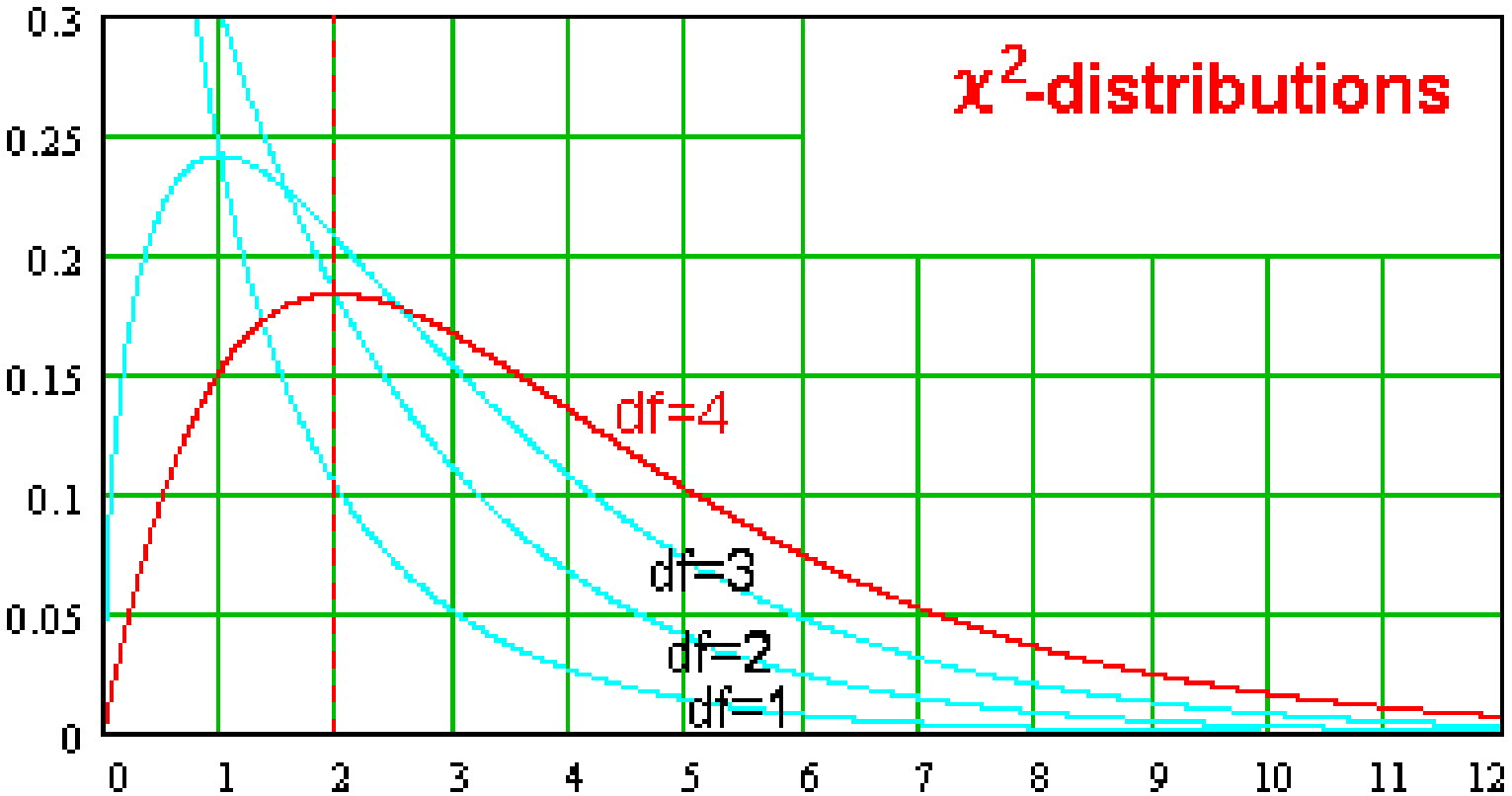
$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

درجه آزادی $(2-1)(2-1) = 1$ خطا= .001 مقدار قابل قبول=10.828

داده ها همبسته هستند

توزیع χ^2

$$f(x) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{\frac{v-2}{2}} e^{-x/2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$



همبستگی در خصیصه‌های عددی

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B},$$

کوواریانس در داده‌های عددی

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n} \qquad E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}.$$

$$Cov(A,B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n},$$

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A\sigma_B}$$

همبستگی در خصیصه‌های عددی

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$$E(AllElectronics) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

$$E(HighTech) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

$$Cov(AllElectroncis, HighTech) = \frac{6\times 20+5\times 10+4\times 14+3\times 5+2\times 5}{5}$$

$$-4\times 10.80 = 50.2 - 43.2 = 7.$$

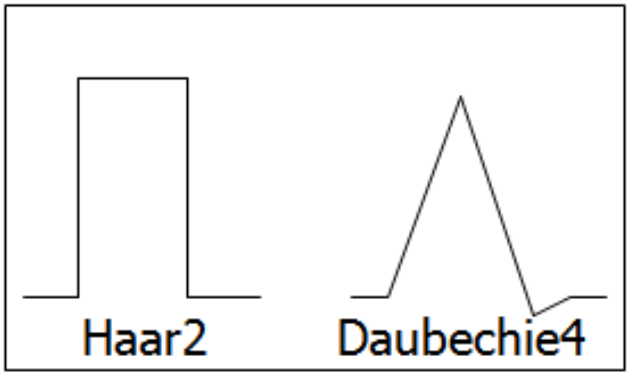
چندگانگی

۱ **چندگانگی:** زمانی که برای یک شی یکسان چند ردیف داده داریم و این زمانی اتفاق می‌افتد که از خصیصه‌هایی مثل آدرس و ... استفاده کنیم.

۲ **تشخیص تعارض:** ممکن است در چند منبع مختلف از واحدها، اندازه‌ها یا کدینگ‌های متفاوتی استفاده کرده باشند.

کاهش داده

تبدیل موجک



۱ $[2, 2, 0, 2, 3, 5, 4, 4] \longrightarrow [2\frac{3}{4}, -1\frac{1}{4}, \frac{1}{2}, 0, 0, -1, -1, 0]$.

Resolution	Averages	Detail Coefficients
8	$[2, 2, 0, 2, 3, 5, 4, 4]$	
4	$[2, 1, 4, 4]$	$[0, -1, -1, 0]$
2	$[1\frac{1}{2}, 4]$	$[\frac{1}{2}, 0]$
1	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$

تحلیل مولفه‌های اصلی PCA

دنبال k بردار یکه متعامدی می‌گردد که در نمایش داده‌ها کمک می‌کند.
مثل حالت انتخاب پایه برای یک فضای برداری.

۱ داده ورودی باید نرمال‌سازی شده باشد.

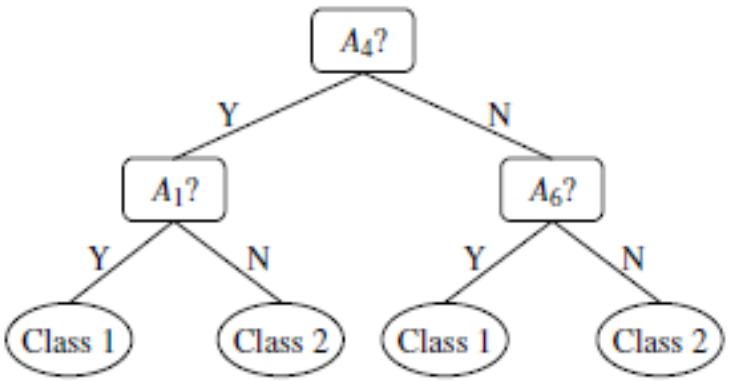
۲ بردارهای متعامدی را می‌یابد که برای ورودی داده‌ها یک پایه می‌باشند. این بردارها را مولفه‌های اصلی می‌گویند.

۳ مولفه‌های اصلی چنان مرتب می‌شوند که میزان اهمیت در واریانس (خود واریانس) به صورت نزولی باشد.

۴ با توجه به ترتیب خصیصه‌ها بر اساس واریانس می‌توان ضعیف‌ترین متغیر را حذف کرد.

انتخاب مجموعه خصیصه‌ها

1. **رو به جلو**
2. **رو به عقب**
3. **ترکیب عقب و جلو**
4. **درخت تصمیم** روش‌هایی چون CART,C4.5,ID3

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

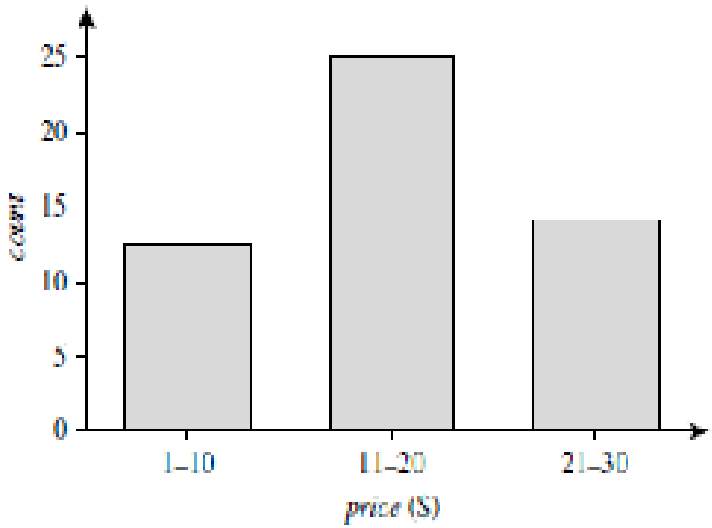
رگرسیون و کاهش داده پارامتری

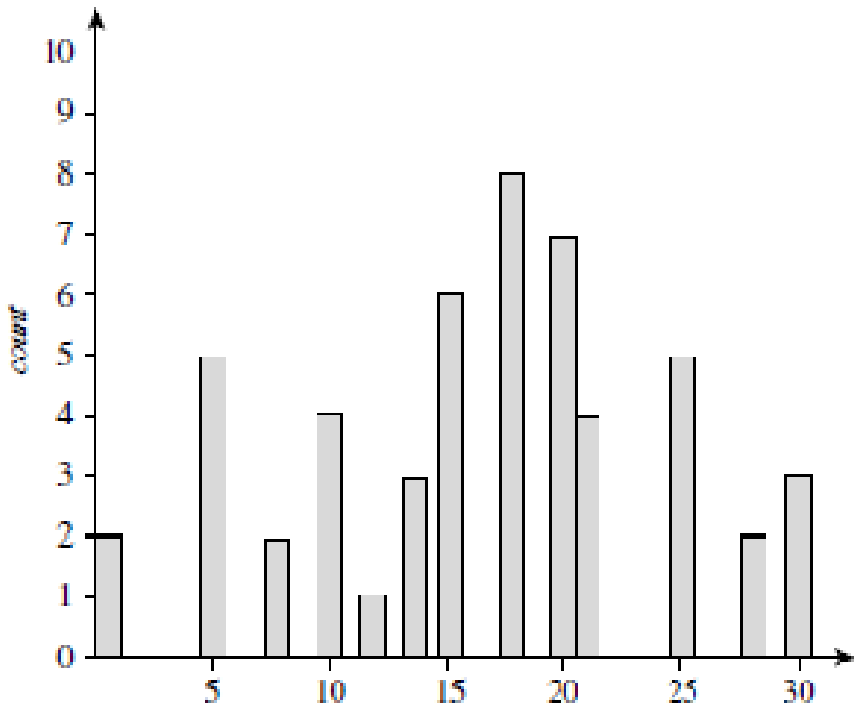
یک توزیع مناسب را می‌یابد که روی همه یا بخشی از خصیصه‌ها خوب عمل می‌کند به جای کل ستون از یک تابع استفاده می‌کنیم.

روش هیستوگرام

با استفاده از جعبه‌هایی توزیع داده را تخمین می‌زند و یک روش معمول در کاهش داده‌ها است

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.





روش خوشه بندی

در فصل‌های ۱۰ و ۱۱ به این موضوع می‌پردازیم.

روش نمونه‌گیری

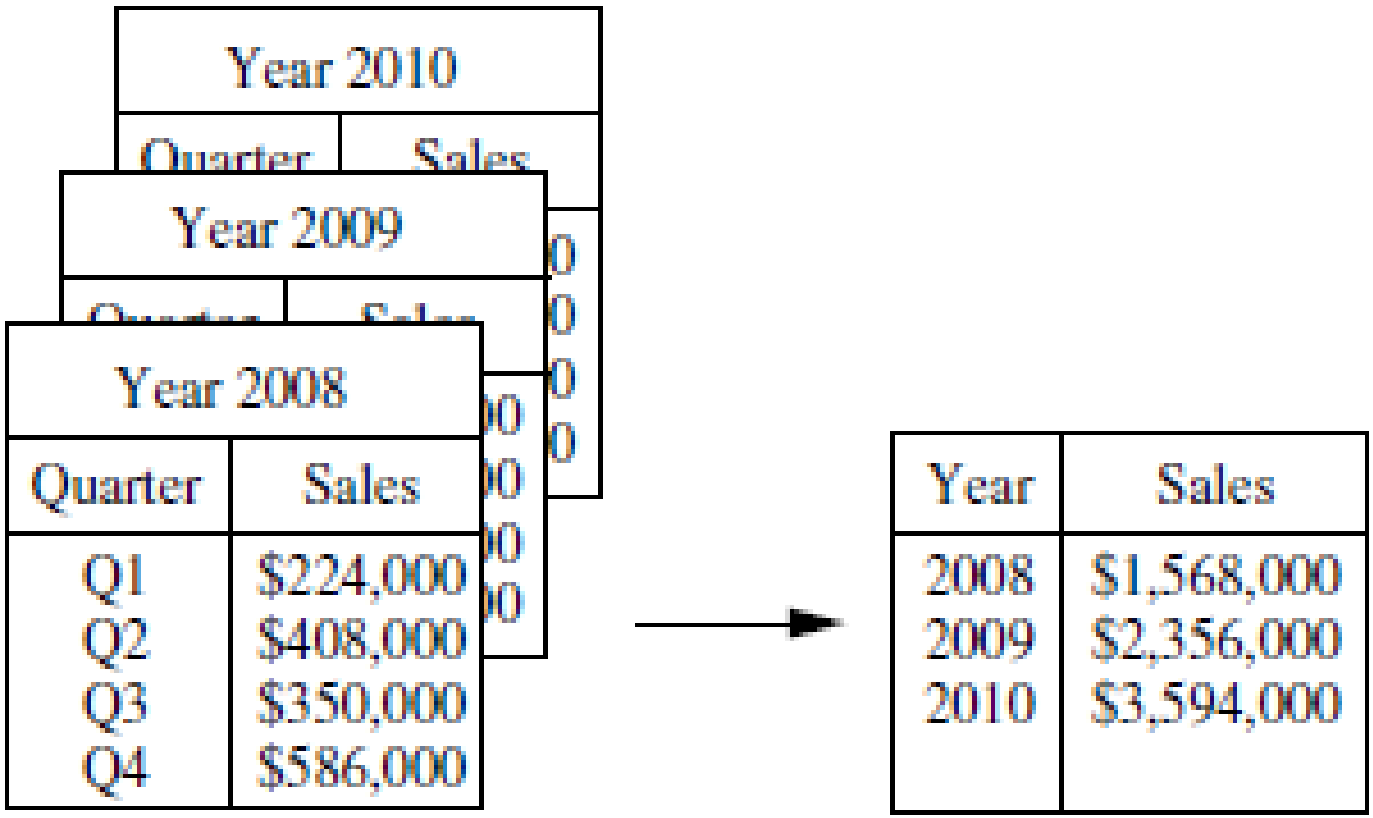
۱ نمونه تصادفی بدون جایگذاری SRSWOR

۲ نمونه تصادفی با جایگذاری

۳ نمونه خوشه‌ای

۴ نمونه لایه‌ای

روش مکعب داده



تبدیل و گسسته‌سازی داده

تبدیل داده

نرمالسازی

۱ نرمالسازی مینی ماکس

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A.$$

۲ نرمال‌سازی میانگین صفر

$$v_i' = \frac{v_i - \overline{A}}{\sigma_A}$$

۳ نرماسازی دهگان

$$v_i' = \frac{v_i}{10^j}$$

گسسته‌سازی

۱ **گسسته‌سازی با جعبه‌ها**

۲ **گسسته‌سازی با تحلیل هیستوگرام**

۳ **گسسته‌سازی با خوشه بندی، درخت تصمیم و تحلیل همبستگی**

تولید مفهوم سلسله مراتبی برای داده‌های اسمی

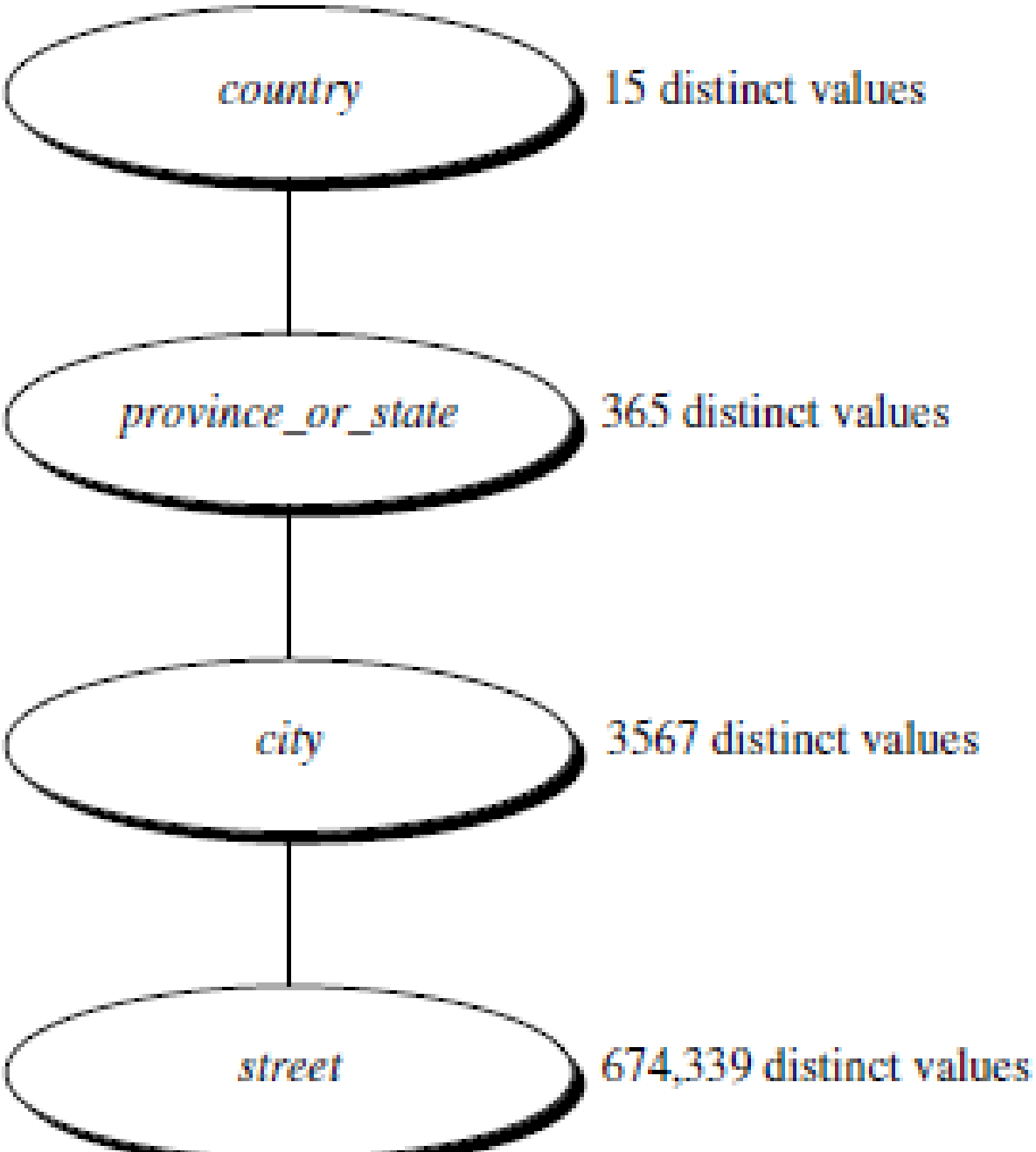
خصیصه‌های اسمی تعداد متناهی ولی احتمالا بسیار زیاد حالت دارد.
ترتیب خاصی وجود ندارد.
تعریف ترتیب به صورت دستی ممکن است بسیار هزینه‌بر باشد.

۱ **یک ترتیب صریح ایجاد می‌کنیم:**
خیابان < شهر < استان < کشور

۲ **یک بخش از سلسله مراتب را با استفاده از گروه بندی صریح داده مشخص می‌کنیم:**
تهران $\subset$ { شهریار، دماوند، شهر ری، تهران } و
ایران $\subset$ { قزوین، مرکزی، تبریز، فارس، تهران }

۳ **تعیین یک دسته از خصیصه‌ها و نه لزوما ترتیب آن‌ها:** پس از انجام این کار توسط کاربر برخی روشها وجود دارند که بقیه کار را انجام دهند (تعداد اجزاء یا دیکشنری).

۴ **تعیین یک ترتیب جزئی روی خصیصه‌ها:** ارتباطات مفهومی از پیش تعیین‌شده



تمرین و پروژه

پروژه

۱ پروژه Gap Miner

۲ پروژه n-Gram

۳ نوشتن برنامه‌ای برای انجام پیش‌پردازش داده‌ها (تمامی حالات)

تمرین

۱ تمرینات فصل سوم

۲ از نرم افزار ویراستیار استفاده کنید (مخصوصا برای پروژه‌های تحویلی)

۳ مراجع بخش سوم را مرور کنید.