بسمه تعالی

داده کاوی

Data mining

مدرس

فاطمه دارائی

f_daraei@semnan.ac.ir

https://fdaraei.profile.semnan.ac.ir

1

فصل دوم: پیش پردازش داده ها

آماده‌سازی داده‌ها به مراحل قبل از داده‌کاوی اطلاق می‌گردد، هرچند از این تکنیک‌ها می‌توان در حین اجرای الگوریتم‌های داده‌کاوی نیز استفاده نمود. این مرحله‌گاهی با نام مرحله‌ی پیش‌پردازش داده‌ها نیزشناخته می‌شود.کیفیت داده‌های ورودی درساده‌ترین تحلیل تا ساخت مدل‌های پیچیده یکی از کلیدهای موفقیت انجام یک پروژه به حساب می‌آید.

کاهش کیفیت داده‌ها بر دقت و اعتبار تحلیل‌های داده‌کاوی تأثیر می گذارند.

2

کیفیت داده

داده‌های ناقص ممکن است ناشی از موارد زیر باشند:
مقدار داده "قابل اعمال نیست" در زمان جمع‌آوری:
تفاوت در ملاحظات بین زمانی که داده جمع‌آوری شده و زمانی که تحلیل می‌شود.
مشکلات انسانی/سخت‌افزاری/نرم‌افزاری

داده‌های نویزی (مقادیر نادرست) ممکن است ناشی از موارد زیر باشند:
ابزارهای جمع‌آوری داده معیوب
خطاهای انسانی یا کامپیوتری در زمان ورود داده‌ها
خطاها در انتقال داده

داده‌های ناسازگار ممکن است ناشی از موارد زیر باشند:
منابع داده متفاوت
نقض وابستگی تابعی (برای مثال، تغییر برخی داده‌های مرتبط)
رکوردهای تکراری نیز نیاز به پاک‌سازی داده دارند.

معیارهای کیفیت داده

دقت: (Accuracy) درست یا غلط، دقیق یا نه

کامل بودن: (Completeness) ثبت نشده، غیر قابل دسترسی

سازگاری: (Consistency) برخی از داده ها اصلاح شده اند اما نه همه

به هنگام بودن: (Timeliness) داده ها به موقع به روزرسانی شده اند؟

باورپذیری: (Believability) اطمینان از درستی داده ها

قابلیت تفسیر: (Interpretability) آیا داده ها به سادگی قابل درک هستند؟

3

وظایف عمده در پیش پردازش داده ها

پاک سازی داده ها Data Cleaning

- پرکردن مقادیر گم شده
- اصلاح داده های نویزی
- شناسایی و حذف داده های دورافتاده
- از بین بردن تناقضات

تجمیع داده ها Data Integration

یک مفهوم ممکن است در پایگاه‌های مختلف نام‌های مختلفی داشته باشد که منجر به ناسازگاری و تکرار می شود.

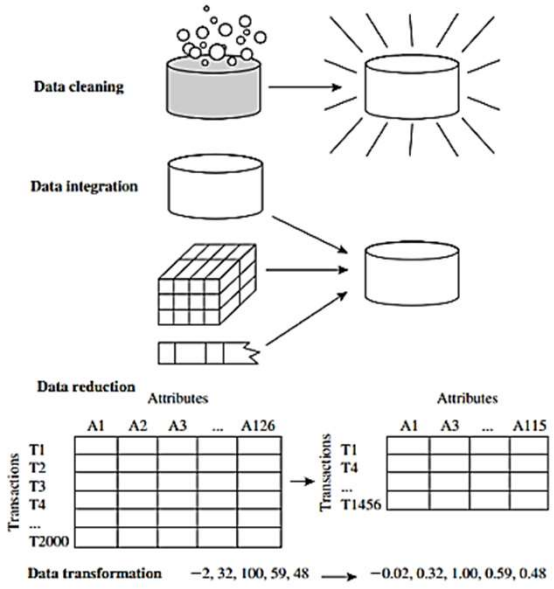
کاهش داده ها Data Reduction

- کاهش ابعاد
- فشرده‌سازی داده‌ها

تغییر شکل داده ها Data Transformation

- نرمال سازی
- گسسته سازی

4



5

پاک سازی داده ها Data Cleaning

داده ها در دنیای واقعی نادرست‌اند: بسیاری از داده ها به طور بالقوه نادرست‌اند، به دلیل ابزار معیوب و ناقص،

خطای انسانی یا کامپیوتری، خطای انتقال.

ناقص بودن: Incomplete فقدان مقادیر برای برخی ویژگی‌ها یا نبود برخی ویژگی‌های مورد نظر، یا داشتن داده‌های کلی و نه جزئی.
مثال: شغل=" " (خالی بودن ویژگی شغل)

نویز: Noisy وجود خطاها یا داده‌های پرت در مجموعه داده.
مثال: حقوق="-۱۰" (وجود مقدار غیر منطقی)

ناسازگار بودن: Inconsistent وجود تناقض در کدها یا نام‌ها.
مثال: سن="۴۲" و تاریخ تولد="۰۳/۰۷/۱۹۹۷" (عدم تطابق بین سن و تاریخ تولد)
مثال: امتیاز قبلاً "۱، ۲، ۳" بوده و اکنون "A، B، C" است.
مثال: وجود اختلاف بین رکوردهای تکراری.

عمدی: **Intentional**
۱ ژانویه روز تولد همه!

این مشکلات می‌توانند تحلیل داده‌ها را دشوار کنند و نیاز به پاک‌سازی داده‌ها دارند.

6

missing Data (داده از دست رفته): ممکن است یک ویژگی از یک داده وجود نداشته باشد یا اندازه‌گیری نکرده باشیم.

راه حل:

۱-آن سطر را حذف می‌کنیم.

۲- پر کردن مقادیر از دست رفته به صورت دستی.

۳- پر کردن خودکار با یک مقدار مشخص. مثلا ناشناخته، Unknown

۳-اگرصورت مسئله دسته بندی است برای داده‌های همکلاس میانگین‌گیری انجام می‌دهیم وآن مقدار را برای داده‌ی از دست‌رفته درنظر می‌گیریم. اگرصورت مسئله دسته بندی نیست میانگین تمام داده‌ها یا داده‌های نزدیک را درنظر می‌گیریم.

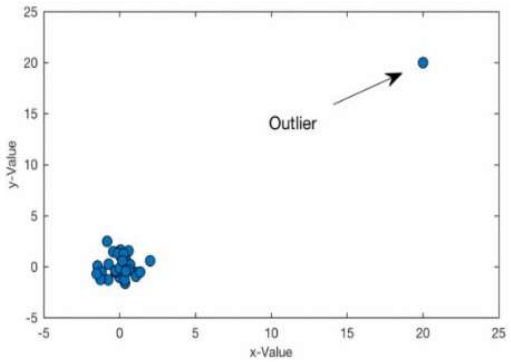
ویژگی 1	ویژگی 2	ویژگی 3	ویژگی 4	برچسب
5	20	100	1	1
9	60	100	2	0
8	70	300	1	1
11	40	120	3	1
7	NULL	200	2	0
6	50	300	1	1

7

Outlier دور افتاده یا خارج از محدوده:

میانگین و انحراف معیارتمام داده ها را حساب می‌کنیم.

داده هایی را که در بازه‌ی Threshold قرارگرفتند را به عنوان داده اصلی درنظر می‌گیریم.



Threshold = Means +(-) 2δ

8

داده های نویزی

نویز: خطای تصادفی یا واریانس در یک متغیر اندازه گیری شده. راه حل:

۱- Binning : ابتدا داده‌ها مرتب شده و تقسیم به دسته‌های با فرکانس تکرار یکسان می‌شوند.

به عنوان مثال داده های طبقه بندی شده برای قیمت (به دلار) را داریم:

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

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 هموارسازی توسط میانگین
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

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

9

تجمیع داده ها Data Integration

ادغام داده ها: ترکیب داده ها از منابع متعدد در یک انباره منسجم.

باید مشخص شود یک ویژگی از یک منبع معادل کدام ویژگی از منبع دوم است.

در یک منبع مقادیر به صورت H و S است در یک منبع دیگر به صورت ۰ و ۱

باید مشخص شود دردو منبع موجودیت هایی را شناسایی کنیم که هر دو یک چیز را توصیف می کنند.

Bill Clinton = William Clinton

افزونگی داده ها: افزونگی در داده ها اغلب هنگام ادغام چندین پایگاه داده رخ می دهد.

ممکن است یک ویژگی یا شیء یکسان در پایگاه داده های مختلف نام های متفاوتی داشته باشد

داده های قابل مشتق شدن: یک ویژگی ممکن است مشتق از یک ویژگی در جدولی دیگر باشد. مانند درآمد سالیانه.

ویژگی های افزونه را می توان با تجزیه و تحلیل همبستگی و تحلیل کوواریانس تشخیص داد.

10

تحلیل همبستگی بر روی داده های اسمی

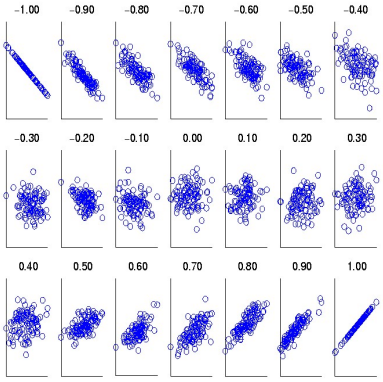
$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n)\sigma_A\sigma_B}$$

که در آن n تعداد ویژگی ها است.

اگر r(A,B)>0، A و B همبستگی مثبت دارند (مقادیر A با B افزایش می یابد). هر چه بیشتر، همبستگی قوی تر است.

r(A,B)=0: A و B مستقل هستند.

r(A,B)<0: A و B همبستگی منفی دارند.



11

تحلیل کواریانس روی داده‌های عددی

کواریانس مشابه همبستگی است.

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

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

کوواریانس مثبت: اگر Cov(A,B) > 0 باشد، A و B هر دو بزرگتر از مقادیر مورد انتظارشان هستند.

کوواریانس منفی: اگر Cov(A,B) < 0 باشد، اگر A بزرگتر از مقدار مورد انتظارش باشد، B احتمالا کوچکتر از مقدار مورد انتظارش خواهد بود.

استقلال: Cov(A,B) = 0 ، A و B مستقل هستند اما عکس آن درست نیست.

مثال: فرض کنید دو سهام A و B در یک هفته دارای مقادیر زیر هستند: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

سوال: اگر سهام تحت تأثیر همان روندهای صنعت قرار گیرند، آیا قیمت آنها با هم افزایش می یابد یا کاهش می یابد؟

$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$

E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4

E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6

Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4

بنابراین، A و B با هم افزایش می یابند.

12

تغییر مقیاس داده ها Data Transformation

اگر دو ویژگی با مقیاس‌های مختلف داشته باشیم (درrangeهای مختلف) بدون پیش پردازش معمولا تاثیرx2 بیشترازx1 خواهد بود. لذا تمام ویژگی ها را دریک بازه تغییرمقیاس می‌دهیم.

مثال:

x_1	x_2
0.0001	100000
0.0002	200000
0.0001	300000

1- مقیاس به [-1,1] :

$$x_i' = \frac{x_i}{Max|x_i|\ i = 1,..,n}$$

2- مقیاس به [0,1] :

$$x_i' = \frac{x_i - Min(x_i)}{Max(x_i) - Min(x_i)}$$

13

3- مقیاس به [-1,1] :

$$x_i' = \frac{x_i - Min(x_i)}{Max(x_i) - Min(x_i)} * 2 - 1$$

4- مقیاس به [L,H]:

$$x_i' = \frac{x_i - Min(x_i)}{Max(x_i) - Min(x_i)} * (H-L) + L$$

5-

$$x_i' = \frac{x_i - Average(X)}{\delta(x)}$$

14

کاهش داده ها Data Reduction

بدست آوردن یک نمایش کاهش یافته از مجموعه داده که حجم بسیار کمتری دارد اما نتایج تحلیلی یکسان (یا تقریباً یکسان) را تولید می کند.

چرا کاهش داده؟

تجزیه و تحلیل داده های پیچیده ممکن است زمان بسیار زیادی طول بکشد تا در تمام مجموعه داده اجرا شود.

زمان و فضای مورد نیاز در داده کاوی را کاهش می دهد.

مصورسلازی ساده تر می شود.

تکنیک های کاهش ابعاد:

- تبدیل موجک Wavelet transforms
- تحلیل مولفه های اساسی Principal Components Analysis (PCA)
- روشهای با نظارت و غیرخطی Feature subset selection

15

Principal Components Analysis (PCA)

یکی از عمومی ترین روش های آماری به منظورکاهش ابعاد داده ها روش تحلیل مولفه های اصلی است. همانطور که از نامش پیداست، می تواند مولفه های اصلی را شناسایی کند. به جای اینکه تمام ویژگی ها را مورد بررسی قرار دهیم، یکسری ویژگی ها را که ارزش تحلیل بیشتری دارند تحلیل کنیم.

در این روش واریانس کل صفات خاصه موجود تحلیل می شود.

مراحل انجام تحلیل مولفه اساسی:

۰- داده ها نرمال می شوند(اختیاری)

۱- ماتریس کواریانس ویژگی ها ساخته می شود.

۲- بردارهای ویژه و مقدارهای ویژه این ماتریس بدست می آید. $C * Evector = \lambda * Evector$

تساوی $|C - \lambda I|=0$ را حل می کنیم تا مقادیر ویژه بدست آیند سپس آن ها را در فرمول بالا جاگذاری می‌کنیم تا به ازای هر مقدار ویژه یکEvector بدست آید.

C: مالتریس کواریانس ویژگی ها λ: مقدار ویژه Evector: بردار ویژه (d * 1)

16

۳- λi ها (مقادیر)ویژه را به صورت نزولی مرتب می‌کنیم.

$$\lambda_1 > \lambda_2 > \ ... > \lambda_d$$

۴- k مناسب توسط کاربر یا رابطه زیر بدست می آید:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq 0.95$$

K = کوچکترین مقداری که با آن رابطه مقابل برقرار باشد.

۵- در نهایت X جدید از فرمول زیر محاسبه می شود:

$$X_{new} = X_{old} \ * \text{ماتریس نگاشت}$$

ماتریس نگاشت یک ماتریسd*k است که در هر ستون آن یک Evector قرار دارد.

17

مثال

	Math	English	Arts
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

Matrix A

$\bar{A}$ = [66 60 60]

Mean of Matrix A

	Math	English	Art
Math	504	360	180
English	360	360	0
Art	180	0	720

Covariance Matrix of A

18

دانشگاه سمنان
Semnan University
پردیس فرزانگان

$$\det(A-\lambda I) = 0 \longrightarrow \det\left(\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$$

$$\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

$$\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

19

دانشگاه سمنان
Semnan University
پردیس فرزانگان

$$\det\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix} \longrightarrow -\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800 = 0$$

با حل معادله بالا مقادیر ویژه به صورت زیر بدست می آیند:

$$\lambda \approx 44.81966\ldots, \lambda \approx 629.11039\ldots, \lambda \approx 910.06995\ldots$$

سپس با قرار دادن مقادیر ویژه در بردارهای زیر، بردارهای ویژه بدست می آیند: C * Evector = λ * Evector

بردارهای ویژه:

$$\begin{pmatrix} -3.75100\ldots \\ 4.28441\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} -0.50494\ldots \\ -0.67548\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} 1.05594\ldots \\ 0.69108\ldots \\ 1 \end{pmatrix}$$

20

دانشگاه سمنان
Semnan University
پردیس فرزانگان

مقادیر ویژه به ترتیب زیر مرتب می شوند:

$$\begin{pmatrix} 910.06995 \\ 629.11039 \\ 44.81966 \end{pmatrix}$$

پس از پیدا کردن k مناسب ماتریس نگاشت به این صورت می شود:

$$\mathbf{W} = \begin{bmatrix} 1.05594 & -0.50494 \\ 0.69108 & -0.67548 \\ 1 & 1 \end{bmatrix}$$

در انتها با استفاده از فرمول زیر ماتریس جدید با دو بعد حاصل می‌شود:

$$X_{new} = X_{old} * \text{ماتریس نگاشت}$$

21