





دانشگاه فنی و حرفه ای استان مرکزی

داده کاوی

مدرس:

فاطمه مقیمی

پاییز ۱۴۰۱



فصل سوم: پیش پردازش داده‌ها

- پیش پردازش داده‌ها: خلاصه‌ای جامع
کیفیت داده‌ها (Data Quality) 
- عملیات اصلی در پیش پردازش داده‌ها
- پاک‌سازی داده‌ها (Data Cleaning)
- تجمیع داده‌ها (Data Integration)
- کاهش داده‌ها (Data Reduction)
- تغییر شکل (Transformation) و گسسته‌سازی (Discretization) داده‌ها
- خلاصه فصل

کیفیت داده: چرا داده‌ها را پیش‌پردازش میکنیم؟

- وضعیت پایگاه داده‌های کنونی: حاوی داده‌های نویزی، داده‌های مفقود، داده‌های ناسازگار
- داده‌های بی‌کیفیت منجر به کاوش‌های بی‌کیفیت خواهد شد
- **معیارهای کیفیت داده:**
 - **دقت (Accuracy)** درست یا غلط، دقیق یا نه
 - **کامل بودن (Completeness)** ثبت نشده، غیر قابل دسترسی، ...
 - **سازگاری (Consistency)** برخی از داده‌ها اصلاح شده‌اند اما نه همه، آویزان، ...
 - **به‌هنگام بودن (Timeliness)** داده‌ها به موقع به روز رسانی شده‌اند؟
 - **باورپذیری (Believability)** اطمینان از درستی داده‌ها
 - **قابلیت تفسیر (Interpretability)** آیا داده‌ها به سادگی قابل درک هستند؟

وظایف عمده در پیش پردازش داده‌ها

1) پاک‌سازی داده‌ها

- پر کردن مقادیر گم شده
- اصلاح داده‌های نویزی
- شناسایی و حذف داده‌های دورافتاده
- و از بین بردن تناقضات

2) تجمیع داده‌ها

- یک مفهوم ممکن است در پایگاه‌های مختلف نام‌های مختلفی داشته باشد که منجر به ناسازگاری و تکرار میشود

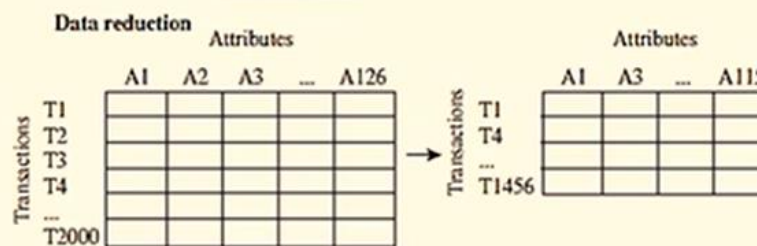
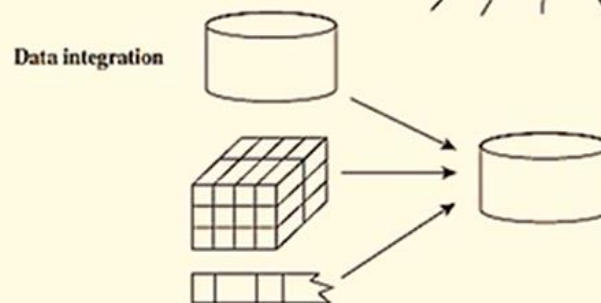
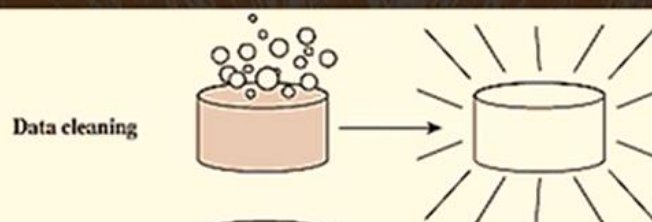
3) کاهش داده‌ها

- کاهش ابعاد (Dimensionality reduction)
- کاهش تکثر (Numerosity reduction)
- فشرده‌سازی داده‌ها (compression)

4) تغییر شکل داده‌ها

- نرمال سازی
- گسسته‌سازی
- تولید سلسله مراتب مفاهیم

انواع پیش پردازش ها



Data transformation $-2, 32, 100, 59, 48 \rightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

پاکسازی داده‌ها

- داده‌ها در دنیای واقعی نادرست هستند: بسیاری از داده‌ها به طور بالقوه نادرست اند، دلایل: ابزار معیوب و ناقص، خطای انسانی یا کامپیوتری، خطای انتقال
- **داده‌های ناکامل: (incomplete)** ویژگی مقدار ندارد، فقدان ویژگی‌های مورد علاقه یا تنها دارای داده‌های تجمیعی
 - به عنوان مثال: شغل = "" (فاقد داده)
- **داده‌های نویز: (noisy)** حاوی نویز، اشتباهات یا داده‌های پرت
 - به عنوان مثال: حقوق = "10-" (خطا)
- **ناسازگاری در داده‌ها: (inconsistent)** شامل اختلاف در کد یا نام، به عنوان مثال
 - سن = "۴۲"، تاریخ تولد = "۱۳۸۹/۴/۲۳"
 - دسته‌بندی قبلی، "1, 2, 3" دسته‌بندی فعلی "A, B, C"
 - تفاوت بین رکوردهای تکراری
- **عمدی: (Intentional)** (پنهان کردن داده‌های از دست رفته)
 - ۱ ژانویه به عنوان روز تولد همه؟

داده‌های ناقص (Missing Values)

- داده‌ها همیشه در دسترس نیستند
 - مثال: بسیاری از رکوردها، مقدار ثبت شده‌ای برای ویژگی‌های مختلف ندارند
- گم‌شدن داده‌ها به دلیل:
 - نقص تجهیزات
 - به دلیل ناسازگاری با دیگر داده‌های ثبت شده حذف شده‌اند
 - داده‌ها به دلیل سوء تفاهم وارد نشده‌اند
 - داده‌های خاصی ممکن است در زمان ورود اطلاعات، مهم قلمداد نشوند
 - نبود تاریخ ثبت یا تغییر داده‌ها

چگونگی رفتار با داده‌های Miss؟

- (1) نادیده گرفتن رکورد: معمولاً زمانی که برچسب کلاس موجود نباشد، انجام می‌شود (در زمان دسته‌بندی) — زمانی که درصد مقادیر از دست رفته برای هر ویژگی، بطور قابل توجهی متفاوت باشد، موثر نیست
- (2) پر کردن مقادیر از دست رفته به صورت دستی: خسته کننده + غیر عملی؟
- (3) پر کردن خودکار با روش‌های زیر:
 - یک ثابت جهانی: به عنوان مثال، نا شناخته "unknown"، یک کلاس جدید!
 - میانگین ویژگی (به عنوان مثال داده‌های نرمال از میانگین و داده‌های چوله از میانه استفاده شود)
 - میانگین ویژگی برای همه نمونه‌های متعلق به همان کلاس: هوشمندانه‌تر
 - محتمل ترین مقدار: مبتنی بر استنتاج همچون فرمول های بیزین یا درخت تصمیم

داده‌های نویزی

• نویز: خطای تصادفی یا واریانس در یک متغیر اندازه‌گیری شده

• مقادیر نادرست ویژگی به دلیل

– نقص در ابزار جمع‌آوری داده‌ها

– مشکلات ورود داده‌ها

– مشکلات انتقال داده‌ها

– محدودیت تکنولوژی

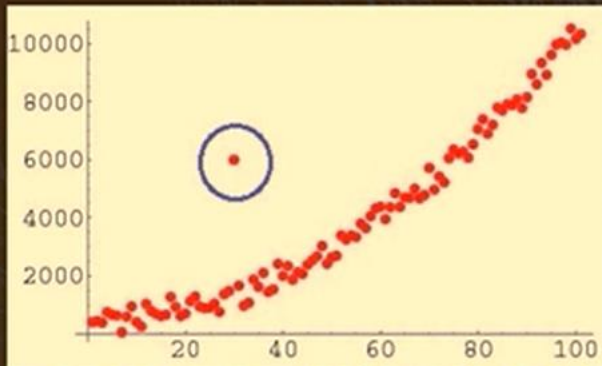
– تناقض در قراردادهای نامگذاری

• سایر مشکلات داده‌ها، نیاز به پاک کردن داده‌ها دارند

– رکورد تکراری

– اطلاعات ناقص

– اطلاعات متناقض



چگونگی برخورد با داده‌های نویزی

Binning (1)

- ابتدا داده‌ها مرتب شده و تقسیم به دسته‌های با فرکانس تکرار یکسان (equal-frequency) می شوند.
- داده‌های طبقه‌بندی شده برای قیمت (به دلار): ۴، ۸، ۹، ۱۵، ۲۱، ۲۱، ۲۴، ۲۵، ۲۶، ۲۸، ۲۹، ۳۴
- ✳ تقسیم‌بندی داده‌ها با فراوانی یکسان bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 2: 26, 28, 29, 34

✳ هموارسازی توسط میانگین:

- Bin 1 : 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3 : 29, 29, 29, 29

✳ هموارسازی توسط میانه مرزها:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

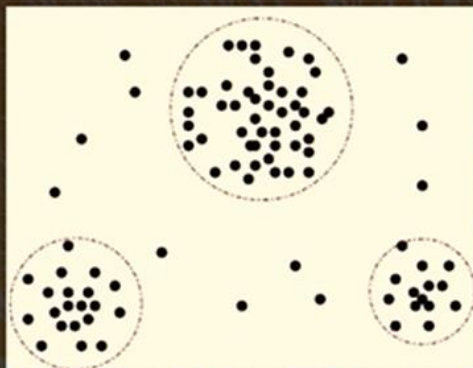
چگونگی برخورد با داده‌های نویزی

(2) رگرسیون (Regression)

– هموار کردن داده‌ها به کمک سازگار کردن آن‌ها با تابع رگرسیون

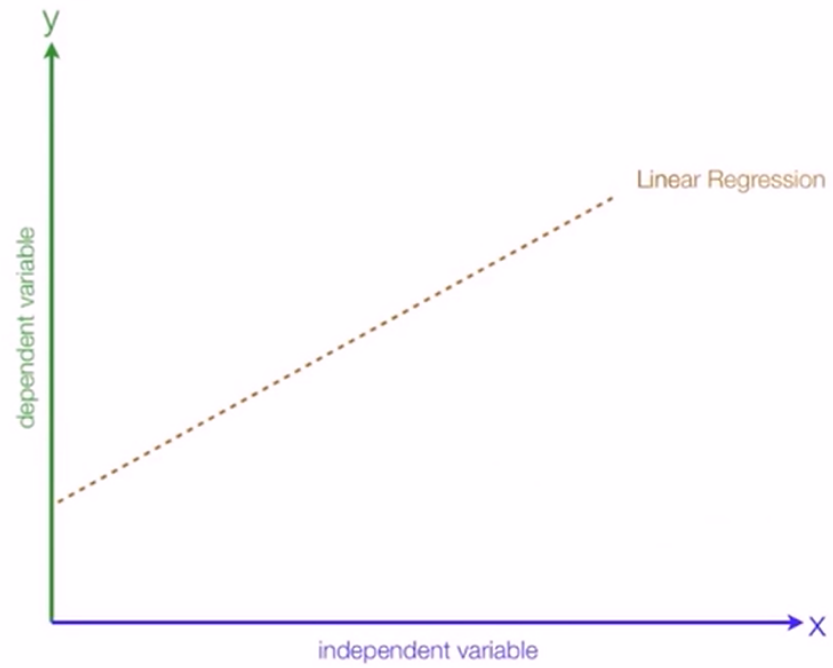
(3) خوشه‌بندی

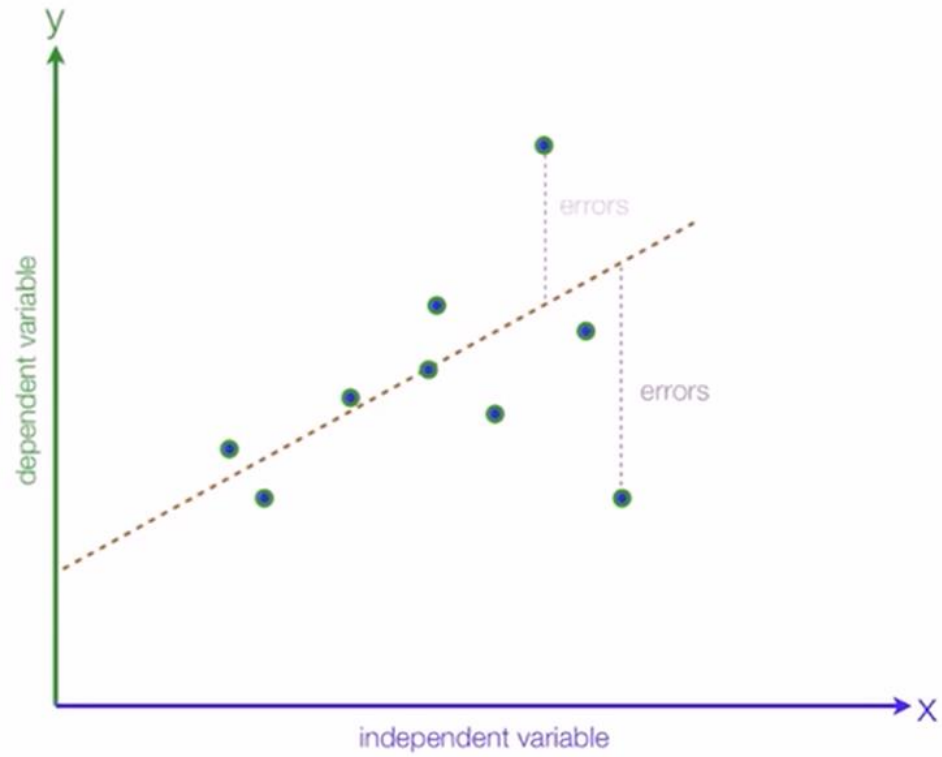
– شناسایی و حذف داده‌های پرت

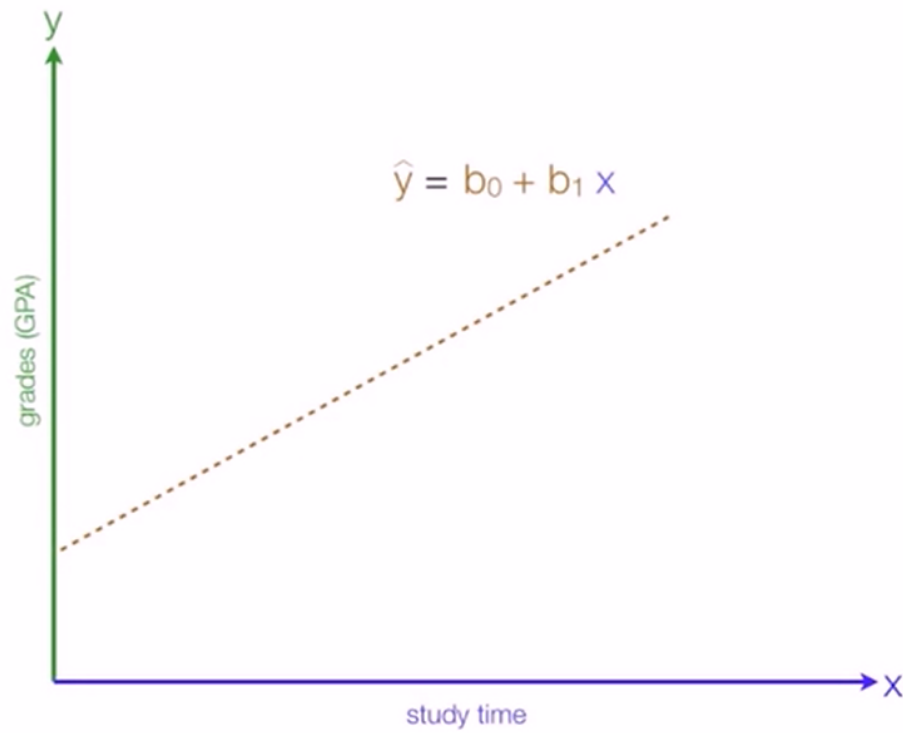


(4) ترکیب بازرسی کامپیوتر و انسان

– تشخیص مقادیر مشکوک بصورت خودکار سپس بررسی آن‌ها توسط عامل انسانی (به عنوان مثال تعامل با داده‌های دورافتاده)







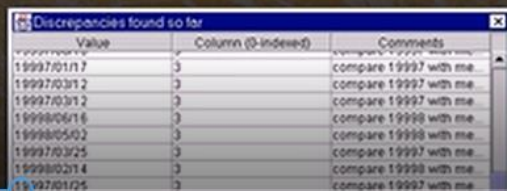
پاکسازی داده‌ها به عنوان یک رویه

• تشخیص اختلاف داده‌ها (Data discrepancy detection)

- استفاده از فراداده Use metadata (به عنوان مثال، دامنه، محدوده، وابستگی، توزیع)
- بررسی قانون منحصر به فرد بودن، قانون توالی و قانون null
- استفاده از ابزارهای تجاری
- تمیزکردن داده: استفاده از دانش زمینه (به عنوان مثال، کد پستی، املاي کلمات) جهت تشخیص خطاها و ایجاد اصلاحات
- حسابرسی داده‌ها: با تجزیه و تحلیل داده‌ها برای کشف قوانین و ارتباط برای شناسایی متخلفان (مثال استفاده از همبستگی و خوشه‌بندی برای پیدا کردن داده‌های پرت)

• مهاجرت و یکپارچه‌سازی داده‌ها (Data migration and integration)

- ابزار مهاجرت داده‌ها: اجازه تحولاتی برای خاص شدن
- ابزارهای ETL (استخراج/انتقال/بارگذاری): اجازه دادن به کاربران برای تعیین تبدیلات از طریق یک رابط کاربر گرافیکی



Value	Column (0-indexed)	Comments
199970117	3	compare 19997 with me
199970312	3	compare 19997 with me
199970312	3	compare 19997 with me
199980616	3	compare 19998 with me
199980502	3	compare 19998 with me
199970325	3	compare 19997 with me
199980214	3	compare 19998 with me
199970125	3	compare 19997 with me

• تجمیع دو فرایند (Integration of the two processes)

- تکراری و تعاملی (به عنوان مثال Potter's wheel)

تجميع داده‌ها

• ادغام داده‌ها

- ترکیب داده‌ها از منابع متعدد در یک انباره منسجم
- ادغام شماها (Schema integration) : مانند $B.cust = A.cust-id$
- ادغام ابرداده‌ها (metadata) از منابع مختلف
- (1) مشکل شناسایی موجودیت (Entity identification problem)
- باید مشخص شود که یک ویژگی از منبع اول معادل کدام ویژگی از منبع دوم است
 - در یک منبع مقادیر به صورت H و S است در منبع دوم 1 و 0
- باید مشخص شود دو منبع موجودیت‌هایی را شناسایی کنیم که هر دو یک چیز را توصیف میکنند
- Bill Clinton = William Clinton
- تشخیص و برطرف نمودن ناسازگاری مقادیر داده
 - برای یک موجودیت دنیای واقعی، مقادیر ویژگی از منابع مختلف، متفاوت است.
 - دلایل احتمالی: نمایش‌های متفاوت، مقیاس‌های مختلف، به عنوان مثال، متریک در مقابل واحد بریتانیا

بررسی افزونگی در تجمیع داده‌ها (۲)

2) افزونگی داده‌ها اغلب هنگامی رخ می‌دهد که پایگاه داده‌های متعدد ادغام می‌گردند.

- **شناسایی شی:** ویژگی‌ها یا اشیا ممکن است نام‌های مختلف در پایگاه داده‌های مختلف داشته باشند.
- **داده قابل اشتقاق:** یکی ویژگی ممکن است مشتق از یک ویژگی در جدولی دیگر باشد. به عنوان مثال، درآمد سالانه
- ویژگی‌های افزونه را می‌توان با تحلیل همبستگی (correlation) یا کوواریانس (covariance) تشخیص داد

بررسی افزونگی در تجمیع داده‌ها

- برخی از افزونگی‌ها را می‌توان با تحلیل همبستگی شناسایی نمود.
 - با داشتن دو ویژگی، چنین تحلیلی می‌تواند میزان وابستگی دو ویژگی را نسبت به هم بسنجد
- (a) برای داده‌های اسمی، ما از تست (کای-دو) χ^2 استفاده می‌کنیم.
- (b) برای ویژگی‌های عددی، ما استفاده می‌کنیم از **correlation coefficient** و کواریانس، هر دو نحوه‌ی ارتباط ویژگی‌ها را نمایان می‌سازند

تحلیل همبستگی (داده‌های اسمی)

- ابتدا جدول **Contingency** را تکمیل می‌کنیم به این صورت که در سطر مقادیر اسمی یک ویژگی و در ستون مقادیر ویژگی دوم قرار می‌گیرند
- خانه‌های جدول تعداد حالات موجود در داده‌ها را نشان می‌دهند
- به طور مثال در جدول زیر ۲۵۰ حالت (**fiction, male**) داریم که به این اعداد فراوانی مشاهده شده **observed frequency** می‌گویند و با **O_{ij}** نشان می‌دهند

	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?

- اعداد داخل پرانتز، فراوانی مورد انتظار (**expected frequency**) هستند که به روش زیر محاسبه می‌شوند

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{n}$$

$$e_{11} = \frac{\text{count}(\text{male}) \times \text{count}(\text{fiction})}{n} = \frac{300 \times 450}{1500} = 90,$$

تست کای-دو

- حال بر اساس این جدول **تست χ^2** محاسبه می شود

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

	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?

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