
بسم الله الرحمن الرحیم

داده کاوی

دکتر حسام عمرانپور

دانشگاه صنعتی نوشیروانی بابل

بارم بندی:

میانترم حل تمرین پایان ترم پروژه

مراجع: Data mining concepts and techniques,Jiawei Han,Michelin Cumber and Jian Pei

داده کاوی مهدی اسماعیلی

تنظیم : فاطمه حیدری

حدیثه پورعلی


سرفصل ها
1
آماده سازی داده ها
2
الگو های مکرر و
قوانین انجمنی
3
روش های طبقه
بندی و تخمین
4
خوشه بندی

فصل 1

آماده سازی داده ها

انواع داده و خصوصیت آن ها

نوع داده ها می تواند ما را در انتخاب تکنیک های داده کاوی کمک کند.صرف نظر از این که داده ها ممکن است ساخت یافته ، نیمه ساخت یافته و یا غیر ساخت یافته باشند ، ممکن است دو گونه از داده ها داشته باشیم ؛ داده های کمی و داده های کیفی.

متغیرهای کمی

هرگاه صفت خاصه و داده مورد نظر را بتوان شمارش و یا اندازی گیری کرد و سپس آن را به صورت عدد بیان کرد ، یک متغیر کمی خواهیم داشت که به آن متغیر عددی نیز گویند.در این نوع داده ها ما دارای دو خاصیت ترتیب و فاصله هستیم.این متغیر ها می توانند پیوسته یا گسسته باشند.

متغیر های پیوسته ، متغیرهایی هستند که می توانند کلیه مقادیر حقیقی بین محدوده ای داشته باشند ، مثل قد و وزن.

متغیر های گسسته ،حوزه مقادیر این متغیر ها مجموعه ای قابل شمارش و محدود است ، مثل تعداد فرزندان یک خانواده.

متغیرهای کیفی

حاصل متغیر های کیفی را نمی توان با عدد نشان داد ، بلکه بر اساس خاصیتی که موردنظر است ، داده ها در طبقات و دسته های مختلفی قرار می گیرند. در این نوع متغیرها فاصله قابل تعریف نیست و در برخی موارد ترتیب معنا دارد (ordinal) و در برخی موارد ترتیب معنا ندارد (nominal).

رابطه ای که در این نوع از داده ها وجود دارد ، مساوی یا نامساوی بودن است.

ابعاد داده ها

در فرآیند داده کاوی با مقدار خیلی زیاد و متنوعی از داده ها رو به رو هستیم ، که هریک دارای رفتار متفاوتی هستند.هرچه تعداد ویژگی ها (ابعاد) بیشتر باشد ، تعداد داده های موجود در باید بیشتر شود تا دقت بیشتر شود.

اگر در یک فضای یک بعدی 100 نمونه داشته باشیم ، با حفظ چگالی تعداد این نمونه ها در یک فضای 5 بعدی به 100^5 می رسد.

$$E_d(p) = p^{1/d}$$

d : تعداد ابعاد داده

p : کسری از نمونه ها که تعداد آن مشخص است.

مثال ابعاد داده

اگر بخواهیم فقط 10 درصد از نمونه ها را جمع آوری کنیم ،سطح های انتخابی در داده هایی با ابعاد 2 ، 3 و 10 در ادامه محاسبه شده اند :

$$E_2(0.10) = (0.10)^{1/2} = 0.32$$

$$E_3(0.10) = (0.10)^{1/3} = 0.46$$

$$E_{10}(0.10) = (0.10)^{1/10} = 0.80$$

این محاسبات نشان می دهد که حجم زیادی از داده های کنار هم ، فقط بخش کوچکی از داده ها را در فضاهایی با ابعاد بالاتر شامل می شوند.

فاصله داده ها در ابعاد d

خصوصیت دیگر داده های حجیم ، فاصله ی بیشتر نمونه ها در ابعاد بالاتر است. برای تعداد n نمونه داده در ابعاد d فاصله ی D به صورت زیر محاسبه می شود:

$$D(d,n) = {}^{1}/_{2} \times ({}^{1}/_{n})^{{}^{1}/_{d}}$$

آمار

در این بخش به طور خلاصه به بیان برخی از شاخص های مرکزی و شاخص های پراکندگی در آمار توصیفی می پردازیم.

- **شاخص های مرکزی :**

میانگین ، میانه ، مد

- **شاخص های پراکندگی :**

دامنه تغییرات و میانگین انحرافات ، انحراف چارکی ، واریانس و انحراف استاندارد

- **کواریانس و همبستگی**

میانگین

یکی از معروف ترین شاخص های مرکزی میانگین است، که انواع آن عبارت است از میانگین حسابی ، میانگین وزن دار ، میانگین هندسی و میانگین هارمونیک.

- **میانگین حسابی :**

$$\mu_A = {}^{1}/_{n} \sum_{i=1}^{n} X_i$$

- **میانگین وزن دار :**

$$\mu_W = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}$$

میانگین

- میانگین هندسی :

$$\mu_G = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$$

- میانگین هارمونیک :

$$\mu_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

$$\mu_H \leq \mu_G \leq \mu_A$$

میانه

میانه عددی است که توزیع داده ها را به دو قسمت مساوی تقسیم می کند، به نحوی که نیمی از داده ها بزرگتر و نیم دیگر کوچکتر از آن هستند.

در مجموعه اعداد مرتب شده داده ی میانی به عنوان میانه لحاظ می گردد. در صورتی که تعداد داده ها زوج باشد، نصف مجموع دو داده ای که در وسط قرار دارند، میانه محسوب می شود.

میانه برای داده های پیوسته ی دسته بندی شده از فرمول زیر محاسبه می شود :

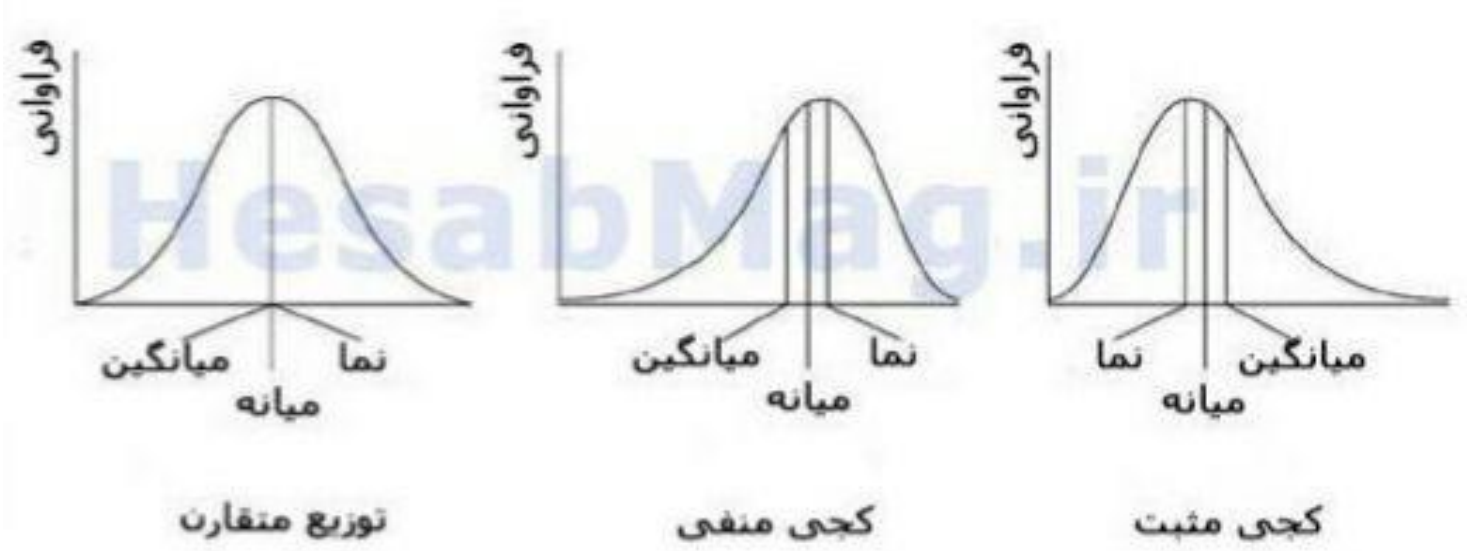
$$m = L + \frac{\left({}^{n}/_{2} - g\right) \times W}{f}$$

L : کران پایین دسته ای که میانه در آن قرار دارد.
n : تعداد داده ها
g : فراوانی تجمیعی دسته ما قبل میانه
f : فراوانی دسته میانه
W : طول هر دسته

مد

داده ای که فراوانی آن از سایر داده ها بیشتر باشد را نما یا مد می نامیم. اگر دو داده ی مجاور دارای بیشترین فراوانی باشند، نصف مجموع آن ها به عنوان مد انتخاب می شود، در غیر این صورت اگر دو داده مجاور نباشند هر دو به عنوان مد انتخاب می شوند.

انواع توزیع فراوانی



دامنه تغییرات و میانگین انحرافات

دامنه تغییرات : تفاوت میان بزرگترین و کوچکترین داده را به عنوان دامنه ی تغییرات داده ها می شناسیم.

$$R = Max(x_1, x_2, \dots, x_n) - Min(x_1, x_2, \dots, x_n)$$

میانگین انحرافات : به منظور دخالت تاثیر تمام داده ها می توان فاصله ی کلیه داده ها با میانگین را محاسبه نمود، که به آن انحراف از میانگین داده ها گوییم.

$$D = \frac{1}{n}\sum_{i=1}^{n} |x_i - \mu_A|$$

انحراف چارکی

در آمار توصیفی به هر یک از سه مقداری که یک مجموعه داده ی مرتب را به چهار قسمت مساوی تقسیم می کنند، چارک گفته می شود.

$$\text{انحراف چارکی} = \frac{|\text{میانگین چارک اول} - \text{میانگین چارک سوم}|}{2}$$

واریانس و انحراف استاندارد

واریانس : میانگین مجذور انحرافات را واریانس گویند.

$$\delta^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_A)^2$$

انحراف استاندارد : با جذر گرفتن از مقدار واریانس انحراف استاندارد بدست می آید.

$$\delta = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_A)^2}$$

ضریب تغییر :

$$\frac{\delta}{\mu}$$

کواریانس

اندازه ی تغییرات هماهنگ دو متغیر تصادفی را کواریانس گویند. اگر دو متغیر یکی باشند، کواریانس برابر با واریانس خواهد بود.

برای مقادیر ویژگی های x و y مقدار کواریانس برابر است با :

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

μ_x و μ_y به ترتیب میانگین برای دو ویژگی x و y را نشان می دهد.

x_i ویژگی x از داده i و y_i ویژگی y از داده i

کواریانس

اگر این مقدار صفر باشد،میان دو ویژگی همبستگی وجود ندارد و دو ویژگی دارای رابطه خطی نیستند. مقدار مثبت آن نشان می دهد با افزایش یکی، دیگری نیز افزایش میابد و مقدار منفی دلالت بر این دارد که افزایش یکی باعث کاهش دیگری می شود.

- برای داده ها با d بعد، یک ماتریس $d \times d$ به نام ماتریس کواریانس ویژگی داریم که:

$$S_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

که در آن x_{ki} و x_{kj} به ترتیب مقدار iام و jام از kامین نمونه داده است.

همبستگی

همبستگی به رابطه بین دو متغیر اشاره می کند، که می توان آن را با کمک نمودار پراکندگی نشان داد. ارزش مقداری آن را ضریب همبستگی می نامیم.

$$cor(x,y) = \frac{cov(x,y)}{\sqrt{cov(x,x)cov(y,y)}} = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

ضریب همبستگی دارای دامنه ای بین 1- و 1 است. مقدار صفر عدم همبستگی را نشان می دهد و مقادیر 1+ و 1- به ترتیب دلالت بر همبستگی کامل مثبت و همبستگی کامل منفی برای دو ویژگی فوق را دارد.