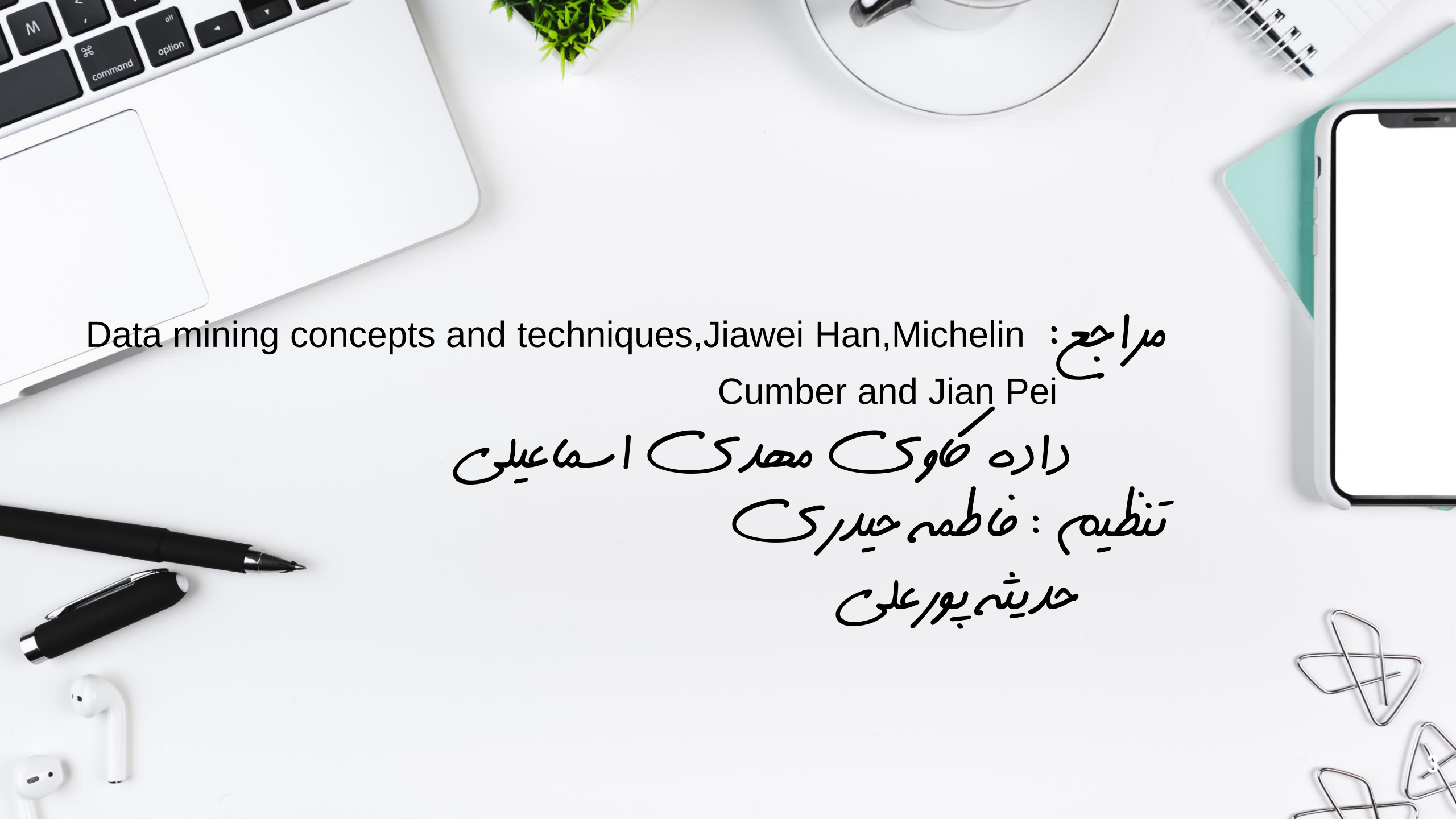

بسم الله الرحمن الرحیم

داده کاوی

دکتر حسام عمرانپور

دانشگاه صنعتی نوشیروانی بابل



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

داده کاوی مهدی اسماعیلی

تنظیم: فاطمه حیدری

حدیثه پورعلی

ادامه فصل 1

آماده سازی داده ها

الگوریتم Principal Component Analysis
تحلیل مولفه های اساسی

0- داده ها نرمال می شوند (اختیاری)

1- ماتریس کواریانس ویژگی ها ساخته می شود.

2- بردار های ویژه و مقدار های ویژه این ماتریس بدست می آید

C * Evector = λ * 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 یک Evector بدست آید.

C: ماتریس کواریانس ویژگی ها λ: مقدار ویژه

Evector: بردار ویژه (d*1)

الگوریتم Principal Component Analysis
تحلیل مولفه های اساسی

3- λ_i ها (مقادیر) ویژه را مرتب میکنیم به صورت نزولی

$\lambda_1 > \lambda_2 > \ \ldots > \lambda_d$

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

k= کوچکترین مقداری که با آن رابطه مقابل برقرار باشد

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

الگوریتم Principal Component Analysis
تحلیل مولفه های اساسی

- $X_{new} = X_{old}$ * ماتریس نگاشت

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

Principal Component Analysis الگوریتم
تحلیل مولفه های اساسی

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

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

Mean of Matrix A

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

Covariance Matrix of A

الگوریتم Principal Component Analysis
تحلیل مولفه های اساسی

ادامه مثال:

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

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

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

Principal Component Analysis الگوریتم
تحلیل مولفه های اساسی

- ادامه مثال:

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

- با حل معادله بالا مقادیر ویژه به صورت زیر به دست می آید:

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

Eigenvalues

الگوریتم Principal Component Analysis
تحلیل مولفه های اساسی

- ادامه مثال:
- سپس با قرار دادن مقادیر ویژه در معادله زیر بردار های ویژه بدست می آیند:

C * Evector = λ * Evector

- بردارهای ویژه:

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

الگوریتم Principal Component Analysis
تحلیل مولفه های اساسی

- ادامه مثال:
- مقادیر ویژه به ترتیب زیر مرتب میشوند:

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

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

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

الگوریتم Principal Component Analysis
تحلیل مولفه های اساسی

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

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

توزیع نرمال

- اغلب متغیرهای تصادفی پیوسته در طبیعت توزیع نرمال دارند.

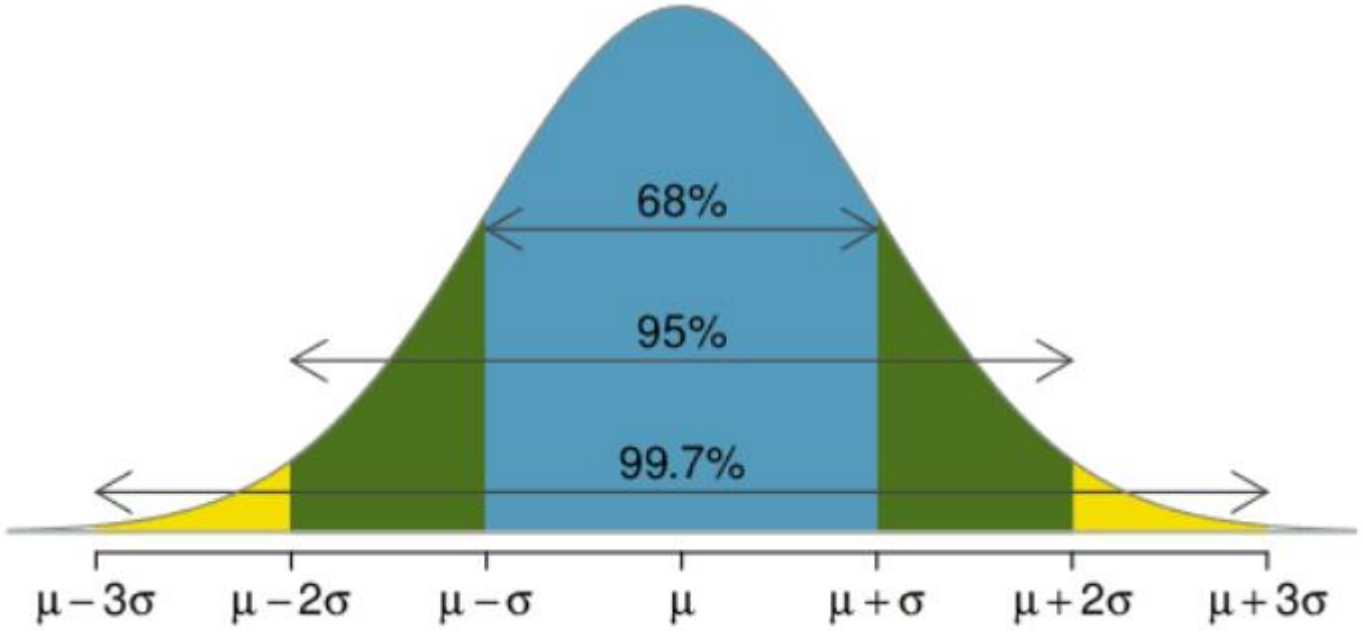
- $P(x)=\frac{1}{\sqrt{2\pi\delta^2}} * e^{\frac{-(x-\mu)^2}{2\delta^2}}$

- μ: میانگین

- δ^2: واریانس

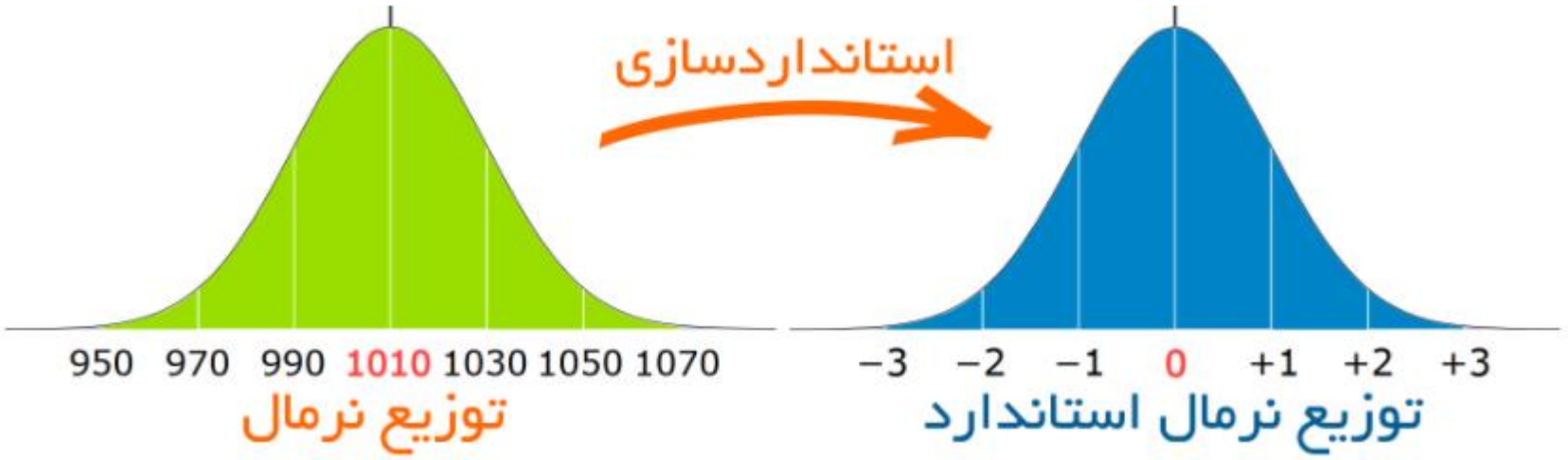
- این توزیع شامل یک max (peak) است.

توزیع نرمال



توزیع نرمال استاندارد

- توزیع نرمال با $\mu=0$ و $\delta=1$ را توزیع نرمال استاندارد می نامند.



(پیش پردازش) Pre processing

1- Data Cleaning

2- Data Integration

3- Data Transformation

4- Data Reduction

1-Data cleaning

1-1: missing Data : ممکن است یک ویژگی از یک داده وجود نداشته باشد (اندازه گیری نکرده باشیم)

راه حل:

1- آن سطر را حذف میکنیم

2- اگر صورت مسئله دسته بندی است برای داده های هم کلاس میانگین گیری انجام می دهیم و آن مقدار را برای داده ی از دست رفته در نظر می گیریم.

(اگر صورت مسئله دسته بندی نیست میانگین تمام داده ها یا داده های نزدیک را در نظر میگیریم)

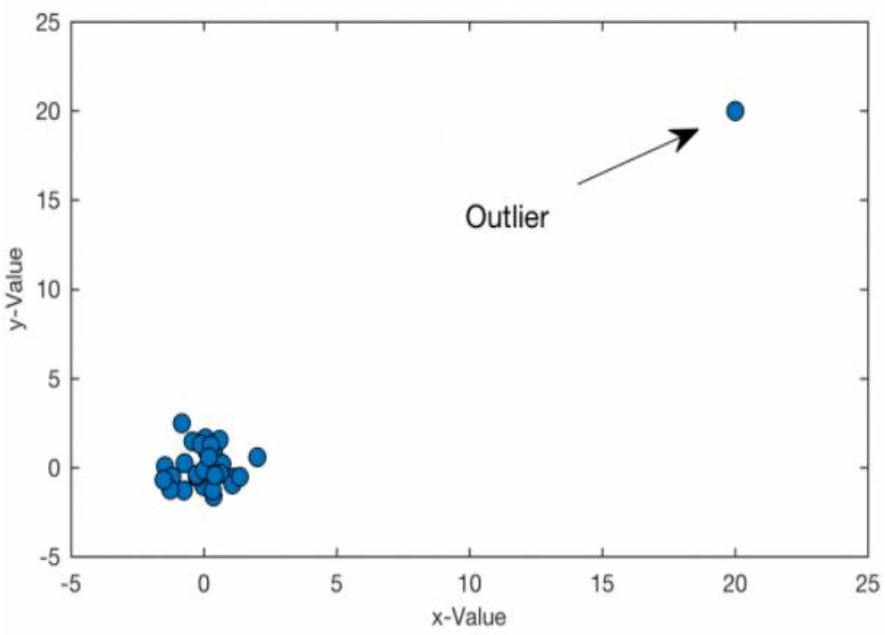
1-Data cleaning

مثال:

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

1-Data cleaning

2-1: outlier (دورافتاده – خارج از محدوده) :



$$\text{Threshold} = \text{Means} +(-)\ 2\delta$$

میانگین و انحراف معیار تمام داده ها را حساب میکنیم. داده هایی را که در بازه ی Threshold قرار گرفتند را به عنوان داده اصلی در نظر میگیریم.

2- Data Integration

وقتی که داده ها زیاد از حد باشند (یعنی داده های شبیه به هم زیاد باشند)

صورت مسئله با داده های نامتوازن

مثال: تشخیص سرطان که در داده ها، 5 نفر فوت شده و 400 نفر فوت نشده داریم، هر روش دسته بندی برچسب این مسئله را فوت نشده در نظر میگیرد.

راه حل: از هر کدام از داده هایی که در یک بازه هستند میانگین میگیریم و به عنوان نماینده آن بازه در نظر میگیریم تا بالانس شود.

3- Data Transformation
تغییر مقیاس داده ها

اگر دو ویژگی با مقیاس های مختلف داشته باشیم (در range های مختلف) بدون پیش پردازش معمولا تاثیر x_2 بیشتر از x_1 خواهد بود. لذا تمام ویژگی ها را در یک بازه تغییر مقیاس می دهیم.

مثال:

x_1	x_2
0.0001	100000
0.0002	200000
0.0001	300000

3- Data Transformation
تغییر مقیاس داده ها

1- مقیاس به [-1,1] :

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

2- مقیاس به [0,1] :

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

3- Data Transformation
تغییر مقیاس داده ها

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]:

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

3- Data Transformation
تغییر مقیاس داده ها

-5

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

4- Data Reduction

4-1: کاهش تعداد نمونه ها (سطرها):

4-1-1: **نمونه گیری تصادفی**: x% از داده ها را به صورت تصادفی انتخاب میکنیم

4-1-2: **نمونه گیری منظم**:

مثال: اگر 100 نمونه داده داشته باشیم از 10 داده اول یکی به صورت تصادفی انتخاب می شود، از 10 داده دوم یکی به صورت تصادفی انتخاب می شود، از 10 داده سوم یکی به صورت تصادفی انتخاب می شود ⟵ در نهایت 10 داده خواهیم داشت.

4-1-2: **نمونه گیری طبقه ای**: از هردسته به صورت تصادفی x% و یا تعداد مشخص انتخاب می شود

4-1-2: **نمونه گیری خوشه ای**: داده ها خوشه بندی می شوند و از هر خوشه یک یا چند داده انتخاب می شود

4- Data Reduction

4-2: کاهش تعداد ویژگی:

4-2-1: Feature Extraction:

4-2-1-1: خطی

بدون ناظر و برچسب: PCA

با ناظر و برچسب: LDA

4-2-1-2: غیرخطی

4- Data Reduction

:Feature Selection :4-2-2

4-2-2-1: بر اساس واریانس:

فرض کنید داده ها در دو کلاس A و B قرار دارند، مقدار زیر را برای هر ویژگی حساب میکنیم، ویژگی که مقدار کمتری دارد می تواند حذف شود(زمانی که توزیع نرمال باشد این روش می تواند روش مناسبی باشد).

$$S = \frac{|Mean(A) - Mean(B)|}{\sqrt{\frac{Var(A)}{N_1} + \frac{Var(B)}{N_2}}}$$

4- Data Reduction

4-2-2-1: بر اساس واریانس:

مثال:

x	y	class
3	7	A
2	9	B
6	6	A
5	5	A
8	7	B
4	9	A

$$S_x = \frac{|4.5 - 5|}{\sqrt{\frac{1.25}{4} + \frac{9}{2}}} = 0.22$$

$$S_y = \frac{|6.75 - 8|}{\sqrt{\frac{2.2}{4} + \frac{1}{2}}} = 1.21$$

4- Data Reduction

4-2-2-2: بر اساس آنتروپی:

شباهت $S_{ij} = e^{-\alpha D_{ij}} = e^{-\frac{D_{ij}}{2}}$

فاصله داده x_i از x_j $D_{ij} = \sqrt{\sum_{k=1}^{d} (\frac{x_{ik} - x_{jk}}{MAX\ k - MIN\ k})^2}$

اگر داده ها غیر عددی باشند $D_{ij} = \sum_{k=1}^{d} \frac{|x_{ik} \neq x_{jk}|}{d}$

$$\text{Entropy} = -\sum_{i=1}^{N-1} \sum_{j=i=1}^{N} [S_{ij} \times \log_2 S_{ij} + (1 - S_{ij}) \times \log_2 (1 - S_{ij})]$$

4- Data Reduction

4-2-2-2: بر اساس آنتروپی:

الگوریتم:

1- مجموعه را با تمام ویژگی ها قرار می دهیم.

2- آنتروپی مجموعه را محاسبه می کنیم.

3- با حذف هر ویژگی از مجموعه مجددا آنتروپی را محاسبه می کنیم.

4- ویژگی که با حذف آن کمترین فاصله بین آنتروپی های مجموعه و 3 حاصل شده است را به عنوان مجموعه جدید انتخاب می کنیم

5- آن قدر ادامه می دهیم تا به تعداد ویژگی مورد نظر برسیم.

- یک روش بالا به پایین است، درواقع در هر مرحله یک ویژگی حذف می گردد.
- ویژگی هایی باقی می مانند که بی نظمی موجود در دیتا را حفظ نمایند.

4- Data Reduction

4-3: گسسته سازی: