بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

درس : *یادگیری ماشین*

استاد : *خانم دکتر زرین بال ماسوله*

موضوع : *normalization methods*

دانشجو : *فرزاد محسنی*

مقطع : *کارشناسی ارشد*

رشته تحصیلی : *مهندسی آیتی*

کد دانشجویی :

دانشگاه : *صنعتی امیرکبیر*

سال تحصیلی : *1404-1403*

زمستان 1403

normalization methods for data

Min-Max Scaling (Rescaling)

- مقیاس بندی داده ها بین یک بازه مشخص **[0,1]**
- نرمال سازی برای هر ویژگی **(Feature)** به صورت جداگانه باید انجام شود.
- **min** و **max** را برای هر ویژگی **(Feature)** بدست می آوریم.
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم.
  - x' : مقدار نرمال‌شده داده پس از اعمال نرمال سازی **Min-Max**
  - x_{min} : کمترین مقدار ویژگی در مجموعه داده
  - x_{max} : بیشترین مقدار ویژگی در مجموعه داده

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Customer ID	Age	Purchase
P1	34	750,000
P2	51	250,000
P3	45	1,200,000
P4	22	500,000

Customer ID	Age	normalized_Age
P4	22	$x' = \frac{22-22}{51-22} = 0$
P1	34	$x' = \frac{34-22}{51-22} = 0.41$
P3	45	$x' = \frac{45-22}{51-22} = 0.79$
P2	51	$x' = \frac{51-22}{51-22} = 1$

Customer ID	Purchase	normalized_Purchase
P2	250,000	$x' = \frac{250000-250000}{1200000-250000} = 0$
P4	500,000	$x' = \frac{500000-250000}{1200000-250000} = 0.26$
P1	750,000	$x' = \frac{750000-250000}{1200000-250000} = 0.52$
P3	1,200,000	$x' = \frac{1200000-250000}{1200000-250000} = 1$

Z-Score Normalization (Standardization)

- بعد از استاندارد سازی برای هر ویژگی (Feature) میانگین صفر و انحراف معیار یک می شود
- نرمال سازی برای هر ویژگی (Feature) به صورت جداگانه باید انجام شود.
- μ (مو) و σ (سیگما) را برای هر ویژگی (Feature) بدست می آوریم.
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم.
  - x' : مقدار نرمال‌شده داده پس از اعمال نرمال سازی
  - μ : میانگین مقادیر ویژگی
  - σ : انحراف معیار مقادیر ویژگی

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2}$$

$$x' = \frac{x - \mu}{\sigma}$$

Customer ID	Age	Purchase
P1	34	750,000
P2	51	250,000
P3	45	1,200,000
P4	22	500,000

$$\mu_{age} = \frac{34 + 51 + 45 + 22}{4} = 38.00$$

$$\mu_{purchase} = \frac{750000 + 250000 + 1200000 + 500000}{4} = 675000.00$$

$$\sigma_{age} = \sqrt{\frac{(34 - 38.00)^2 + (51 - 38.00)^2 + (45 - 38.00)^2 + (22 - 38.00)^2}{3}} = 12.78$$

$$\sigma_{purchase} = \sqrt{\frac{(750000 - 675000.00)^2 + (250000 - 675000.00)^2 + (1200000 - 675000.00)^2 + (500000 - 675000.00)^2}{3}} = 405174.86$$

Customer ID	normalized_Age	normalized _Purchase
P1	$\frac{34-38}{12.78} = -0.31$	$\frac{750000-675000}{405174.86} = 0.18$
P2	$\frac{51-38}{12.78} = 1.01$	$\frac{250000-675000}{405174.86} = -1.04$
P3	$\frac{45-38}{12.78} = 0.54$	$\frac{1200000-675000}{405174.86} = 1.29$
P4	$\frac{22-38}{12.78} = -1.25$	$\frac{500000-675000}{405174.86} = -0.43$

$$\mu_{Z-Score\_Age} = \frac{Z_1 + Z_2 + Z_3 + Z_4}{4} = 0$$

$$\mu_{Z-ScorePurchase} = \frac{Z_1 + Z_2 + Z_3 + Z_4}{4} = 0$$

$$\sigma_{Z-Score\_Age} = \sqrt{\frac{(Z_1 - 0.0000000000)^2 + (Z_2 - 0.0000000000)^2 + (Z_3 - 0.0000000000)^2 + (Z_4 - 0.0000000000)^2}{3}} = 1$$

$$\sigma_{Z-ScorePurchase} = \sqrt{\frac{(Z_1 - 0.0000000000)^2 + (Z_2 - 0.0000000000)^2 + (Z_3 - 0.0000000000)^2 + (Z_4 - 0.0000000000)^2}{3}} = 1$$

Decimal Scaling

- بعد از استاندارد سازی برای هر ویژگی **(Feature)** تمام مقادیر در بازه **[−1,1]** یا **[0,1]** (بسته به این که داده منفی هم داریم یا نه) قرار می گیرند.
- نرمال سازی برای هر ویژگی **(Feature)** به صورت جداگانه باید انجام شود.
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم
  - x' : مقدار نرمال‌شده داده پس از اعمال نرمال سازی
  - j : تعداد ارقام بزرگ ترین مقدار مطلق در داده (تعداد ارقام عدد **M** را (در مبنای **10**) بشمارید؛ همان تعداد (یا اگر عدد اعشاری بود، قسمت بالاتر یا همان **⌈·⌉** را به عنوان **j** در نظر بگیرید یعنی)
    - به زبان ساده تر:
    - اگر عددی دارای بخش صحیح باشد، تعداد ارقام بخش صحیح را مبنا قرار می دهیم
    - اگر همه مقادیر عددی دارای مقدار اعشار باشند (مانند **0.02**، **0.1**، **0.003**) ، مقدار **j** بر اساس اولین رقم غیرصفر پس از ممیز انتخاب می‌شود

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

P1	P2
-0.34	750,000,000,000
0.034	-250,000
-0.0034	1,200,000
0.00034	-500,000

P1 = [−0.34 , 0.034 , −0.0034 , 0.00034]

- **|−0.34| = 0.34**
- **|0.034| = 0.034**
- **|−0.0034| = 0.0034**
- **|0.00034| = 0.00034**

Maximum absolute value: 0.34

j = 1

P2 = [750,000,000,000 , −250,000 , 1,200,000 , −500,000]

- **|750,000,000,000|=750,000,000,000**
- **|−250,000|=250,000**
- **|1,200,000|=1,200,000**
- **|−500,000|=500,000**

Maximum absolute value: 750,000,000,000

j = 12

Column	Maximum Absolute Value	j (Number of Digits)
P1	0.34	**1** (اولین رقم مهم بعد از اعشار)
P2	750,000,000,000	**12** (قسمت بدون اعشار 12 رقمی)

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

P1	normalized_P1
-0.34	$x' = \frac{-0.34}{10^1} = -0.034$
0.034	$x' = \frac{-0.034}{10^1} = 0.0034$
-0.0034	$x' = \frac{-0.0034}{10^1} = -0.00034$
0.00034	$x' = \frac{-0.00034}{10^1} = 0.000034$

P2	normalized_P2
750,000,000,000	$x' = \frac{750,000,000,000}{10^{12}} = 0.75$
-250,000	$x' = \frac{-250,000}{10^{12}} = -0.000000025$
1,200,000	$x' = \frac{1,200,000}{10^{12}} = 0.0000012$
-500,000	$x' = \frac{-500,000}{10^{12}} = -0.00000005$

Log Scaling (Log Transformation)

- نرمال سازی برای هر ویژگی **(Feature)** به صورت جداگانه باید انجام شود.
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم
  - x' : مقدار نرمال‌شده داده پس از اعمال نرمال سازی
  - b : پایه لگاریتم است (معمولاً از لگاریتم طبیعی **(e)** یا پایه ۱۰ استفاده می‌شود)
    - پایه **e** :
      - اگر داده های شما شامل مقادیر بسیار بزرگ و بسیار کوچک است
      - اگر می‌خواهید داده‌ها را برای درک بهتر انسان‌ها نمایش دهید
    - پایه **10** :
      - اگر داده های شما دارای رشد یا کاهش نمایی هستند
      - اگر مدل های آماری یا احتمال را بررسی می‌کنید
  - c : یک مقدار کوچک مثبت است که اضافه می شود تا از مشکلات ناشی از اعداد صفر یا منفی جلوگیری شود (لگاریتم برای مقادیر صفر و منفی تعریف نشده است)

$$x' = \log_b(x + c)$$

- $x' = \log_{10}(x + c)$
- $x' = \ln(x + c)$, $e(Euler's\ number) \approx 2.718$

مدیریت مقادیر منفی و صفر (اگر داده ها شامل مقدار صفر یا منفی باشند) :

- چون لگاریتم برای مقادیر منفی و صفر تعریف نشده است، باید همه داده‌ها را شیفت کنیم تا مثبت شوند
- $\mathbf{min}(X)$: کمترین مقدار موجود در ستون
- $|\mathbf{min}(X)| + 1$: اضافه می شود تا تمام داده ها مثبت شوند

$$x' = \log_{10}(x + |\mathbf{min}(X)| + 1)$$

P1	P2
-0.34	750,000,000,000
0.034	250,000
-0.0034	1,200,000
0.00034	500,000

P1 = [−0.34 , 0.034 , −0.0034 , 0.00034]

- **Minimum value: (P1) = −0.34**
- **Shift value: |min(P1)| + 1 = 0.34 + 1 = 1.34**

P1	normalized_P1
-0.34	$x' = \log_{10}(-0.34 + \lvert -0.34 \rvert + 1) = \log_{10}(1.00) = 0$
0.034	$x' = \log_{10}(0.034 + \lvert -0.34 \rvert + 1) = \log_{10}(1.374) = 0.138$
-0.0034	$x' = \log_{10}(-0.0034 + \lvert -0.34 \rvert + 1) = \log_{10}(1.3366) = 0.126$
0.00034	$x' = \log_{10}(0.00034 + \lvert -0.34 \rvert + 1) = \log_{10}(1.34034) = 0.127$

P2 = [750,000,000,000 , 250,000 , 1,200,000 , 500,000]

P2	normalized_P2
750,000,000,000	$x' = \log_{10}(750,000,000,000) = 11.87$
250,000	$x' = \log_{10}(250,000) = 5.39$
1,200,000	$x' = \log_{10}(1,200,000) = 6.07$
500,000	$x' = \log_{10}(500,000) = 5.69$

Robust Scaling

- مقیاس بندی مقاوم (Robust Scaling) یکی از روش های نرمال سازی ویژگی‌ها است که در برابر مقادیر پرت (Outliers) مقاوم است
- نرمال سازی برای هر ویژگی (Feature) به صورت جداگانه باید انجام شود.
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم
  - x' : مقدار نرمال‌شده داده پس از اعمال نرمال سازی
  - $\text{median}(X)$: میانه (مرکز داده‌ها) را نشان می دهد
  - $\text{IQR}(X)$: بازه بین چارکی (Interquartile Range)
    - مقیاسی برای توزیع داده ها استفاده می‌شود و نشان می‌دهد که ۵۰٪ میانی داده ها (بین چارک اول و سوم) در چه محدوده ای قرار دارند
      - Q1 (چارک اول) :
        - مقدار ۲۵٪ پایین داده ها مرز بین ۲۵٪ کمترین مقادیر و ۷۵٪ بقیه
      - Q3 (چارک سوم) :
        - مقدار ۷۵٪ پایین داده ها مرز بین ۷۵٪ کمترین مقادیر و ۲۵٪ بیشترین مقادیر
      - IQR :
        - فاصله بین چارک اول و چارک سوم که محدوده ای برای بخش میانی داده ها را نشان می دهد
      - $\text{IQR}(X) = Q3 - Q1$

$$x' = \frac{x - \text{median}(X)}{\text{IQR}(X)}$$

P1	P2
0.56	75
0.28	-70
-0.0006	80
0.32	0
-0.89	-500
0	

P1 = [0.56 , 0.28 , −0.0006 , 0.32 ,−0.89 ,0]

- مرتب سازی داده ها بترتیب صعودی
  - **−0.89 , −0.0006 , 0 , 0.28 , 0.32 , 0.56**
- یافتن میانه
  - $\text{Median} = \frac{0+0.28}{2} = 0.14$
- محاسبه چارک ها
  - مقدار میانه از نیمه پایینی داده ها : چارک اول **(Q1)**
    - **(-0.89 , -0.0006 , 0):**
      - **Q1 = −0.0006**
  - مقدار میانه از نیمه بالایی داده‌ها : چارک سوم **(Q3)**
    - **(0.28 , 0.32 , 0.56):**
      - **Q3 = 0.32**
- محاسبه **IQR**
  - **IQR = Q3 − Q1 = 0.32 − (−0.0006) = 0.3206**

$$x' = \frac{x - 0.14}{0.3206}$$

P1	normalized_P1
0.56	$x' = \frac{0.56 - 0.14}{0.3206} = 1.31$
0.28	$x' = \frac{0.28 - 0.14}{0.3206} = 0.44$
-0.0006	$x' = \frac{-0.0006 - 0.14}{0.3206} = -0.44$
0.32	$x' = \frac{0.32 - 0.14}{0.3206} = 0.56$
-0.89	$x' = \frac{-0.89 - 0.14}{0.3206} = -3.21$
0	$x' = \frac{0 - 0.14}{0.3206} = -0.44$

P2 = [75 , −70 , 80 , 0 , −500]

- مرتب سازی داده ها بترتیب صعودی
  - **−500 , −70 , 0 , 75 , 80**
- یافتن میانه
  - **Median = 0**
- محاسبه چارک ها
  - مقدار میانه از نیمه پایینی داده ها : چارک اول **(Q1)**
    - **(-500 , -70 , 0):**
      - **Q1 = −70**
  - مقدار میانه از نیمه بالایی داده‌ها : چارک سوم **(Q3)**
    - **(0 , 75 , 80):**
      - **Q3 = 75**
- محاسبه **IQR**
  - **IQR = Q3 − Q1 = 75 − (−70) = 145**

$$x' = \frac{x - 0}{145}$$

P2	normalized_P2
75	$x' = \frac{75 - 0}{145} = 0.52$
-70	$x' = \frac{-70 - 0}{145} = -0.48$
80	$x' = \frac{80 - 0}{145} = 0.55$
0	$x' = \frac{0 - 0}{145} = 0.00$
-500	$x' = \frac{-500 - 0}{145} = -3.45$

Max-Abs Scaling

- نرمال سازی برای هر ویژگی **(Feature)** به صورت جداگانه باید انجام شود.
- خروجی داده ها همیشه در بازه **[−1,1]** قرار دارد
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم
  - x' : مقدار نرمال‌شده داده پس از اعمال نرمال سازی
  - $\max|X|$: بزرگ ترین مقدار مطلق در ویژگی است

$$x' = \frac{x}{\max|X|}$$

P1	normalized_P1
-50	$x' = \frac{-50}{100} = -0.50$
-20	$x' = \frac{-20}{100} = -0.2$
0	$x' = \frac{0}{100} = 0.00$
10	$x' = \frac{10}{100} = 0.10$
30	$x' = \frac{30}{100} = 0.30$
100	$x' = \frac{100}{100} = 1.00$

P1 = [−50 , −20 , 0 , 10 , 30 , 100]

max|X| = 100

$$x' = \frac{x}{100}$$

L1 Normalization (Manhattan Scaling)

- نرمال سازی **L1** که با نام مقیاس بندی منهتن **(Manhattan Scaling)** نیز شناخته می‌شود ، یک روش نرمال سازی ویژگی ها است که بردار داده ها را طوری مقیاس بندی می‌کند که نُرم **L1** (مجموع قدرمطلق مقادیر) برابر ۱ شود
- نرمال سازی برای هر ویژگی **(Feature)** به صورت جداگانه باید انجام شود
- مجموع قدرمطلق مقادیر در بردار نرمال شده برابر ۱ می شود
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم
  - x' : مقدار نرمال شده داده پس از اعمال نرمال سازی
  - $\| X \|_1$: نرم **L1** (نرم منهتن) داده ها

$$x' = \frac{x}{\| X \|_1}$$

$$\| X \|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

P1	normalized_P1
-52	$x' = \frac{-52}{109.45} = -0.4751$
-0.25	$x' = \frac{-0.25}{109.45} = -0.0023$
0	$x' = \frac{0}{109.45} = 0$
32	$x' = \frac{32}{109.45} = 0.2924$
25.2	$x' = \frac{25.2}{109.45} = 0.2302$

P1 = [–52 , –0.25 , 0 , 32 , 25.2]

- $\| X \|_1 = |-52| + |-0.25| + |0| + |32| + |25.2| = 52 + 0.25 + 0 + 32 + 25.2 = 109.45$
- $x' = \frac{x}{109.45}$

$$|-0.4751| + |-0.0023| + |0| + |0.2924| + |0.2302| = 1$$

Unit Vector Scaling (L2 Normalization)

- مقیاس بندی بردار واحد **(L2 Normalization)** یک روش نرمال سازی ویژگی‌ها است که یک بردار را به گونه‌ای تبدیل می‌کند که نرم **L2** (یا نرم اقلیدسی) آن برابر ۱ شود
- نرمال سازی برای هر ویژگی **(Feature)** به صورت جداگانه باید انجام شود.
- مجموع مربع مقادیر نرمال شده برابر ۱ می شود
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم
  - x' : مقدار نرمال شده داده پس از اعمال نرمال سازی
  - $\| X \|_2$: نرم **L2** (نرم اقلیدسی) داده ها

$$x' = \frac{x}{\| X \|_2}$$

$$\| X \|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

P1	normalized_P1
-52	$x' = \frac{-52}{66.025} = -0.787$
-0.25	$x' = \frac{-0.25}{66.025} = -0.0038$
0	$x' = \frac{0}{66.025} = 0$
32	$x' = \frac{32}{66.025} = 0.484$
25.2	$x' = \frac{25.2}{66.025} = 0.382$

P1 = [–52 , –0.25 , 0 , 32 , 25.2]

$$\| X \|_2 = \sqrt{(-52)^2 + (-0.25)^2 + (0)^2 + (32)^2 + (25.2)^2} = \sqrt{4363.1025} = 66.025$$

$$x' = \frac{x}{\| X \|_2} = \frac{x}{66.025}$$

$$(-0.787)^2 + (-0.0038)^2 + (0)^2 + (0.484)^2 + (0.382)^2 = 1$$

Mean Normalization

- نرمال سازی برای هر ویژگی **(Feature)** به صورت جداگانه باید انجام شود
- نرمال‌سازی میانگین باعث می‌شود که مقادیر حول صفر متمرکز شوند
- داده‌های نرمال‌شده در بازه $[-1,1]$ قرار می‌گیرند (اگر داده ها متقارن باشند)
- حالا استفاده از رابطه زیر :
  - x : مقدار اصلی داده ای که می‌خواهیم نرمال کنیم
  - x' : مقدار نرمال شده داده پس از اعمال نرمال سازی
  - μ : میانگین مقادیر ویژگی
  - x_{max} : بزرگ ترین مقدار در ویژگی
  - x_{min} : کوچک ترین مقدار در ویژگی

$$x' = \frac{x - \mu}{x_{max} - x_{min}}$$

P1	normalized_P1
-52	$x' = \frac{-52 - 1.99}{84} = -0.641$
-0.25	$x' = \frac{-0.25 - 1.99}{84} = -0.027$
0	$x' = \frac{0 - 1.99}{84} = -0.024$
32	$x' = \frac{32 - 1.99}{84} = 0.357$
25.2	$x' = \frac{25.2 - 1.99}{84} = 0.276$

P1 = [–52 , –0.25 , 0 , 32 , 25.2]

- $\boldsymbol{\mu = \frac{-52+(-0.25)+0+32+25.2}{5} = \frac{4.95}{5} = 1.99}$
- $\boldsymbol{x' = \frac{x-1.99}{x_{max}-x_{min}}}$
- $\boldsymbol{x_{max} = 32, \quad x_{min} = -52}$
- $\boldsymbol{\frac{x-1.99}{32-(-52)} = \frac{x-1.99}{84}}$