TEXT

On Given dataset apply different normalization techniques

2 June

* Independent Component Analysis: ICA

Cocktail party problem

Changing signals: Signals that change over time

- Dimensionality reduction by identifying independent components

observed signal $x = As$ → mixing matrix, → source

Recovering original: $s = Wx$ → demixing/weight.

$(W = A^{-1})$

* Data Normalization:

1. Min-max normalization $[0, 1]$.

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

Data: $x' = 200, 300, 400, 600, 1000$ (min, max)

$$x' = \frac{200-200}{1000-200}, \frac{300-200}{1000-200}, \frac{400-200}{1000-200}, \frac{600-200}{1000-200}, \frac{1000-200}{1000-200}$$

$$x' = 0, \frac{100}{800}, \frac{200}{800}, \frac{400}{800}, 1$$

$$x' = 0, 0.125, 0.25, 0.5, 1$$

2. z-score normalization: Mean- standard deviation:

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

$x = 200 ; 300 ; 400 ; 600 ; 1000 \quad \mu = 500.$

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}} = \sqrt{\frac{90000 + 40000 + 10000 + 10000 + 250000}{5}} = \sqrt{80000}$$

$\sigma = 283.$

$$z = \frac{200 - 500}{283} ; \frac{300 - 500}{283} ; \frac{400 - 500}{283} ; \frac{600 - 500}{283} ; \frac{1000 - 500}{283}.$$

$z = -1.06 ; -0.71 ; -0.35 ; 0.35 ; 1.77.$

3. z-score normalization: Mean - Absolute deviation:

$$z = \frac{x - \mu}{A}.$$

$$A = |200 - 500| + |300 - 500| + |400 - 500| + |600 - 500| + |1000 - 500| \; / 5$$

$A = 300 + 200 + 100 + 100 + 500 / 5 = 1200 / 5 = 240.$

$$z = \frac{200 - 500}{240} ; \frac{300 - 500}{240} ; \frac{400 - 500}{240} ; \frac{600 - 500}{240} + \frac{1000 - 500}{240}.$$

$z = -1.25 ; -0.83 ; -0.42 ; 0.42 ; 2.1.$

4. Normalization by Decimal Scaling:

- find value of j.
- Smallest integer j such that

$$\text{Max}\left(\frac{|v_i|}{10^j}\right) \leq 1$$

$\Rightarrow v = -200 ; 300 ; 400 ; 600 ; 1000$

$\frac{200}{10^3} \leq 1 = 0.2$ $\quad \frac{400}{10^3} \leq 1 = 0.4$ $\quad \frac{1000}{10^3} = 1$

$\frac{300}{10^3} \leq 1 = 0.3$ $\quad \frac{600}{10^3} \leq 1 = 0.6$

$v' = 0.2 ; 0.3 ; 0.4 ; 0.6 ; 1.$

* Binning:
- A form of quantization.
- Putting into buckets.
- Reduces overfitting.
- Smoothening data.

1- Equal frequency binning:

Input: [5, 10, 11, 13, 15, 35, 50, 55].

↳ Bin 1: [5, 10, 11, 13]

Bin 2: [35, 35, 50, 55]

2- Equal width binning: $w = \frac{\text{max} - \text{min}}{\text{no. of bins}}$

Bin 1: [min + w].

Bin 2: [min + 2w]

Bin n: [min + nw].

3 Customized binning:
Custom.

* Outliers:
Deviated datapoints.

Graphically: Scatter plot, box plot.

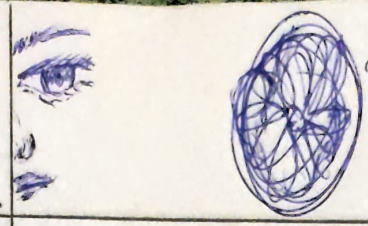
→ IQR: Inter Quartile Range.

1. Sort the data from low to high.
2. Identify the first Quartile (Q_1), the median, third Quartile (Q_3).
3. Calculate $IQR = Q_3 - Q_1$.
4. Calculate upper fence = $Q_3 + (1.5 * IQR)$.
5. Calculate lower fence = $Q_1 - (1.5 * IQR)$.
6. Use fences to highlight any outliers, all values that fall outside the fences.
The outliers are those values that are greater than the upper fence and lesser than the lower fence.

ex 1. ~~Sorted data: 22 24 26~~

26 37 24 28 35 22 31 53 41 64 29

1. Sorted: 22 24 26 28 29 31 35 37 41 53 64
2. $Q_1 = 26$, $Q_3 = \frac{41+37}{2} = 39$
3. $IQR = 39 - 26 = 13$.
4. Upper = $39 + (1.5 \times 13) = 58.5$; Lower = $26 - (1.5 \times 13) = 6.5$
5. Outliers = 64



__/__/__

29 jan

* Z-score based Outliers or Anomaly Detection.

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

Z threshold = ±3, ±2, ±3.5, ±4.

Qu.

100	150	120	125	140	130	140	135	130	150
140	100	95	80	120	125	130	100	140	135
130	145	110	120	130	135	140	125	130	120.

μ = 124.64. σ = 16.96. $= \sqrt{\frac{\Sigma(x_i-\mu)}{N}}$.

$Z_{100} = \frac{100-124.64}{16.96} = -1.45$. $Z_{125} = 0.02$.

$Z_{150} = 1.49$. $Z_{120} = -0.275$. $Z_{140} = 0.904$.

$Z_{130} = 0.314$. $Z_{110} = -0.865$. $Z_{135} = 0.61$.

$Z_{95} = -1.45$. $Z_{80} = -2.639$. $Z_{145} = 1.198$.

Threashold ±3 : No outliers.

±2 : 80 is the outlier.

* LOF :

Qu. K = 2 nearest neighbours.

A(1,2) B(2,3) C(3,4) D(10,10).

① Euclidean distance formula : $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$.

__/__/__

A(1,2) B(2,3) C(3,4) D(10,10)

$AB = \sqrt{(2-1)^2+(3-2)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.4$.

$BC = \sqrt{(3-2)^2+(4-3)^2} = \sqrt{1+1} = \sqrt{2} = 1.4$

$CD = \sqrt{(10-3)^2+(10-4)^2} = \sqrt{49+36} = 9.22$.

$AC = \sqrt{(3-1)^2+(4-2)^2} = \sqrt{4+4} = 2.83$.

$AD = \sqrt{(10-1)^2+(10-2)^2} = \sqrt{81+64} = 12.04$.

$BD = \sqrt{(10-2)^2+(10-3)^2} = \sqrt{64+49} = 10.63$.

① Finding nearest neighbours:

A: B & C. C: B & A.

B: C & A. D: C & B.

② Reachability: reach dist (p,q) = max (dist (p,q), k distance). (k distance = neighbourhood distance)

reach_dist (A,B) = max (1.4, 2.83) = 2.83.

reach_dist (A,C) = max (2.83, 1.4) = 2.83

~~reach_dist (A,D) = max (12.04,~~

reach_dist (B,A) = max (1.4, 2.83) = 2.83

reach_dist (B,C) = max (1.4, 1.4) = 1.4

reach_dist (C,B) = 1.4; reach_dist (C,A) = 2.83.

reach_dist (D,C) = max (9.22, 10.63) = 10.63;

reach_dist (D,B) = max (10.63, 1.4) = 10.63.

? AD or CD?

④ $$LRD(p) = \frac{1}{\frac{1}{k}\sum_{q \in N_k(p)} reach\text{-}dist(p,q)}$$

$$LRD(A) = \frac{1}{\frac{1}{2}[1.4+2.83]} = \underline{0.47}$$

$$LRD(B) = \frac{1}{\frac{1}{2}[1.4+2.83]} = \underline{0.47}$$

$$LRD(C) = \frac{1}{\frac{1}{2}[2.83+2.83]} = \underline{0.35}$$

$$LRD(D) = \frac{1}{\frac{1}{2}[9.22+10.63]} = \underline{0.1}$$

⑤ $$LOF(p) = \frac{\frac{1}{k}\sum_{q \in N_k(p)} LRD(q)}{LRD(p)}$$

* Modified z-score:

$$M_i = \frac{0.6745 \times (x_i - Median)}{MAD}$$

where x_i - data point.

MAD - Median ($|x_i - Median|$).

- Median Absolute Deviation.

0.6745 is a scaling factor.

__/__/____

Qn. Data points = 10 12 14 15 100.

Median = 14.

MAD = Median ($|10-14|, |12-14|, |14-14|, |15-14|, |100-14|$)

MAD = Median (4, 2, 0, 1, 86) ⇒ Median (0, 1, 2, 4, 86).

MAD = 2.

Mi~~s~~ score = $\frac{0.6745 \times (10-14)}{2} = -1.349$.

$M_2 = \frac{0.6745 \times (12-14)}{2} = -0.6745$.

$M_3 = 0$. $M_4 = \frac{0.6745 \times (15-14)}{2} = 0.33725$.

$M_5 = \frac{0.6745 \times (100-14)}{2} = 29.0035$.

Qn. n=7; 2D plane. k=2.

X = A(1,1) B(2,2) C(2,3) D(3,3)
E(3,4) F(8,8) G(100,100).

① Euclidean distance:

$AB = \sqrt{(2-1)^2+(2-1)^2} = 1.414$. $AE = \sqrt{(3-1)^2+(4-1)^2} = 3.6$.

$AC = \sqrt{(2-1)^2+(3-1)^2} = 2.24$. $AF = \sqrt{(8-1)^2+(8-1)^2} = 9.9$.

$AD = \sqrt{(3-1)^2+(3-1)^2} = 2.83$. $AG = \sqrt{(100-1)^2+(100-1)^2} = 140.007$.

$BC = \sqrt{0+1^2} = \sqrt{1} = 1$. $BD = \sqrt{1^2+1^2} = 1.414$. $BE = \sqrt{1^2+2^2} = 2.24$.

$BF = \sqrt{6^2+6^2} = 8.5$. BG = ~~[illegible]~~ 138.6.

$CD = \sqrt{1^2+0} = 1$. ~~[illegible]~~ $CE = \sqrt{1^2+1^2} = 1.414$.

$CF = \sqrt{6^2+5^2} = 7.81$. CG = ~~[illegible]~~ 137.88.

__/__/__

DE = $\sqrt{0+1^2}$ = 1. DF = $\sqrt{5^2+5^2}$ = 4.1. DG = 137.[illegible]2.

EF = $\sqrt{5^2+4^2}$ = 6.4 EG = 136.47. FG = 130.1.

② Nearest neighbours:

A = B, C = [1.414, 2.24].
B = A, C = [1.414, 1].
C = B, D = [1, 1].
D = C, E = [1, 1].
E = C, D = [1.414, 1].
F = D, E = [4.1, 6.4]
G = E, [illegible]F = [130.1, 136.47]. [136.47, 130.1].

③ Reach distance: (p,q) = max (dist(pq), q's nearest neighbour) [first occurrence]

A → B = max(1.414, 1) = 1.414.
A → C = max(2.24, 2[illegible]) = 2.24.

B → A = 1.414 ; B → C = max(1, [illegible] 2.24) = 2.24.

C → B = 1.414 ; C → D = max(1, 1) = 1.

D → C = 2.24 ; D → E = max(1, 1[illegible]) = 1.

E → [illegible]C = max(1.414, [illegible]) = [illegible] ; E → D = 1.

F → D = max(4.1, 1) = 4.1 ; F → E = max(6.4, 1) = 6.4.

G → E = max(136.47, 1) = 136.47. ; G → F = max(130.1, 130.1) = 130.1

4 Feb

④ $LRD = \dfrac{1}{\frac{1}{k}\sum \text{reachdist}(p,q)}$

A: $1/\frac{1.414+2.24}{2} = 1/1.827 = 0.547$

B: $1/\frac{1.414+2.24}{2} = 1/1.827 = 0.547$

C: $1/\frac{1.414+1}{2} = 1/1.207 = 0.8285$

D: $1/\frac{2.24+1}{2} = 1/1.62 = 0.6173$

E: $1/\frac{2.24+1}{2} = 1/1.62 = 0.6173$

F: $1/\frac{7.1+6.4}{2} = 1/6.75 = 0.148$

G: $1/\frac{136.47+130.1}{2} = 1/133.285 = 0.0075$

⑤ $LOF = \dfrac{\frac{1}{k}\sum LRD(q)}{LRD(p)}$

A: $\frac{1}{2}\left[\frac{LRD(B)+LRD(C)}{LRD(A)}\right] = \frac{1}{2}\left(\frac{0.547+0.83}{0.547}\right) = 1.26$

B: $\frac{1}{2}\left(\frac{A+C}{B}\right) = \frac{1}{2}\left(\frac{0.547+0.83}{0.547}\right) = 1.26$

C: $\frac{1}{2}\left(\frac{B+D}{C}\right) = \frac{1}{2}\left(\frac{0.547+0.6173}{0.8285}\right) = 0.7$

D: $\frac{1}{2}\left(\frac{C+E}{D}\right) = \frac{1}{2}\left(\frac{0.8285+0.6173}{0.6173}\right) = 1.171$

E: $\frac{1}{2}\left(\frac{C+D}{E}\right) = \frac{1}{2}\left(\frac{0.8285+0.6173}{0.6173}\right) = 1.171$

F: $\frac{1}{2}(D,E)/F = \frac{1}{2}(0.6173+0.6173)/0.148 = 4.171$

G: $\frac{1}{2}(E,F)/G = \frac{1}{2}(0.6148+0.148)/0.0075 = 51.05$

on. LDA, PCA, performance evaluations, overfitting/underfitting-theory..

__/__/__

∴ The outliers are F & G. (mostly G).

* ~~Types of Kernels~~:

Lower dimension → higher dimension, and the linearly separate it. Used for classification.

- Types of Kernels:

① Linear Kernel: $K(x,y) = x^T y$.

$K(x,y) = \phi(x) \cdot \phi(y) = x^T y$.

② Polynomial kernel: $K(x,y) = (x^T y)^q$.

q - degree of polynomial.

Homogeneous kernel.

For inhomogeneous kernel:

$K(x,y) = (c + x^T y)^q$.

c - constant.

q - degree of polynomial.

③ Gaussian / RBF Kernel: Radial basis functions.

$K(x,y) = e^{-(x-y)^2 / 2\sigma^2}$.

σ - constant.

If γ = small, RBF similar to SVM.

γ = large, kernel is influenced by more support vectors.

④ Sigmoid kernel: $K(x_i, y_i) = \tanh(\kappa x_i y_i - \sigma)$

Qn. Consider the 2 points $x(1,2)$, $y(2,3)$; $\sigma = 1$.
Apply RBF kernel, find the value of RBF kernel for these points.

$$K(x,y) = e^{-\frac{(x-y)^2}{2\sigma^2}}$$

$$= e^{-\frac{(1,2)-(2,3)}{2}} = e^{-\frac{(1-1)^2}{2}} = e^{-\frac{0}{2}} = e^{0} = 1.$$

$(1-2)^2 + (2-3)^2 = 2$. $e^{-2/2} = \underline{0.36}$.

Qn. $x(1,2)$ $y(2,3)$. $c=1$. $\therefore q=2$.
Linear, homo and inhomo

Linear = $x^T y = \begin{bmatrix} 1 \\ 2 \end{bmatrix} [2 \ 3] = \underline{8}$.

Homogeneous = $(x^T y)^2 = 8^2 = \underline{64}$.

Inhomogeneous = $(c + x^T y)^2 = (1+8)^2 = 9^2 = \underline{81}$.

6 Feb.

Q1: i. Consider:

y	3	5	7	2	6
ŷ	2.8	5.2	6.9	2.1	5.7

Compute the evaluation metrics including MAE, MSE, RMSE & RelMSE.

ii. Apply LDA on:

$x_1 = \{(2,3)\ (3,4)\ (4,5)\ (5,6)\}$

$x_2 = \{(7,8)\ (8,9)\ (9,10)\ (10,11)\}$.

iii. y: true label: 1 0 1 0 1 1 0 1 0 1

y: ~~predicted~~ pred: 1 1 1 0 0 1 0 1 0 1

Evaluate confusion matrix, accuracy, precision, recall and F1 score.

→ i. $MAE = \frac{1}{n}\sum_{i=0}^{n-1} |y_a - y_p| = \frac{1}{5}[0.2 + (-0.2) + 0.1 + 0.1 + 0.3]$

$= 0.18$.

$MSE = \frac{1}{n}\sum_{i=0}^{n-1} (y_a - y_p)^2 = \frac{1}{5}(0.2^2 + (-0.2)^2 + 0.1^2 + (-0.1) + 0.3^2)$

$= \frac{1}{5}(0.19) = 0.038$.

$RMSE = \sqrt{0.038} = 0.195$

Ans = 16.

$RelMSE = \frac{\sum (y_a - y_p)^2}{\sum (y_a - y_{avg})^2} = \frac{0.19}{(2.56 + 0.16 + 5.76 + 6.76 + 1.96)}$

$= \frac{0.19}{17.2} = 0.011$.

ii. LDA: $x_1 = \{(2,3)\ (3,4)\ (4,5)\ (5,6)\}$ N=4.

$x_2 = \{(7,8)\ (8,9)\ (9,10)\ (10,11)\}$.

① $\mu_1 = 1/N \sum x_i = 1/4(14, 18) = (3.5, 4.5)$.

$\mu_2 = 1/N \sum x_i = 1/4(34, 38) = (8.5, 9.5)$.

② $S_W = S_1 + S_2$.

$S_i = \frac{1}{N-1} \sum (x-\mu_i)(x-\mu_i)^T$.

$= 1/3 \left[\begin{bmatrix} -1.5 \\ -1.5 \end{bmatrix} \begin{bmatrix} -1.5 & -1.5 \end{bmatrix} + \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix} \begin{bmatrix} -0.5 & -0.5 \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} + \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix} \begin{bmatrix} 1.5 & 1.5 \end{bmatrix} \right]$.

$= \frac{1}{3} \begin{bmatrix} 2.25 & 2.25 \\ 2.25 & 2.25 \end{bmatrix} + \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix} + \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix} + \begin{bmatrix} 2.25 & 2.25 \\ 2.25 & 2.25 \end{bmatrix}$.

$= \frac{1}{3} \begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix} = \begin{bmatrix} 1.67 & 1.67 \\ 1.67 & 1.67 \end{bmatrix}$.

$S_2 = \frac{1}{N-1} \sum (x-\mu_2)(x-\mu_2)^T$.

$= \begin{bmatrix} 1.67 & 1.67 \\ 1.67 & 1.67 \end{bmatrix}$.

$S_W = S_1 + S_2 = 2 \begin{bmatrix} 1.67 & 1.67 \\ 1.67 & 1.67 \end{bmatrix} = \begin{bmatrix} 3.34 & 3.34 \\ 3.34 & 3.34 \end{bmatrix}$.

③ $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \begin{bmatrix} -5 \\ -5 \end{bmatrix} \begin{bmatrix} -5 & -5 \end{bmatrix}$.

$= \begin{bmatrix} 25 & 25 \\ 25 & 25 \end{bmatrix}$.

④ Eigen values: $|S_W^{-1} S_B - \lambda I| = 0$.

$|S_W^{-1} S_B - \lambda I| \Rightarrow$ Since $\det(S_W) = 0$, S_W^{-1} does not exist.

iii. y-true : 1 0 1 0 1 1 0 1 0 1.
y-pred : 1 1 1 0 0 1 0 1 0 1

Confusion matrix :

		True 0	True 1	
Predicted	0	3 TN	1 FN	4
	1	1 FP	5 TP	6
		4	6	10

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{3+5}{10} = \frac{8}{10} = 0.8 = 80\%.$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{5}{5+1} = \frac{5}{6} = 0.83 = 83\%.$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{5}{5+1} = \frac{5}{6} = 0.83 = 83\%.$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot 0.83 \cdot 0.83}{0.83 + 0.83} = 0.83 = 83\%.$$