Preprocessing

Data Transformation: Normalization/Scaling/Standardization

Data Cleaning: Incomplete, Noisy, and Inconsistent data

Scaling and Normalization

Why scale or normalize data?

Example:

	Marks1	Marks2	Marks3
Student 1	280	70	60
Student 2	200	60	55
Student 3	270	40	30

When you pass this data to a Machine Learning algorithm like kNN, kMeans or Neural Networks, the model would look at Marks1 and see that it is a higher value than Marks2 or Marks3 all around, so model thinks maybe it is of higher importance, which may not be the actual case.

Scaling and Normalization

Why scale or normalize data?

Example:

	Marks1	Marks2	Marks3
Student 1	280	70	60
Student 2	200	60	55
Student 3	270	40	30

Euclidean distance (s1,s2) = 80.77

Euclidean Distance (s1,s3) = 43.58

Euclidean Distance (s2,s3) = 71.58

Euclidean distance between Student-1 and Student-2 is being dominated by the marks1.

Same is the case for Euclidean Distance calculation between student-2 and student-3. There also the e. distance is being high because of difference in the marks-1 attribute.

If we see the e. distance for student-1 and student-3, there the distance is not high because marks-1 are close to each each other (viz. 280 and 270). Unlike 200 & 270 or 200 & 280.

Why scale or normalize data?

You have data of a person.

Usecase: Eligibility for a loan

Features:
- age [20-60]
- years of experience [0-40]
- salary [In lacs]

Salary would start dominating in the calculations of ML model if you don’t scale it or normalize it to the same scale of age and Years of experience.

MinMax Scaler

Why scale data?

Example:

	Marks1	Marks2	Marks3
Student 1	280	70	60
Student 2	200	60	55
Student 3	270	40	30

Euclidean distance(s1,s2) = 80.77
Distance(s1,s3) = 43.58
Distance(s2,s3) = 71.58

Formula for scaling all data into range [0, 1] => [newmin, newmax]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

For M1, S1: v’ = ((280 - 200)/(280 - 200))(1-0) + 0 = 1

MinMax Scaler

Why scale data?

Example:

	Marks1	Marks2	Marks3
Student 1	280 (1)	70 (1)	60 (1)
Student 2	200 (0)	60 (0.66)	55 (0.83)
Student 3	270 (0.875)	40 (0)	30 (0)

Euclidean distance (s1,s2) = 80.77 (1.069)
Euclidean Distance (s1,s3) = 43.58 (1.419)
Euclidean Distance (s2,s3) = 71.58 (1.89)

Normalization

Why normalize data?

Example:

	Marks1	mean	sd
Subject 1	70	60	15
Subject 2	72	68	6

Does it mean the student has done better in subject 2?

This can be found out using the Z-score normalization.

In z-score normalization (or *zero-mean normalization*), the values for an attribute, *A*, are normalized based on the mean and standard deviation of *A*. A value, *v*, of *A* is normalized to *v'* by computing:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

Normalization

Why normalize data?

Example:

	Marks1	mean	sd
Subject 1	70	60	15
Subject 2	72	68	6

z-score is 0.67 in both cases.

Both the points are 0.67 std deviation away from the mean.

Normalization

Normalize the two variables based on *z-score normalization*.

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

In z-score normalization (or *zero-mean normalization*), the values for an attribute, *A*, are normalized based on the mean and standard deviation of *A*. A value, *v*, of *A* is normalized to *v’* by computing:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

Hint:
For age:
Std. deviation = 12.84
Mean=46.44

For %fat:
Std. deviation = 8.99
Mean=28.78

Normalize the two variables based on *z-score normalization*.

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

A value, *v*, of *A* is normalized to *v'* by computing: $v' = \frac{v - \bar{A}}{\sigma_A}$

age	23	23	27	27	39	41	47	49	50
z-age	-1.83	-1.83	-1.51	-1.51	-0.58	-0.42	0.04	0.20	0.28
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
z-%fat	-2.14	-0.25	-2.33	-1.22	0.29	-0.32	-0.15	-0.18	0.27

age	52	54	54	56	57	58	58	60	61
z-age	0.43	0.59	0.59	0.74	0.82	0.90	0.90	1.06	1.13
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7
z-%fat	0.65	1.53	0.0	0.51	0.16	0.59	0.46	1.38	0.77

MinMax Scaler

Scale using min-max the following group of data:

200, 300, 400, 600, 1000

set *min* = 0 and *max* = 1

Min-max normalization maps a value, *v*, of *A* to *v'* in the range [new_min_A, new_max_A] by computing

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

$original\ data$	200	300	400	600	1000
$[0,1]\ normalized$	0	0.125	0.25	0.5	1

Normalization

Normalize using z-score the following group of data:
200, 300, 400, 600, 1000

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

The variance of N observations is:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}\left(\sum x_i\right)^2\right]$$

Square root of the variance is called **standard deviation.**

Normalization

Normalize using z-score the following group of data:
200, 300, 400, 600, 1000

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

$original\ data$	200	300	400	600	1000
z-$score$	-1.06	-0.7	-0.35	0.35	1.78

Aside: How many modes are there?
“Mode” is the value that is occurring most number of times.

Thank You!