Machine Learning (SS 23) Assignment 01: Preprocessing

Team Members:

- Likhit Jain, 3678905, M.Sc. Computer Science
- Tareq Abu El Komboz, 3405686, M.Sc. Informatik
- Serge Kotchourko, 3309449, M.Sc. Informatik

Problem 1: Labels

(a)

The euclidean distance metric $||\cdot||_2$ for points from $p, q \in \mathbb{R}$ is defined as $d(p,q) = \sqrt{(p-q)^2} = |p-q|$

$\|\|\cdot\|\|_2$	Circle $\mapsto 0$	Square $\mapsto 1$	Triangle $\mapsto 2$	Pentagon $\mapsto 3$
Circle $\mapsto 0$	0	1	2	3
Square $\mapsto 1$	1	0	1	2
Triangle $\mapsto 2$	2	1	0	1
Pentagon $\mapsto 3$	3	2	1	0

Table 1: Euclidean Distance for Classes with label $y \in \mathbb{R}$

The problem arises from, that different class-pairs have varying distances and do not encode meaningful information. For example, the euclidean distance between (circle, pentagon) is 3, while (circle, triangle) is 2. Hence the pair (circle, triangle), with this labeling, is somehow more closely related than the pair (circle, pentagon).

(b)

Let $m = 4$ and hence a labeling will be $y \in \mathbb{R}^4$. We assign the following labels to the classes defined in the exercise:

$$\text{Circle} \mapsto \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \text{Square} \mapsto \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \text{Triangle} \mapsto \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \text{Pentagon} \mapsto \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

From this, we get the following euclidean distance between between all pairs of classes (see Table 2).

(c)

The difference here, is that instead of a hard assignment of labels a soft assignment is needed, as a destination might have multiple months where travel might be recommended, i.e. a destination might be assigned labels *January*, *February* and *March*.

A possible labeling could be the following vector $y \in \mathbb{R}^{12}$, where each vector-entry corresponds to the month (i.e. y_1 is *January*, ... y_{12} is *December*) and the value of this entry is $y_i \in [0,1]$, describing the likelihood/recommendation for this destination for this month.

$\|\cdot\|_2$	Circle	Square	Triangle	Pentagon
Circle	0	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{2}$
Square	$\sqrt{2}$	0	$\sqrt{2}$	$\sqrt{2}$
Triangle	$\sqrt{2}$	$\sqrt{2}$	0	$\sqrt{2}$
Pentagon	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{2}$	0

Table 2: Euclidean Distance for Classes with label $y \in \mathbb{R}^4$

Problem 2: Dataset Preprocessing Techniques

Problem 2.1: Handling Missing Data, Scaling, and Feature Selection

(a)

As the missing data are grades, the missing data can be completed in a multitude of way, e.g. mean, median, (mode) or some kind of interpolation (e.g. quadratic), or considering other methods from the given resource[1] with forward/backward fill.

There is a lot of discussion, if and how grades are distributed, but an initial (un)educated guess would say, that they follow some normal distribution. Hence, we use the mean of the column to fill the missing data for each column respectively. Also, as the data seems to be elements of $\mathbb{N}$, i.e. no real numbers, we round to the nearest integer.

(b)

Age	Gender	Course 1 Grade	Course 2 Grade	Course 3 Grade
18	Male	85	92	78
19	Female	84	83	90
20	Male	90	88	92
18	Female	78	85	80
19	Male	82	90	85

Table 3: Students' grades dataset, fixes

(c)

Consider the following definition of min-max scaling taken from Wikipedia[2] for some feature x:

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

First, find the minimal and maximal value of the feature (e.g. for *Temperature (C)* minimum is 1 and maximum is 8). Now, for each value (cf. x) of the feature found in the "series", calculate the new value x' using the above formula, e.g. for the first value of *Temperature (C)* we have:

$$x' = \frac{x - \min x}{\max x - \min x} = \frac{5 - 1}{8 - 1} = \frac{4}{7} \approx 0.571$$

[1] https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/
[2] https://en.wikipedia.org/wiki/Feature_scaling

Note, that the denominator can be pre-computed and stored. Now, the final values will be in range $[0, 1]$

(d)

Date	Temperature (C)	Precipitation (mm)	Humidity (%)
1-Jan-21	$\frac{4}{7} \approx 0.571$	$\frac{5}{25} = 0.2$	$\frac{10}{25} = 0.4$
2-Jan-21	$\frac{7}{7} = 1$	$\frac{0}{25} = 0.0$	$\frac{0}{25} = 0.0$
3-Jan-21	$\frac{2}{7} \approx 0.286$	$\frac{15}{25} = 0.6$	$\frac{20}{25} = 0.8$
4-Jan-21	$\frac{0}{7} = 0.0$	$\frac{25}{25} = 1.0$	$\frac{25}{25} = 1.0$
5-Jan-21	$\frac{5}{7} \approx 0.714$	$\frac{10}{25} = 0.4$	$\frac{15}{25} = 0.6$

Table 4: Weather dataset, normalized using min-max scaling

(e)

Again, considering the given resource[3], using correlation as the main indicator for feature elimination, we can select any feature of the three given ones, as they all correlate pairwise. For example, if the *Precipitation (mm)* rises, the value of *Temperature (C)* falls and *Humidity (%)* grows. Hence, features *Temperature (C)* and *Humidity (%)* do not contain/add additional information.

Additionally, assuming for this question that the prediction comparison is made between one and all three features, the answer is yeas, as the data correlates and one of the features is enough to make the same prediction. In general, *Precipitation (mm)* nor any other of these three features would be enough to make a sufficient prediction, as its data resolution is not high enough (does not catch weather changes during the day)

Problem 2.2: Feature Engineering

(a)

No, there are no missing data entries.

(b)

For feature scaling, we chose min-max scaling, as in Exercise 2.1 (c) and (d).

Make	Model	Year	Engine Size	Horsepower	Mileage	Price
Ford	Fiesta	2018	$= 0.000$	$= 0.000$	≈ 0.222	12000
Toyota	Corolla	2017	≈ 0.308	≈ 0.143	≈ 0.555	15000
Honda	Civic	2018	≈ 0.192	≈ 0.332	≈ 0.333	22000
Chevrolet	Camaro	2015	$= 1.000$	$= 1.000$	$= 1.000$	18000
Ford	Mustang	2019	≈ 0.500	≈ 0.942	$= 0.000$	27000

Table 5: Car prices dataset, "Engine Size", "Horsepower", and "Mileage" features min-max scaled

[3]https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

(c)

We use the following formula to create a the new feature *Age* in the dataset:

$$Age_i = \text{Current-Year} - Year_i$$

Example for first entry (i.e. $i = 1$): $Age_i = \text{Current-Year} - Year_i = 2023 - 2018 = 5$. The table 5 is extended by this new feature:

Make	Model	Year	Engine Size	Horsepower	Mileage	Price	Age
Ford	Fiesta	2018	$= 0.000$	$= 0.000$	≈ 0.222	12000	5
Toyota	Corolla	2017	≈ 0.308	≈ 0.143	≈ 0.555	15000	6
Honda	Civic	2018	≈ 0.192	≈ 0.332	≈ 0.333	22000	5
Chevrolet	Camaro	2015	$= 1.000$	$= 1.000$	$= 1.000$	18000	8
Ford	Mustang	2019	≈ 0.500	≈ 0.942	$= 0.000$	27000	4

Table 6: Car prices dataset, min-max scaled and *Age* Feature extended