University of Stuttgart | ANALYTIC COMPUTING

Machine Learning (SS 23)

Assignment 01: Preprocessing (Solution)

Mojtaba Nayyeri
Mojtaba.Nayyeri@ipvs.uni-stuttgart.de

Akram Sadat Hosseini
Akram.Hosseini@ipvs.uni-stuttgart.de

Nadeen Fathallah
Nadeen.Fathallah@ipvs.uni-stuttgart.de

Rodrigo Lopez Portillo Alcocer
rodrigo.lopez-portillo-alcocer@ipvs.uni-stuttgart.de

Tim Schneider
timphillip.schneider@ipvs.uni-stuttgart.de

Osama Mohammed
osama.mohammed@ipvs.uni-stuttgart.de

Daniel Frank
daniel.frank@ipvs.uni-stuttgart.de

Submit your solution in ILIAS as a single PDF file.[1] Make sure to list the full names of all participants, matriculation number, study program, and B.Sc. or M.Sc on the first page. Optionally, you can *additionally* upload source files (e.g. PPTX files). If you have any questions, feel free to ask them in the exercise forum in ILIAS.

Submission is open until Monday, 01.05.2023, 12:00 noon.

[1] Your drawing software probably allows exporting as PDF. An alternative option is to use a PDF printer. If you create multiple PDF files, use a merging tool (like pdfarranger) to combine the PDFs into a single file.

1. Labels

Assume your are given a classifiation problem with classes ●, ■, ▲, and ⬟. Associated with each class is a label $y \in \mathbb{R}^m$ that needs to be defined. Throughout this semester you will learn about different machine learning algorithms for classification, i.e. find $f : x \mapsto \hat{y}$ which for each data point maps input features $x \in \mathbb{R}^n$ to a prediction $\hat{y} \in \mathbb{R}^m$ in a corresponding label space.

(a) **Task** Assume you assigned the labels ● $\mapsto 0 \in \mathbb{R}$, ■ $\mapsto 1 \in \mathbb{R}$, ▲ $\mapsto 2 \in \mathbb{R}$, and ⬟ $\mapsto 3 \in \mathbb{R}$. Compute the euclidean distance metric $\|\cdot\|_2$ between all pairs of classes. Why might this kind of labeling problematic for training an Machine Learning algortihm?

Solution:

$$\begin{aligned}
\|0-1\|_2 &= \sqrt{(-1)^2} = 1 \\
\|0-2\|_2 &= \sqrt{(-2)^2} = 2 \\
\|0-3\|_2 &= \sqrt{(-3)^2} = 3 \\
\|1-2\|_2 &= \sqrt{(-1)^2} = 1 \\
\|1-3\|_2 &= \sqrt{(-2)^2} = 2 \\
\|2-3\|_2 &= \sqrt{(-1)^2} = 1
\end{aligned}$$

The problem is that the classes have different distances to each other which implies an unintended preference to the algorithm.

(b) **Task** Come up with a different labeling $y \in \mathbb{R}^m$ such that the euclidean distance $\|\cdot\|_2$ between all pairs of classes is uniform.

Solution:

$$\begin{aligned}
\text{●} &\mapsto [1,0,0,0]^T \in \mathbb{R}^4 \\
\text{■} &\mapsto [0,1,0,0]^T \in \mathbb{R}^4 \\
\text{▲} &\mapsto [0,0,1,0]^T \in \mathbb{R}^4 \\
\text{⬟} &\mapsto [0,0,0,1]^T \in \mathbb{R}^4
\end{aligned}$$

and verify that it is indeed uniform

$$\begin{aligned}
\|[1,0,0,0]^T - [1,0,0,0]^T\|_2 &= \sqrt{0+0+0+0} = \sqrt{0} = 0 \\
\|[1,0,0,0]^T - [0,1,0,0]^T\|_2 &= \sqrt{2} \\
\|[1,0,0,0]^T - [0,0,1,0]^T\|_2 &= \sqrt{2} \\
\|[1,0,0,0]^T - [0,0,0,1]^T\|_2 &= \sqrt{2}
\end{aligned}$$

(c) **Task** Now, assume you are training a machine learning algorithm for your travel blog. Given the characteristics of a destination, your algorithm should recommend the best time to travel there from *January* to *December*. How does this differ from predicting ●, ■, ▲, and ⬟? Again, find a labeling that suits the topology of such these classes.

Solution:

$$\phi(m) = \left[\sin\frac{m\pi}{12}, \cos\frac{m\pi}{12}\right]^T \quad \forall m \in \{0,\dots,11\} \tag{1}$$

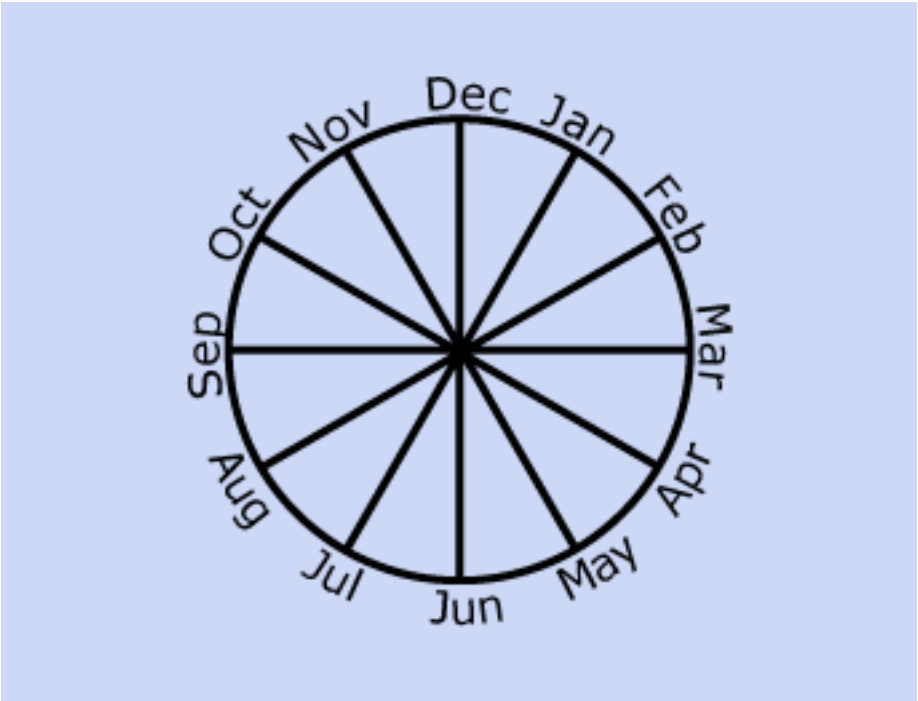

Figure 1 Visualization of the class embeddings.

2. Dataset Preprocessing Techniques

2.1. Handling Missing Data, Scaling, and Feature Selection

In this exercise, you are going to explore different basic steps that are performed when dealing with a new dataset, namely, data preprocessing.

(a) **Task** The first thing to do when dealing with a new dataset is to handle missing data[2]. Consider the following dataset of students grades, you will notice two points of data are missing. Think about a way to complete those two points and explain briefly how to do so.

Table 1 Students' grades dataset

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

Solution: One way to complete the missing data in the dataset is to average all the existing data points in a certain feature and use the mean as the missing data point.

(b) **Task** Write down the values of the missing two data points using your answer from the previous task.

Solution:

Age	Gender	Course 1 Grade	Course 2 Grade	Course 3 Grade
18	Male	85	92	78
19	Female	**83.75**	83	90
20	Male	90	88	92
18	Female	78	85	80
19	Male	82	90	**85**

(c) **Task** Sometimes, the features of a certain dataset can have different ranges. In order to have each feature equally affecting the output, we should perform data scaling techniques such as normalization or standardization for the data[3]. One way to do this is to use min-max scaling. Consider the following dataset where we have several features describing the weather status on different days using three features: temperature, precipitation, and humidity.

Table 2 Weather dataset

Date	Temperature (C)	Precipitation (mm)	Humidity (%)
1-Jan-21	5	10	80
2-Jan-21	8	5	70
3-Jan-21	3	20	90
4-Jan-21	1	30	95
5-Jan-21	6	15	85

[2] https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/
[3] https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html

As can be seen, each feature has a completely different range from the other. Explain briefly how would you perform min-max scaling on a single feature.

Solution: Using min-max scaling on one feature is performed by applying the following equation:

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

(d) **Task** Perform min-max scaling to have the same range for all the features. It is sufficient to only show your calculation for the Temperature feature and write down the final answers for the other two features.

Solution:

For the Temperature feature, $x_{min} = 1$, while $x_{max} = 8$. So the denominator in the previous equation is $x_{max} - x_{min} = 8 - 1 = 7$.
For the first data point, $x_{scaled} = \frac{5-1}{7} = \frac{4}{7} = 0.5714$
For the second data point, $x_{scaled} = \frac{8-1}{7} = \frac{7}{7} = 1$
For the third data point, $x_{scaled} = \frac{3-1}{7} = \frac{2}{7} = 0.2857$
For the fourth data point, $x_{scaled} = \frac{1-1}{7} = \frac{0}{7} = 0$
For the fifth data point, $x_{scaled} = \frac{6-1}{7} = \frac{5}{7} = 0.7143$
Here is the whole dataset scaled.

Date	Temperature (C)	Precipitation (mm)	Humidity (%)
1-Jan-21	0.5714	0.2	0.4
2-Jan-21	1	0	0
3-Jan-21	0.2857	0.6	0.8
4-Jan-21	0	1	1
5-Jan-21	0.7143	0.4	0.6

(e) **Task** Another important data preprocessing technique is to do feature selection. This is typically done when there are a lot of features and it is useful to select the most relevant features. Feature selection can be done using correlation analysis, or recursive feature elimination[4]. Consider the weather dataset from the previous example in Table 2. Perform feature selection by choosing only one feature from the dataset. Do you think only one feature is sufficient to predict the weather on a certain day?

Solution:

If we look at the dataset, we notice that the three features are correlated. For example, the higher the temperature, the lower the Precipitation, and so on. We can have good accuracy for predicting the weather by only choosing the temperature from this dataset.

Date	Temperature (C)
1-Jan-21	0.5714
2-Jan-21	1
3-Jan-21	0.2857
4-Jan-21	0
5-Jan-21	0.7143

2.2. Feature Engineering

There exist several other data preprocessing techniques. In this exercise, you are going to encounter one more preprocessing technique called feature engineering[5]. First, use the knowledge from previous questions in order

[4] https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/
[5] https://www.heavy.ai/technical-glossary/feature-engineering

to perform the following preprocessing techniques. You have the dataset Car Prices as follows.

Table 3 Car prices dataset

Make	Model	Year	Engine Size	Horsepower	Mileage	Price
Ford	Fiesta	2018	1.0	100	5000	12000
Toyota	Corolla	2017	1.8	132	8000	15000
Honda	Civic	2018	1.5	174	6000	22000
Chevrolet	Camaro	2015	3.6	323	12000	18000
Ford	Mustang	2019	2.3	310	3000	27000

For this dataset do the following:

(a) **Task** Are there any missing data? If so, handle these missing data and complete the dataset.

Solution: No missing data.

(b) **Task** Perform feature scaling for the "Engine Size", "Horsepower", and "Mileage" features.

Solution:

Make	Model	Year	Engine Size (scaled)	Horsepower (scaled)	Mileage (scaled)	Price
Ford	Fiesta	2018	0	0	0.222	12000
Toyota	Corolla	2017	0.308	0.144	0.556	15000
Honda	Civic	2018	0.192	0.332	0.333	22000
Chevrolet	Camaro	2015	1	1	1	18000
Ford	Mustang	2019	0.5	0.942	0	27000

(c) **Task** Perform feature engineering to create a new feature (age). In this exercise, you are going to use the existing feature(s) in order to engineer a new feature that could be more useful in a new format. The new feature can be easily computed from the Year feature in the dataset. Show an example of your calculations.

Solution:

Since we are currently in 2023, the Age feature can be calculated by subtracting the Year of the car from 2023. For example, the first car "Ford" has an Age value of $2023 - 2018 = 5$

Make	Model	Year	Engine Size (scaled)	Horsepower (scaled)	Mileage (scaled)	Price	Age
Ford	Fiesta	2018	0	0	0.222	12000	5
Toyota	Corolla	2017	0.308	0.144	0.556	15000	6
Honda	Civic	2018	0.192	0.332	0.333	22000	5
Chevrolet	Camaro	2015	1	1	1	18000	8
Ford	Mustang	2019	0.5	0.942	0	27000	4