University of Stuttgart | ANALYTIC COMPUTING

Machine Learning (SS 24)

Assignment 01: Preprocessing and K-Nearest Neighbors

Yi Wang
yi.wang@ki.uni-stuttgart.de

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

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

Farane Jalali-Farahani
farane.jalali-farahani@ki.uni-stuttgart.de

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

Daniel Frank
daniel.frank@ki.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, 22.04.24, 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. Preprocessing (50 point)

Imagine we have a small dataset from a survey about personal transportation preferences, with the following data:

ID	Age	Income (K$)	Owns Car	Number of Vehicles	Preferred Transport Mode
1	25	50	Yes	2	Car
2	NaN	40	No	0	Public Transport
3	35	NaN	Yes	1	Car
4	45	70	Yes	NaN	Car
5	30	60	No	0	Bike

Table 1 Sample Table

(a) **Handling Missing Values (15 point).** Given the dataset above, identify the columns with missing values and apply the following imputation techniques:

- For the "Age" column, replace missing values with the mean age.
- For the "Income (K$)" column, replace missing values with the median income.
- For the "Number of Vehicles" column, replace missing values with the mode of the column.

Calculate the new values that will replace the NaNs in the dataset.

(b) **Feature Scaling (20 point).** Apply Min-Max scaling to the "Income (K$)" column of the updated dataset from Question (a). Calculate the scaled values for the "Income (K$)" column.

(c) **Encoding Categorical Data (15 point).** Encode the "Owns Car" and "Preferred Transport Mode" columns using the following encoding techniques:

- For "Owns Car", use binary encoding where "Yes" is 1 and "No" is 0.
- For "Preferred Transport Mode", apply one-hot encoding.

List the resulting columns and their corresponding encoded values.

2. K-Nearest Neighbors (50 point)

Consider a small dataset with three features (X1, X2, X3) and a binary class label (Y). Let's assume we're working on a classification task with two classes: 0 and 1.

Observation	X1	X2	X3	Y
1	2	3	1	0
2	5	4	2	0
3	1	2	3	1
4	3	1	2	1
5	4	3	3	1

(a) **Distance Calculation (20 point).** Suppose you have a new observation with features (X1=3, X2=3, X3=2), and you want to classify this observation using the KNN algorithm with K=3. Calculate the Euclidean distance between the new observation and each observation in the dataset. Which are the three nearest neighbors, and what class would the new observation be assigned?

(b) **Impact of K (20 point).** Using the same new observation (X1=3, X2=3, X3=2), how would the assigned class change if K=1 and K=5? Explain the potential benefits and drawbacks of using a smaller vs. larger K value in KNN.

(c) **Distance Weighting (10 point).** Describe how the classification of the new observation (X1=3, X2=3, X3=2) might change if distance-weighted voting is applied instead of uniform voting. Assume K=3 for your calculation.