University of Stuttgart | ANALYTIC COMPUTING

Machine Learning (SS 24)

Assignment 01: Preprocessing and K-Nearest Neighbors (Solution)

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.

Solution:

- **Age**: The mean age is calculated to be 33.75 years. The missing value in the "Age" column will be replaced with this mean.
- **Income (K$)**: The median income is $55K$\$[2]. The missing value in the "Income (K$)" column will be replaced with this median.
- **Number of Vehicles**: The mode of the number of vehicles is 0. The missing value in the "Number of Vehicles" column will be replaced with this mode.

After replacing missing values, the dataset is updated as follows:

ID	Age	Income (K$)	Owns Car	Number of Vehicles	Preferred Transport Mode	Income (K$) Scaled
1	25	50	Yes	2	Car	0.333333
2	33.75	40	No	0	Public Transport	0.000000
3	35	55	Yes	1	Car	0.500000
4	45	70	Yes	0	Car	1.000000
5	30	60	No	0	Bike	0.666667

(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.

Solution:

$$\text{Scaled Value} = \frac{\text{Value} - \text{Min}}{\text{Max} - \text{Min}}$$

[2] K$ means thousands of dollars.

Using Min-Max scaling on the "Income (K$)" column with the minimum income of 40K$ and the maximum of 70K$, the scaled values are calculated as shown in the "Income (K$) Scaled" column of the updated dataset above.

(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.

Solution:

Binary Encoding for "Owns Car"

- Yes $\rightarrow$ 1
- No $\rightarrow$ 0

One-Hot Encoding for "Preferred Transport Mode"

- "Car" $\rightarrow$ (1, 0, 0)
- "Public Transport" $\rightarrow$ (0, 1, 0)
- "Bike" $\rightarrow$ (0, 0, 1)

The dataset would be extended with additional columns to accommodate the one-hot encoded "Preferred Transport Mode", resulting in a structure like this:

Owns Car (Encoded)	Car	Public Transport	Bike
1	1	0	0
0	0	1	0
1	1	0	0
1	1	0	0
0	0	0	1

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?

Solution:

1. Calculate the Euclidean distance using the formula: $\sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + (x_{3i} - x_{3j})^2}$
- Distance to Observation 1: 1.414
- Distance to Observation 2: 2.236
- Distance to Observation 3: 2.449
- Distance to Observation 4: 2.000
- Distance to Observation 5: 1.414

2. Identify the three nearest neighbors based on the smallest distances.
- The three nearest neighbors are Observations 1, 4, and 5, with distances of 1.414, 2.000, and 1.414, respectively.

3. Determine the majority class among these neighbors.
- Observation 1's class: 0
- Observation 4's class: 1
- Observation 5's class: 1
- The majority class among the nearest neighbors is Class 1.

Therefore, if we classify the new observation using the KNN algorithm with K=3, it would be assigned to Class 1.

(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.

Solution:

For this question, we analyze how the classification of the new observation (X1=3, X2=3, X3=2) changes with K=1 and K=5, and discuss the implications of choosing different K values.

When K=1

1. Find the nearest single neighbor: From the previous calculations, we know that the nearest neighbors (closest distances) are Observations 1 and 5, both with distances of 1.414. For K=1, we can choose either of them (let's use Observation 1 for this example).
2. Class of the nearest neighbor: Observation 1 belongs to Class 0.
Classification with K=1: The new observation would be classified as Class 0.

When K=5

1. Find the five nearest neighbors: All observations in our dataset will be considered, as we only have five observations.
2. Determine the majority class among these neighbors: We have two observations belonging to Class 0 (Observations 1 and 2) and three belonging to Class 1 (Observations 3, 4, and 5)
Classification with K=5: The new observation would be classified as Class 1.

Benefits and Drawbacks:

- Smaller K (K=1): More sensitive to noise in the dataset; the classification is based on the nearest observation only, which can lead to overfitting.
- Larger K (K=5): Tends to smooth out the classification and reduce the effect of noise. However, it may include points that are farther away, which could lead to underfitting or misclassifying the new observation if the boundary between classes is complex.

(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.

Solution:

With distance-weighted voting, the influence of each neighbor on the classification decision is weighted by the inverse of their distance to the point being classified, giving closer neighbors more influence.

1. Calculate weighted influence for the three nearest neighbors:
- Weight for Observation 1: 0.707
- Weight for Observation 4: 0.5
- Weight for Observation 5: 0.707

2. Calculate weighted votes for each class:
- Total weighted vote for Class 0: 0.707 (from Observation 1)
- Total weighted vote for Class 1: 1.207 (from Observations 4 and 5)

The new observation would be classified as Class 1, because the total weighted vote for Class 1 is higher than that for Class 0.