Homework 6

45 points
Due Wednesday, April 10 at 11:59 PM

For this assignment, you will investigate patterns of bicycle rentals using K-means clustering. You will use an adaptation of the Bike Rental Data Set from Kaggle[1]. **Please use this version from the course dataset repository**: `https://raw.githubusercontent.com/CUNY-CISC-3225/datasets/main/bike_rentals.csv`. This version has been edited to remove unnecessary columns and has altered some values to make preprocessing easier.

The dataset contains data from 2 years of bicycle rental data. Each row represents an hour of a day, and contains several attributes about the rental time including the season, weather, and what kind of day it is (working day and holiday). It also contains a "count" column, which indicates the number of bikes rented over the hour.

Note that the "temp" column contains the actual temperature in Celsius, and "atemp" contains an estimate of the sensation of temperature in Celsius.

1 Analysis (10 points)

Perform a basic exploratory analysis of the dataset. The goal of this analysis should be to identify whether each column is continuous or categorical.

1. Are there any missing values?

2. Regardless of your answer above, would missing values be a problem for clustering? Why or why not?

3. Which columns contain continuous values? For these columns, use `describe()` to show basic summary statistics about them.

4. Which columns contain categorical values? Use `value_counts()` to show the unique values contained in each column, and how often they appear.

5. Of the categorical columns, which contain binary yes/no values? What do the yes/no values mean?

2 Preprocessing (6 points)

Perform any steps necessary to preprocess the dataset to prepare it for clustering. This could include one-hot encoding, producing binary 0/1 values, or standardizing with Z-score normalization.

3 K=2 (8 points)

Perform K-means clustering with K=2 (3 points). Answer the following questions:

1. (1 point) How many different rental periods are represented in each cluster?

2. (4 points) Using averaged columns within each cluster, give a profile of a low-count rental period and a high-count rental period. Why do you think fewer people are renting bikes in the low-count rental period?

[1] https://www.kaggle.com/datasets/aguado/bike-rental-data-set-uci

4 K=3 (10 points)

Perform K-means clustering with K=3 (3 points). Answer the following questions:

1. (1 point) How many different rental periods are represented in each cluster?
2. (4 points) Using averaged columns within each cluster, give a profile of a low-count rental period and a high-count rental period. Why do you think fewer people are renting bikes in the low-count rental period?
3. (2 points) Based on what you've seen so far, which value of K (K=2 or K=3) provides more useful insight into bike rental patterns? Why?

5 Elbow Method (11 points)

Use the elbow method with inertia scores to approximate an ideal value of K. Once you have done this, perform a K-means clustering with this value of K you discovered (6 points). Answer the following questions:

1. (1 point) How many different rental periods are represented in each cluster?
2. (4 points) Using averaged columns within each cluster, give a profile of a low-count rental period and a high-count rental period. Why do you think fewer people are renting bikes in the low-count rental period?

Submission Instructions

In Blackboard, submit written responses in an appropriate text format (PDF, Word, LibreOffice, etc) and your *.ipynb file(s). Do not submit a share link.