Bike Sharing Demand

YAO YANG

Chongqing University of Posts and Telecommunications test

July 14, 2023

Overview

Problem

Defn The goal of this project is to forecast bike rental demand given the input feature like the duration of travel, departure location, arrival location, and time elapsed.

Defn Evaluation metrics: RMSLE(Root Mean Squard Logarithmic Error) is required to evaluate the model.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[log(p_i+1)-log(a_i+1)]^2}$$

n is the number of test set samples, pi is the test value, and ai is the actual value. When the root mean square error is smaller, it means that the fitting effect of the data is better and the test value is closer to the actual value.

Data clean

Defn

You are provided hourly rental data spanning two years. For this competition, the training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour111covered by the test set, using only information available prior to the rental period.

- train.csv It contains a training set of target variables.
- test.csv It does not contain a training set of target variables.
- sampleSubmission.csv It is a properly formatted sample submission file.

Date fields

- datetime - hourly date + timestamp
- season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday - whether the day is considered a holiday
- workingday - whether the day is neither a weekend nor holiday
- weather - 1: Clear, Few clouds, Partly cloudy, Partly cloudy

  2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

  3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

  4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp - temperature in Celsius
- atemp - "feels like" temperature in Celsius
- humidity - relative humidity
- windspeed - wind speed
- casual - number of non-registered user rentals initiated
- registered - number of registered user rentals initiated
- count - number of total rentals

Implementation process

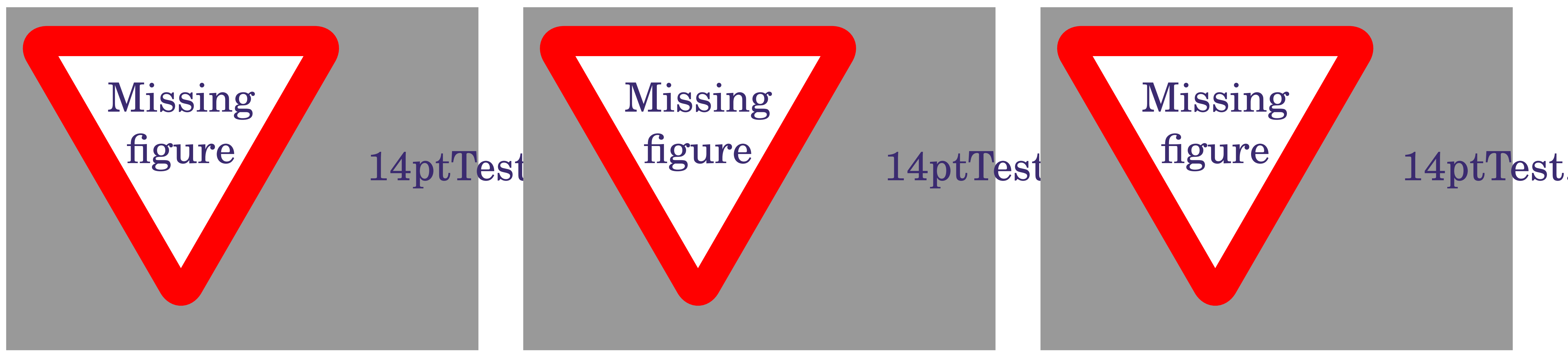
Step One - Data preprocessing

- Suppose f_1, f_2, f_3 are three features of G_q.

  f_1: $\{x_1, x_2, x_3, x_4, x_5, x_2, x_3, x_4, x_1, x_2\}$

  f_2: $\{y_2, y_2, y_1, y_2, y_3, y_3, y_5, y_4, y_4, y_2\}$

  f_3: $\{z_1, z_4, z_2, z_4, z_5, z_3, z_1, z_2, z_4, z_2\}$



(a) f_1 (b) f_2 (c) f_3

Figure 1: Histogram of G_q on three features

Step Two - Feature engineering

- Calculate Earth Mover Distance
  - Represent one feature among different groups
  - Purpose: calculate the minimum mean distance

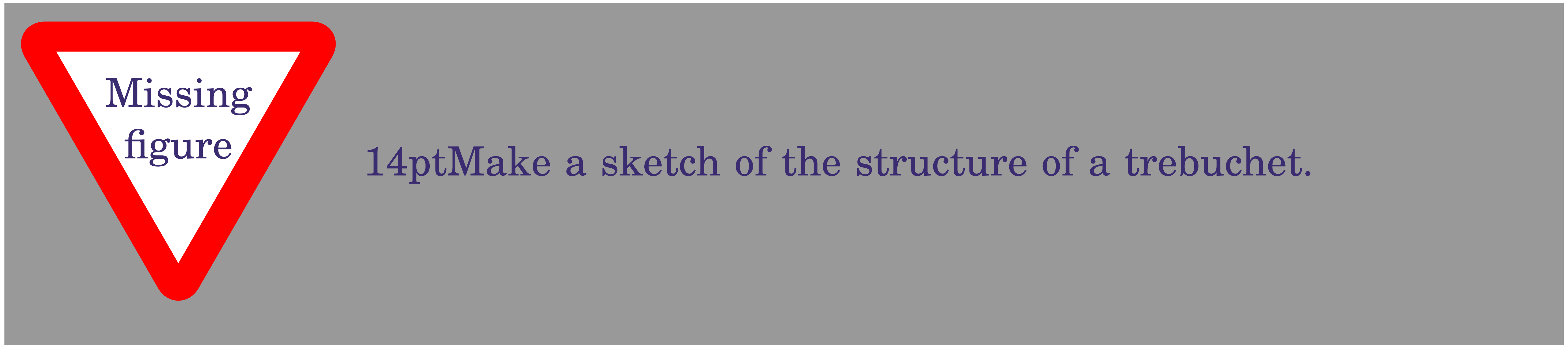


Figure 2: EMD of one feature

Step Three - Buliding models to make predictions

- Calculate the outlying degree

$$OD(G_q) = \sum_1^n EDM(h_{q_s}, h_{k_s})$$

- n ⇔ the number of contrast groups.
- h_{k_s} ⇔ the histogram representation of G_k in the subspace s.

- Identify group outlying aspects mining based on the value of outlying degree.
- The greater the outlying degree is, the more likely it is group outlying aspect.

Prediction results

Evaluation

- $Accuracy = \frac{P}{T}$

  P: Identified outlying aspects

  T: Real outlying aspects

- Synthetic Dataset and Ground Truth

Table 1: Synthetic Dataset and Ground Truth

Query group	$\mathbf{F_1}$	$\mathbf{F_2}$	F_3	$\mathbf{F_4}$	F_5	F_6	F_7	F_8
i_1	**10**	**8**	9	**7**	7	6	6	8
i_2	**9**	**9**	7	**8**	9	9	8	9
i_3	**8**	**10**	8	**9**	6	8	7	8
i_4	**8**	**8**	6	**7**	8	8	6	7
i_5	**9**	**9**	9	**7**	7	7	8	8
i_6	**8**	**10**	8	**8**	6	6	8	7
i_7	**9**	**9**	7	**9**	8	8	8	7
i_8	**10**	**9**	10	**7**	7	7	7	7
i_9	**9**	**10**	8	**8**	7	6	7	7
i_{10}	**9**	**9**	7	**7**	7	8	8	8