Group 1 Presentation

---Zekun Wang, Jinwen Xu,Yuman Wu, Yidan Huo, Yicheng Du

Background

Data Source: We use the dataset named ***Amazon US Customer Reviews Dataset*** from Kaggle.

Dataset Introduction: The dataset encompasses a wide range of customer opinions and experiences, which has a total size of 54.41GB, providing detailed insights into user perspectives on various products listed on Amazon.

Our main questions:

1. How can we know ***the 5 fields with the best development*** and ***the 5 fields with the worst development*** in recent years?

2. How can we ***create new variables*** that ***capture the full information of a time series*** for further effective analysis?

A detailed description of the dataset:

This dataset contains ***37 distinct folders***, and each folder includes ***15 variables***. For our analysis, ***we focused on two specific variables***: ***'star_rating'*** and **'review_date'**.

The specific variable explanations are as follows:

(1) "***Star_rating***" is the 1-5 star rating of the review. Its data type is ***int***.

(2) "***Review_date***" is the date when the review is published. Its data type is ***datetime***.

Task 1

Question: Calculate the mean star-rating of all the 37 categories of products on Amazon.

Method: Calculate the means through parallel computing.

Results:

Top 5 mean star-ratings:

Gift card	**4.731372**
Digital Music Purchase	**4.642895**
Music	**4.436622**
Grocery	**4.312221**
Multilingual	**4.306758**

Bottom 5 mean star-ratings:

Digital Software	**3.539226**
Software	**3.566997**
Major Appliances	**3.716390**
Mobile Electronics	**3.763225**
Digital Video Games	**3.852957**

Task 2

Question: How can we create new variables that capture the full information of a time series for further effective analysis?

Method: Calculate the auto-covariance function matrix of the time series

Computational steps:

(1) Suppose there is a time series, $X_1, X_2, ..., X_t, ..., X_n$, then calculate the mean of the sequence.

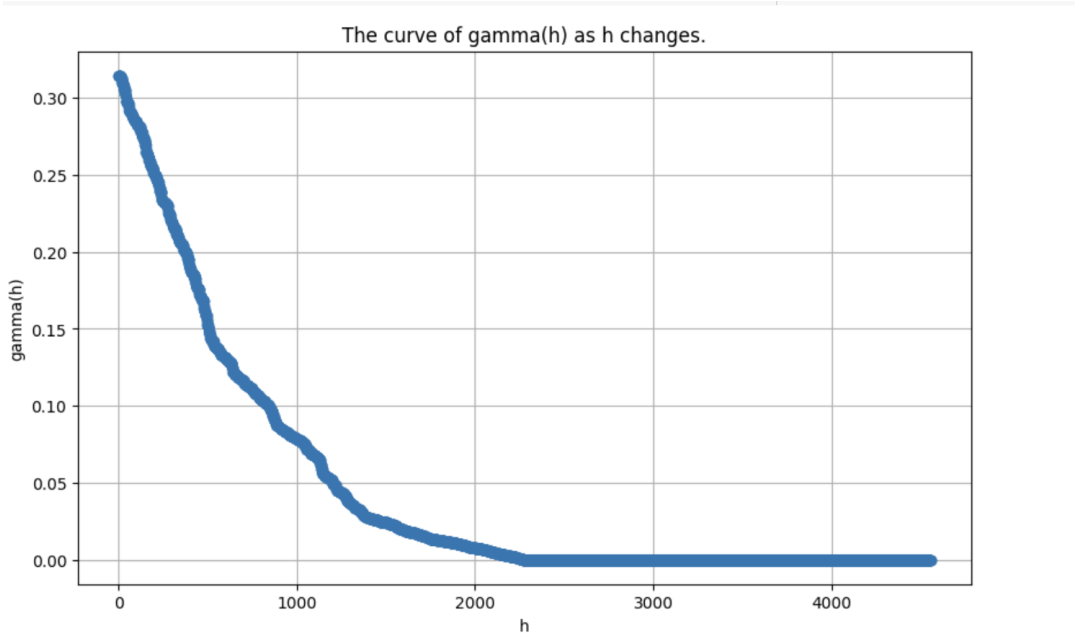
(2) For each possible lag value h, compute the sample covariance between the lagged sequence and the original sequence.

$$\gamma(h) = n^{-1} \sum_{j=1}^{n-h} (x_{j+h} - \bar{x})(x_j - \bar{x}), 0 \leq h < n$$

(3) Auto-covariance function matrix: $\Gamma(m) = [\gamma(0), \gamma(1) \ldots, \gamma(m)]$

Partial calculation results:

Taking 'Musical_Instruments' as an example, select to calculate the average daily score of different products, with a data length of 4558. To visually demonstrate, we create a curve of $\gamma(h)$ as h changes.

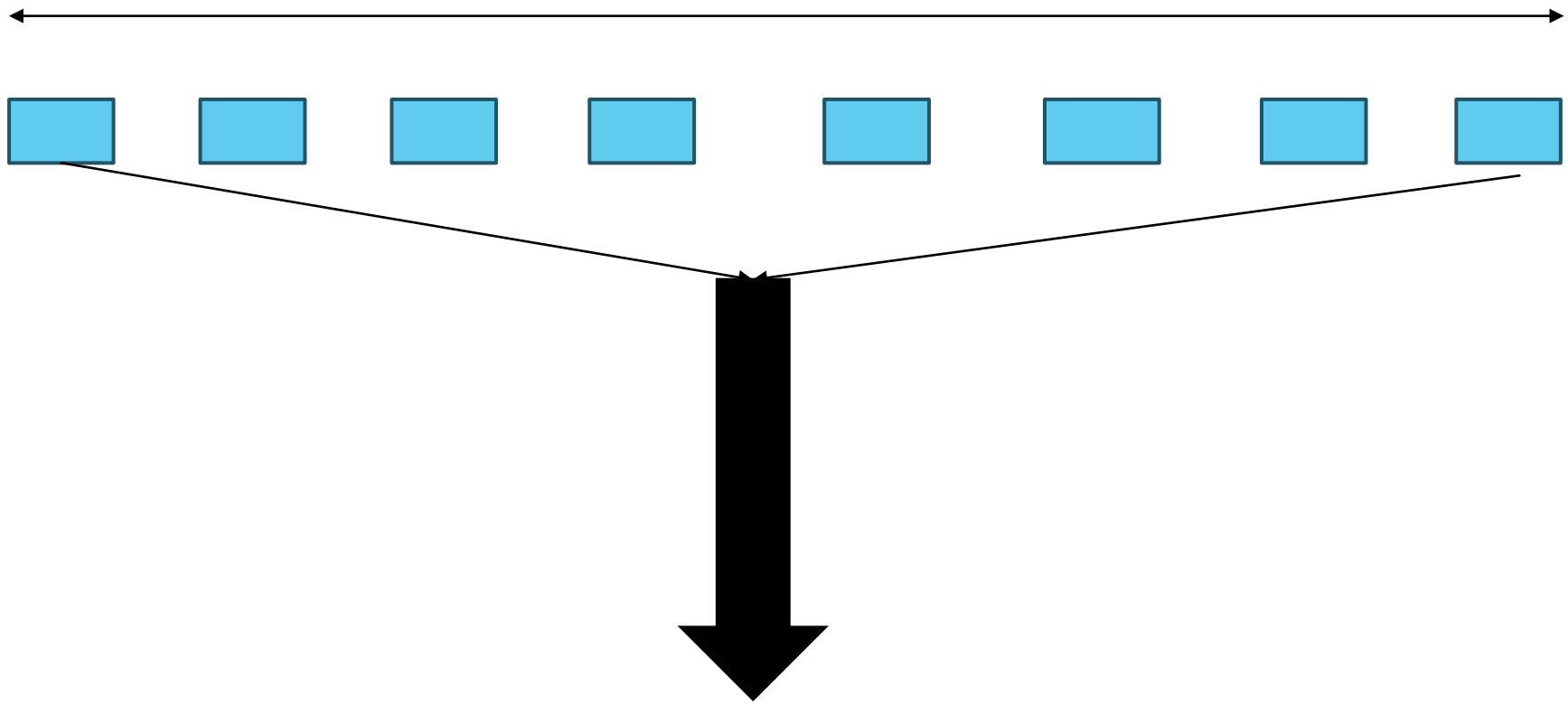


Application:

1. In machine learning: we use the auto-covariance function matrix to represent the features of the time series, which can be used for classification and clustering. In this case, dimensionality reduction is usually used in combination.

2. In the field of deep learning: using the self-covariance function matrix to optimize the transformer-based framework, reducing model complexity and enhancing interpretability.

37 categories files in total, each can be as large as 10 GBs.



Then we begin to conduct our data processing jobs to deal with these files.

Job 1 (Calculating the mean)

It consists of 4 files:

1. Pre1.sh: It splits the file into 100 subfiles and cuts the column “star-rating” of interest.
2. Job1.sh
3. Job1.sub

These two files run all the 100 parallel jobs to calculate the total amount of “star-rating”.

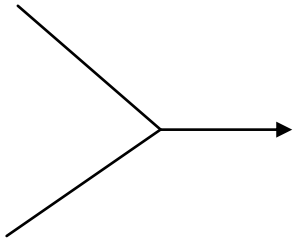
4. Post1.sh

It calculates the sum of the sub-total of the 100 files and total number of reviews , and get the average value of “star-rating”.

Job 2 (Calculating the auto-correlation)

It consists of 3 files:

1. Job2.R: It makes use of the average “star-rating” in Job 1 to calculate the auto-correlation coefficients of the time series.
2. Job2.sh
3. Job2.sub



These two files send jobs to calculate, on which we could possibly make further improvements.

Thank you!