Weicong Feng

Professor Chia-Ling Tsai

CSC74020 Machine Learning

09/18/2021

Assignment1

**Report**

This assignment aims to familiar with coding with matrix operations instead of nested loop and compare the performances between them.

1. Problem1: compare the performances of computing the Euclidean distance by using nested loop and matrix operations
   1. Definition: Euclidean distance is a measure of the true straight line distance between two points in Euclidean space ([Stanford](https://hlab.stanford.edu/brian/euclidean_distance_in.html).) It is broadly used in image processing, physics and machine learning (detail topics in [ScienceDirect](https://www.sciencedirect.com/topics/mathematics/euclidean-distance).)

Diagram, schematic

Description automatically generated (1.1)

In those,

 (1.2)

or . (1.3)

* 1. Solution
     1. Nested loop

For computing the distance between two points, the function accepts two vectors as input and output the distance. In practice, it is necessary to calculate two batches of vectors simultaneously. So, the function should accept two matrices as input, in which each row corresponds to a vector. The output is a matrix whose dimension is the first dimension of the first inputted matrix x the first dimension of the second inputted matrix.

Hence, the first step is checking if the two inputted matrices have same second dimension because it is impossible to calculate the distance between two vectors in various dimension spaces.

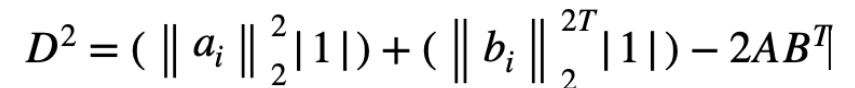
Then, initiate the result matrix with zero by using the first dimensions of the inputted matrices. Finally, two nested loops are used to traverse the Euclidean distance matrix. Inside the loops, the 1.3 is used to fill the matrix.

* + 1. Matrix operation

From a matrix perspective, the Euclidean Distance Matrix can be represented as:

A picture containing text

Description automatically generated (1.4)

or  (1.5)

So, if A and B are two matrices with same 2nd dimension, Euclidean Distance Matrix

Z = A2+B2-2ABT  (1.6)

* 1. Experiment

Various number of rows (10 to 501 in step of 50) of matrices with random integer number are produced and feed into two functions (loop and matrix). The number of rows of inputted matrices decides the number of columns of output matrix, that means the computation. In order to obtain reasonable result, same inputs are feed into two functions for 10 times to get the mean and standard deviation of running time.

As shown from Figure-1, as the number of columns in Euclidean Distance Matrix increases, the running time of nested loop function rise rapidly, and higher variation, while that of matrix operation function rise is almost negligible, and no variation can be found.

Chart, line chart

Description automatically generated

Figure-1

1. Problem2: compare the performances of computing the correlation matrix by using nested loop and matrix operations
   1. Definition: The correlation matrix is a (K × K) square and symmetrical matrix whose ij entry is the correlation between the columns i and j of X ([ScienceDirect](https://www.sciencedirect.com/topics/mathematics/correlation-matrix).)

Chart, scatter chart

Description automatically generated (2.1)

Diagram

Description automatically generated with low confidence

(2.2)

* 1. Solution

In the functions that calculate correlation matrix, only one parameter as input. The input is a matrix need to be return corresponding correlation matrix which represents the correlation coefficient between each pair vectors. In practice, the inputted matrix typically includes m instances (row), in which there are n-dimension features (column).

* + 1. Nested loop

In order to obtain the correlation matrix R, the covariance matrix must be calculated first, which is the sum of the product of two difference between variables and column mean. It means it needs to calculate:

column mean 🡺 the difference between variables and column mean and 🡺 the product of them 🡺 sum all of the products

Due to the same column mean being used during each Sij calculation, the column means were be calculated and stored in a 1xj matrix in advance to avoid from repeating calculation.

When calculating the covariation matrix S, two nested loops are used to traverse the matrix, and each cell can be obtained by above Sij formular 2.2 (additional loop is needed to reckon Sij). At the same time, the standard deviation matrix (1xj) can be calculated in the same nested loop.

Finally, two nested loops can be used to figure out the correlation matrix R by formular 2.1. Theoretically, the correlation matrix R computation can be done in above nested loop, in which the covariation matrix and standard deviation matrix are computed. Considering making the logic clear, the final computation was separated out from covariation matrix computation loops, sacrificing a bit performance.

* + 1. Matrix operation

With matrix operations, mathematics underlying is complex, however, the calculation is simple.

First, the difference between variables and the corresponding column is easy to compute by X-u because of the broadcasting in matrix.

Second, the covariance matrix can be obtained by . The square root of the diagonal of the covariance matrix can get the standard deviation matrix.

Finally, the correlation matrix can be obtained by .

* 1. Experiments

The experiment is similar as that of Euclidean distance. The difference is the control variable is the number of rows of the inputted matrix instead of columns of the result matrix. The fact that influences the shape of the correlation matrix is the second dimension of the inputted matrix, rather than the first dimension. Although the first dimension of the inputted matrix doesn’t affect the size of shape of output matrix, it seriously impacts on the computation performance.

From Figure-2, it shows that the running time in nested loop linearly rise as the number of rows of the inputted matrix increase, while that of matrix operation totally lies on the ground.

Chart, line chart

Description automatically generated

Figure-2

1. Problem 3: Compute the pairwise distance matrix and correlation matrix for the three datasets from sklearn.datasets: Iris, Breast cancer, Digits.

Applying above functions on three different real data sets, the similar results are shown as below:

Table

Description automatically generated

Table-1

Obviously, the matrix operation is significantly efficient in time than nested loop. With the increasing of the amount of input data, the difference between the performance of these two algorithms is further enlarged.

The conclusion is try to use matrix operation instead of nested loop for performance.

Ob