National Taiwan University
Institute of Industrial Engineering
Jakey BLUE

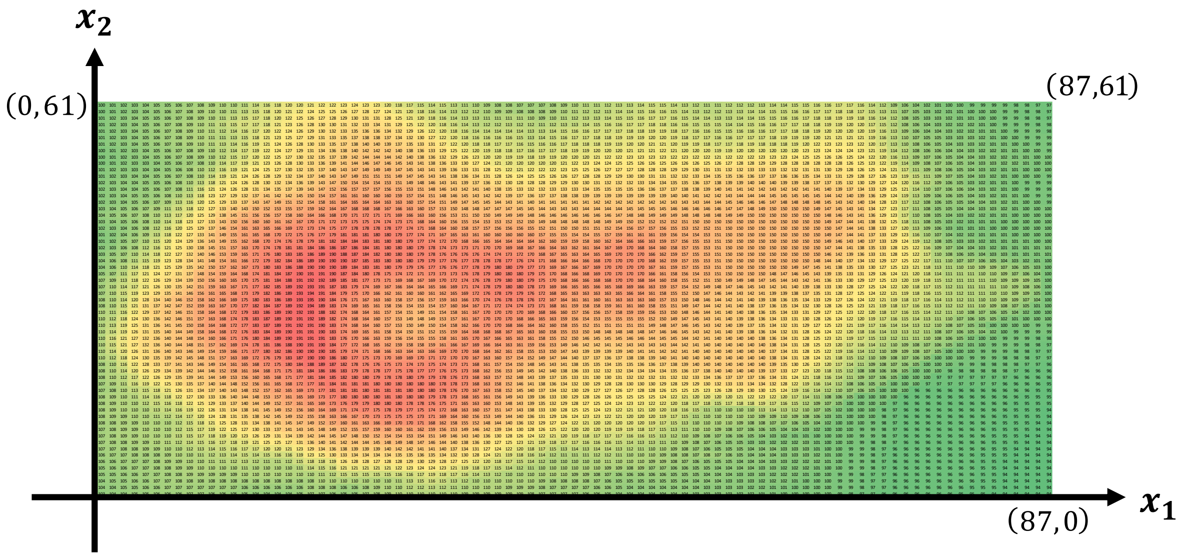
Data Analytics

110-2 Homework #02

Due at 23h59, March 6, 2022; files uploaded to NTU-COOL

1. Download the ORL faces dataset. There are 400 faces, of the dimensions 46 × 56, from 40 people.
   a. (15%) Read the 400 images into one data matrix 400 × 2576. Create an additional label column indicating the physical gender, e.g., $\{0 = \text{female}, 1 = \text{male}\}$.
   b. (10%) Regress the gender label on all the 2576 pixels? What do you observe?
   (15%) Perform the stepwise regression from a null model to find the important pixels. Plot the chosen pixels on a 46 × 56 canvas.

2. (20%) The volcano dataset can be visualized as the contour below.



We can simply assign the grid coordinates as: $x_1 = \{1, 2, \dots, 87\}$; $x_2 = \{1, 2, \dots, 61\}$.

Design an iterative algorithm based on repeating "multiple regressions" to arrive the highest point of this volcano, given the starting point at $(87, 1)$, i.e., the right-bottom corner.

Hint: you can consider a smaller domain of x_1 and x_2 to build a regression model, such that the linear hyperplane can show you an improving direction to another domain.

3. Simulate a "multiple" regression with two predictors problem by yourself, with sufficient sample size, e.g., 50000 samples.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \epsilon \sim N(0, \sigma^2).$$

   a. (5%) Use the regression package in your preferable environment (R or Python) to analyze the problem and review the results.
   b. (15%) From the perspective of "machine learning," code the gradient descent method to optimize (minimize) the error function and get the coefficients. Do you get the same results as those in (a)? Demonstrate the evolution of the iterative errors and the searching path in the domain of the error function.

*Make any necessary assumptions by yourself if not mentioned above.