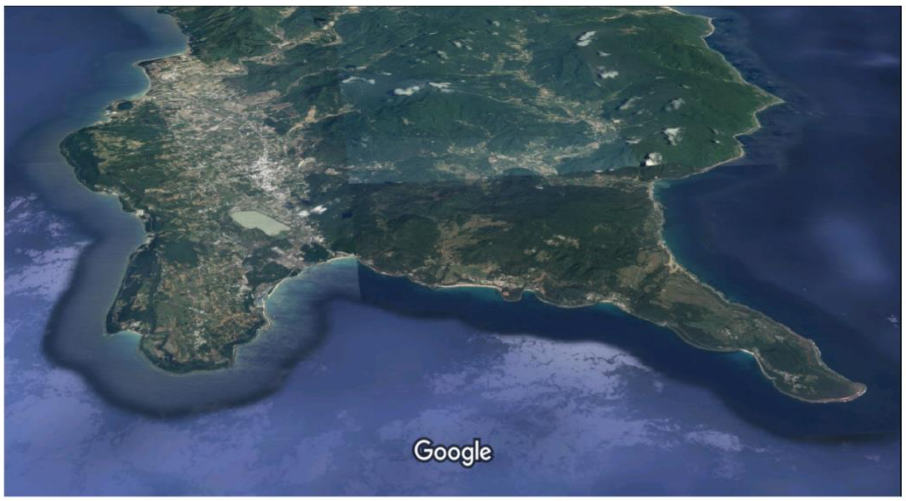
Machine Learning 2017

HW-1：Linear Regression

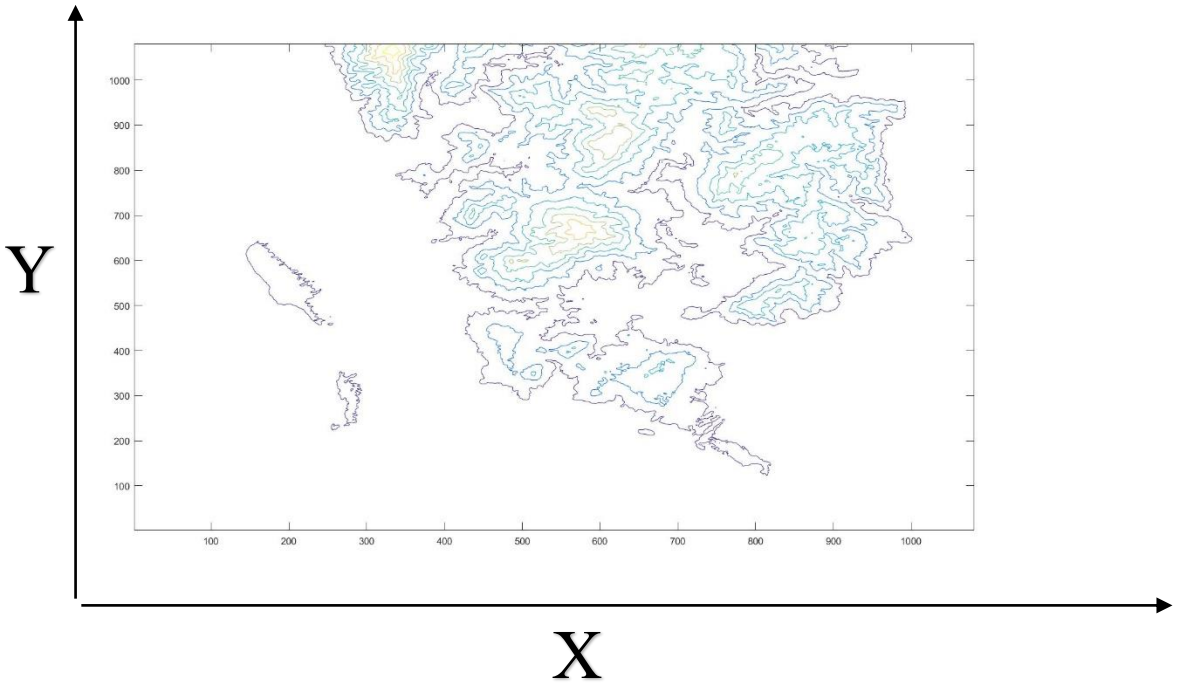
Deadlines: **2017.03.21**-23:59:59

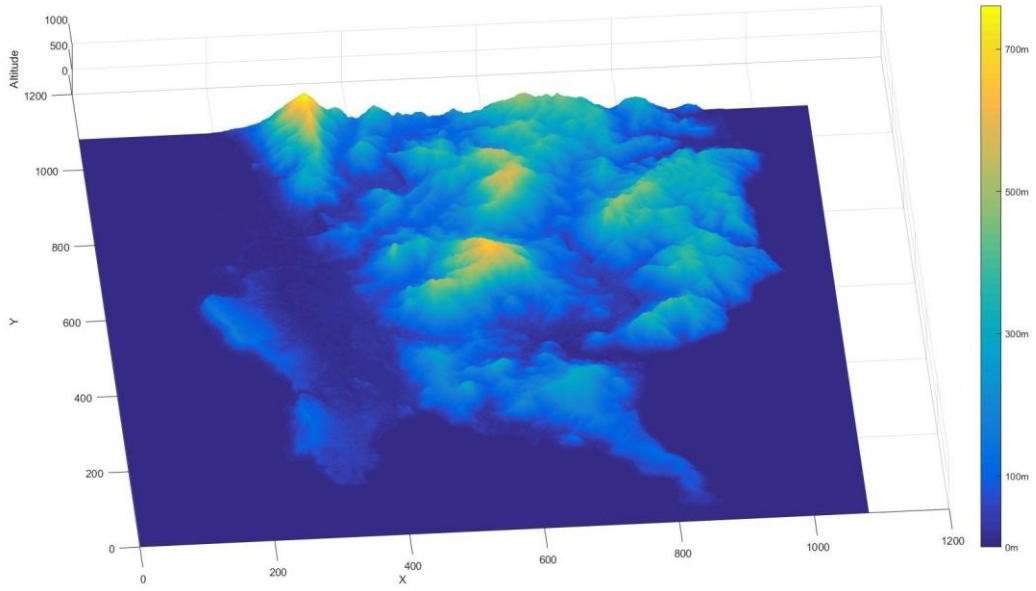
In this homework, you are asked to rebuild the **Height** map of the southern Taiwan based on the training data. The following three approaches need to be realized respectively:

- **Maximum likelihood approach (ML)**
- **Maximum a posteriori approach (MAP)**
- **Bayesian approach**



(The following results are built from full ground truth data, so your result may not be as great as they are)





◆ Target

The range of X and Y of this area will be 1 to 1081; each pair of the coordinate will have a corresponding height value. The height value have minimum value 0m. You will get a dataset of the sampling data from the original data. Please use these data to do the linear regression so the height of the unknown coordinates, which are not in the dataset you got, in this area can be known.

Training data

- **X_train** is a 40000x2 matrix, whose first column and second column record the X coordinate and Y coordinate of the 40000 training points, respectively.
- **T_train is** a 40000x1 vector, which records the heights (target values) of the training points.

Demo

- You will get a testing set named "**X_test**", please predict the corresponding height based on the coordinates in the testing set and submit the result of your 3 predictors (ML, MAP, Bayesian) with your source code and report, we will use the result you submit to evaluate your predictors.
- Don't use the testing set in your model designing.
- There will be totally 10,000 coordinates in the set.
- All coordinates will be in integer, range from 1 ~ 1081
- You may need to explain how your program works

◆ Model

In this homework, your model should be implemented by the **feature vector**, which is defined as

$$\boldsymbol{\phi}(\mathrm{x}) = [\phi_1(x), \phi_2(x), \ldots, \phi_N(x), \phi_{bias}(x)\ ]$$

Using the linear combination of feature functions of the feature vector to predict the data and minimizing the MSE(mean square error)function, y is your predicted value and t is the ground truth

$$E(\boldsymbol{w}) = \frac{1}{2K}\sum_{k}^{K}\|y(\boldsymbol{x}(k), \boldsymbol{w}) - t(k)\|^2$$

For example, we uniformly place N Gaussian basis functions over the spatial domain, with $N = O_1 \times O_2$. Here, O_1 and O_2 denote the number of locations along the X and Y directions, respectively.

For $1 \le n \le N$, we define the Gaussian basis functions as

$$\phi_n(x) = \exp\left(-\frac{(x_1 - \mu_i)^2}{2s_{1n}^2} - \frac{(x_2 - \mu_j)^2}{2s_{2n}^2}\right), for\ 1 \le i \le O_1, 1 \le j \le O_2$$

where $$n = O_2 \times (i-1) + j.$$

The example is only for reference, which means your model doesn't have to be the same! Please design the appropriate feature vector to predict the height as precise as possible.

◆ Tasks (All the tasks below should be included in your report)

1. ML approach/MAP approach/Bayesian approach

A. Use ML,MAP,and Bayesian to construct 3 predictors to predict the heights, explain clearly why and how do you design your predictors.

B. You may need to use some part of the training data as your testing data, and use those testing data to evaluate your methods.

C. Compare the difference and performance among the 3 methods.

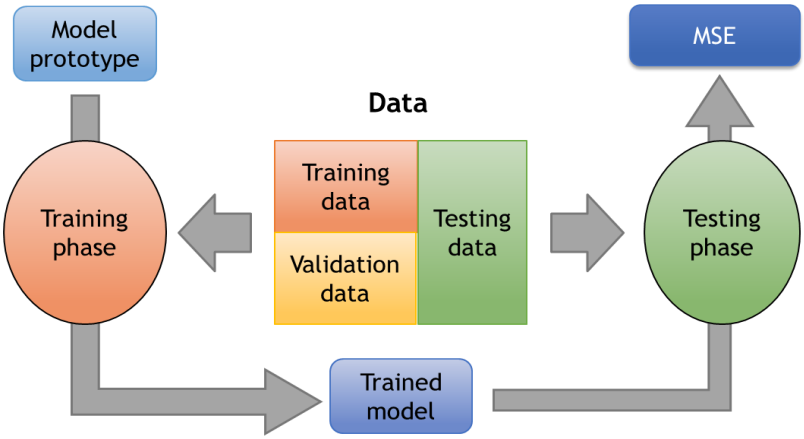
D. Try to make your predictor be underfitting and overfitting for at least one methods and do some discussion

E. Visualize your height map will help you on the discussion

F. Other discussion may help you to get higher score in this Homework

2. Cross validation

Apply N-fold cross-validation in your training stage to select at least one hyperparameter for at least one methods and do some discussion.



Requirements for demonstration:

- ☐ Predict the height according to the X_test.csv by 3 methods, direction of the coordinate are just like that in the 1st page. Save the 3 result in ML.csv/ MAP.csv/ Bayesian.csv file, respectively.
- ☐ Arrange your result into 1 columns, please follow the order of X_test.csv.
- ☐ Explanation of your methods.
- ☐ Five minutes for each person.

Hints:

- ☐ You don't really have to use all the data on training to do the discussion, since it may need to consume lots of computation resource.
- ☐ Study clearly the characteristic of the features you choose before using it.
- ☐ You can modify the data before inputting it if needed, just make sure each coordinate correspond to the correct height.

Reminders:

- ☐ **Your report should be within 12 pages**
- ☐ **Using python is encouraged for you, especially for the machine learning area(see here)**

- ☐ **Don't use high level function/tools (e.g. sklearn) except for reading and writing the files.**
- ☐ **DO NOT COPY!!!**（懶人包\考古題等禁止、其他資料有引用請註明）

Machine Learning 2017
Grading Policy & Homework Rules

- Homework will be graded by
  Completeness
  Correctness
  Algorithm description
  Discussion

- You should upload homework files to the FTP site
  Sever: 140.113.238.220
  Username: ML2017
  Password: ML2017
  Port: 634

- Homework Rules
  File Name: hw1_StudentID.zip/rar (e.g. hw1_1234567.zip)
  Code with comments
  You can use any programing language to finish your homework
  Report (.pdf format)
  ReadMe.txt (describes how to run your code)
  Hand in a hardcopy report on the due day.

- Deadline
  Late Submission (1-7 days): 70% score
  Don't accept after 7 days.