Homework 2

Part II Due on Monday, **4/23/2018**, at **10:00 AM** in class.

Part III Due on Monday, **4/23/2018**, at **9:00 AM** through E-campus platform.

Be noted that late homework will **NOT** be accepted!!

Grading and Submission Policy

- Please work all the problems. We definitely will grade the computer assignment in Part III. However, we will randomly pick only two problems in Part II to grade. The total is 80 points.
- You need to provide detailed proof or derivations that lead to your final answers. Add comments to your programming codes. You will receive no points if your answers are not supported by detailed explanations. So, do not skip steps.
- Partial points will be credited to you when a wrong answer is accompanied by detailed correct reasoning.
- For computer assignment, please encapsulate all the files (including codes and discussions) into a single compressed file. Name the compressed file by your student ***i.d. number*** and your ***name***. Then, upload it through the NCTU e-campus platform. (You don't need to print out hard copies for your codes.)
- Discussions are encouraged. But plagiarism is strictly prohibited. **You will fail the course as penalty for plagiarism!!**

Part I -- Reading Assignment

Plunge into Chapter 3 and start reading Chapter 4.

Part II -- Problem Assignment

1. (10 points)
   Exercise 2.26 and Exercise 3.11.

2. (10 points)
   Exercise 3.2.

3. (10 points)
   Exercise 3.8.

4. (10 points)
   Please explain the mathematical details behind Fig. 3.15 of the textbook. What are the math equations used to plot the circle and the ellipse? Why are the two axes of the ellipse $\mathbf{u}_1$ and $\mathbf{u}_2$? How to measure the $\mathbf{u}_1$-intercept and $\mathbf{u}_2$-intercept of the ellipse?

```
Part III -- Computer Assignment
```

In this problem, you need to apply the Maximum Likelihood (ML)and Bayesian linear regression methods to train a linear model in order to predict the relative humidity of a location in the United States.

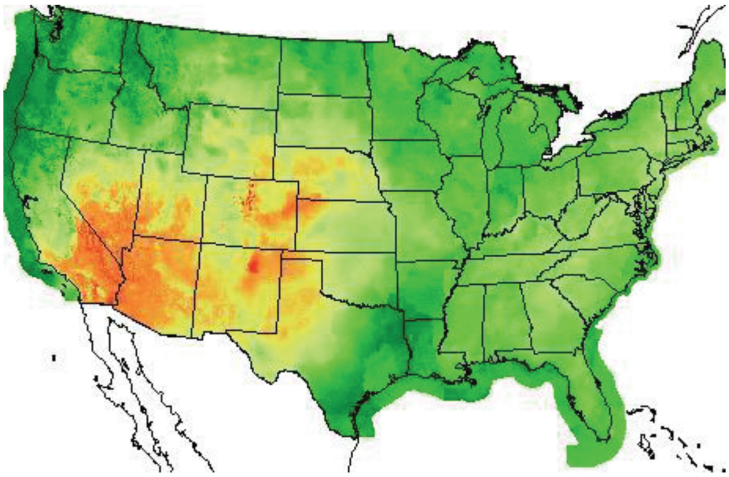

Figure 1: The illustration of humidity values across the United States.

Data

Data, contained in the two MAT files TrainingDataHW2.mat and TestingDataHW2.mat, in this problem are the measured relative humidity across the United States. More detailed descriptions are given below:

— In TrainingDataHW2.mat,
X_train is a 400 × 3 matrix, whose first column and second column record the vertical position x_1 and horizontal position x_2 of the 400 locations (i.e., the training points), respectively. The third column x_3 is an indicator to whether the location is at sea or on land, i.e., for each data point, $x_3 = 1$ for sea and $x_3 = 0$ for land.

T_train is a 400 × 1 vector, which records the relative humidity (target values) of the corresponding locations.

x1_bound is a vector recording the min/max values of the vertical position.

x2_bound is a vector recording the min/max values of the horizontal position

— In TestingDataHW2.mat,
X_test is a 53200 × 3 matrix, whose first column and second column record the vertical position and horizontal position of 53200 testing points, respectively. The third column records the sea indicator.

T_test is a 53200 × 1 vector, which records the relative humidity (target values) of the testing points.

The ground truth in the test data can be plotted in MATLAB using the following two lines:

```
imshow(reshape(T_test,x1_bound(2),x2_bound(2)));
colormap(jet);
```

Feature Vector

In this problem, we utilize the Gaussian basis function and the sea indicator to form the feature vector, denoted as

$$\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_P(\mathbf{x}), \phi_{P+1}(\mathbf{x}), \phi_{P+2}(\mathbf{x})]^T$$

where we place P Gaussian basis functions uniformly over the spatial domain with $P = O_1 \times O_2$, $\mathbf{x} = (x_1, x_2, x_3)$ is the input data (location together with the sea indicator), and O_1 and O_2 denote the number of locations along the horizontal and vertical directions, respectively, that you choose for your model in the prediction. (That is, you need to discuss the impact of different choices of O_1 and O_2.)

More specifically, for $1 \leq k \leq P$, the Gaussian basis function is defined as

$$\phi_k(\mathbf{x}) = \exp\left\{-\frac{(x_1 - \mu_i)^2}{2s_1^2} - \frac{(x_2 - \mu_j)^2}{2s_2^2}\right\}, \quad \text{for} \quad 1 \leq i \leq O_1, 1 \leq j \leq O_2,$$

where the subscript is $k = O_2 \times (i-1) + j$,

$$\mu_i = \left(\frac{x_{1\_\max} - x_{1\_\min}}{O_1 - 1}\right) \times (i-1), \quad \mu_j = \left(\frac{x_{2\_\max} - x_{2\_\min}}{O_2 - 1}\right) \times (j-1)$$
$$s_1 = \frac{x_{1\_\max} - x_{1\_\min}}{O_1 - 1}, \quad s_2 = \frac{x_{2\_\max} - x_{2\_\min}}{O_2 - 1}.$$

Finally, the last two components of the feature vector are $\phi_{P+1}(x) = x_3$ (sea indicator) and $\phi_{P+2}(x) = 1$ (bias).

Problem

(a) (60 Points)
Please employ the linear model

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{P+2} w_j \phi_j(\mathbf{x})$$

to predict the humidity of the locations given in the test data $\mathbf{x}$.

Please use 1) Maximum Likelihood and 2) Bayesian approach to train the model. Then, use your trained linear model to predict the relative humidity for each point in X_test. Plot the distribution of the predicted relative humidity $y(\mathbf{x})$ and the squared error distribution $(y(\mathbf{x}) - t(\mathbf{x}))^{\mathbf{2}}$ for all locations in the test data $\mathbf{x}$ across the United States. Explain the models you design in detail and discuss the differences among these three approaches.

For the Bayesian approach, you can assume the prior $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_0 = \mathbf{0}, \mathbf{S}_0^{-1} = \mathbf{I}\right)$.

(b) (Bonus Points)
You are encouraged to design and to discuss your own regression method to obtain more accurate prediction results.