Entrance Challenge: When Will the Sakura Bloom?

Name： Nguyen Van Tuong

Submission Date：

In [ ]:

```

```

0. Basics of the Sakura Bloom-cycle (5pts total)

In a year, sakura trees basically go through 4 phases: energy production, hibernation, growth, and of course flowering. These phases roughly follow the seasons, but not exactly.

Production phase： Initial development of the buds（Summer-Fall）
Hibernation phase： Bud growth stops while the tree goes into hibernation（Late Fall-Winter）
Growth phase： Buds once again continue to grow when the tree comes out of its winter hibernation（Late Winter-Spring）
Flowering phase： The buds finally bloom in spring (as climate conditions allow), once they have been able to fully develop.（Spring）

Each year, near the end of winter but before the trees finally bloom, the hibernation period ends. The sakura that rested through the winter once gain become metabolically active, and the buds continue to grow (though we may not immediately notice when this happens.) However, the cycle is not simply clockwork- for example, in places where the temperature is above 20℃ year-round, the trees are unable to hibernate sufficiently, and thus cannot blossom.

In this challenge, we have outlined the basic mechanism by which the sakura reach their eventual bloom-date. We consider building a bloom-date prediction model for the case of sakura in Tokyo, with the data split as follows:

Test years： 1966, 1971, 1985, 1994, and 2008
Training years: 1961 to 2017 (Excluding the test years)

You should fit the model to the data from the training years, then use the model to predict the bloom-date for each of the test years. The 3 models to be applied to the data are described below.

Problem 0-1: (5pts)

Acquire data of sakura blooming date (桜の開花日) for Tokyo from 1961 to 2017 using the Japanese Meteorological Agency website (気象庁).

In [ ]:

1. Prediction using the "600 Degree Rule" (15pts total)

For a rough approximaton of the bloom-date, we start with a simple "rule-based" prediction model, called the "600 Degree Rule". The rule consists of logging the maximum temperature of each day, starting on February 1st, and sum these temperatures until the sum surpasses 600˚C. The day that this happens is the predicted bloom-date. This 600˚C threshold is used to easily predict bloom-date in various locations varies by location. However, for more precise predictions, it should be set differently for every location. In this challenge, we verify the accuracy of the "600 Degree Rule" in the case of Tokyo.

Problem 1-1: (5pts)

From here-on, we refer to the bloom-date in a given year j as BD_j. For each year in the training data, calculate the accumulated daily maximum temperature from February 1st to the actual bloom-date BD_j, and plot this accumulated value over the training period. Then, average this accumulated value as T_{mean}, and verify whether we should use 600˚C as a rule for Tokyo.

In [17]:

Problem 1-2: (10pts)

Use the average accumulated value T_{mean} calculated in 1-1 to predict BD_j for each test year, and show the error from the actual BD_j. Compare to the prediction results when 600˚C is used a threshold value, and evaluate both models using the coefficient of determination (R^2 score).

In [1]:

2. Linear Regression Model: Transform to Standard Temperature (30pts total)

The year to year fluctuation of the bloom-date depends heavily upon the actual temperature fluctuation (not just the accumulated maximum). In order to get to a more physiologically realistic metric, Sugihara et al. (1986) considered the actual effect of temperature on biochemical activity.

They introduced a method of "standardizing" the temperatures measured, according to the fluctuation relative to a standard temperature.

In order to make such a standardization, we apply two major assumptions, outlined below.

1) The Arrhenius equation:

The first assumption, also known in thermodynamics as the "Arrhenius equation", deals with chemical reaction rates and can be written as follows:

$$k = A \exp\left(-\frac{E_a}{RT}\right)$$

Basically, it says that each reaction has an activation energy, E_a and a pre-exponential factor A. Knowing these values for the particular equation, we can find the rate constant k if we know the temperature, T, and applying the universal gas constant, $R = 8.314[\mathrm{J/K} \cdot \mathrm{mol}]$.

2) Constant output at constant temperature:

The second assumption, is simply that the output of a reaction is a simple product of the duration and the rate constant k, and that product is constant even at different temperatures.

$$tk = t'k' = t''k'' = \cdots = \mathrm{const}$$

Making the assumptions above, we can determine a "standard reaction time", t_s required for the bloom-date to occur. We can do so in the following way:

$$t_s = \exp\left(\frac{E_a(T_{i,j} - T_s)}{RT_{i,j}T_s}\right)$$

We define $T_{i,j}$ as the daily average temperature, and use a standard temperature of $T_s = 17^{\circ}\mathrm{C}$. For a given year j, with the last day of the hibernation phase set as D_j, we define the number of "transformed temperature days", DTS_J, needed to reach from D_j to the bloom-date BD_j with the following equation:

$$DTS_j = \sum_{i=D_j}^{BD_j} t_s = \sum_{i=D_j}^{BD_j} \exp\left(\frac{E_a(T_{i,j} - T_s)}{RT_{i,j}T_s}\right)$$

From that equation, we can find the average DTS for x number of years (DTS_{mean}) as follows:

$$
\begin{aligned}
DTS_{\text{mean}} &= \frac{1}{x}\sum_{j}^{x} DTS_j \\
&= \frac{1}{x}\sum_{j}^{x}\sum_{i=D_j}^{BD_j} \exp\left(\frac{E_a(T_{i,j} - T_s)}{RT_{i,j}T_s}\right)
\end{aligned}
$$

In this exercise, we assume that DTS_{mean} and E_a are constant values, and we use the data from the training years to fit these 2 constants. The exercise consists of 4 steps:

1. Calculate the last day of the hibernation phase D_j for every year j.
2. For every year j, calculate DTS_j as a function of E_a, then calculate the average (over training years) DTS_{mean} also as a function of E_a.
3. For every year j, and for every value of E_a, accumulate t_s from D_j and predict the bloom date BD_j^{pred} as the day the accumulated value surpasses DTS_{mean}. Calculate the bloom date prediction error as a function of E_a, and find the optimal E_a value that minimizes that error.
4. Use the previously calculated values of D_j, DTS_{mean}, and E_a to predict bloom-day on years from the test set.

Problem 2-1: (5pts)

According to Hayashi et al. (2012), the day on which the sakura will awaken from their hibernation phase, D_j, for a given location, can be approximated by the following equation:

$$D_j = 136.75 - 7.689\phi + 0.133\phi^2 - 1.307\ln L + 0.144T_F + 0.285T_F^2$$

where ϕ is the latitude [°N], L is the distance from the nearest coastline [km], and T_F is that location's average temperature [°C] over the first 3 months of a given year. In the case of Tokyo, $\phi = 35°40'$ and $L = 4\text{km}$.

Find the D_j value for every year j from 1961 to 2017 (including the test years), and plot this value on a graph.

(In Problem 1, we had assumed a D_j of February 1st.)

In [1]:

Problem 2-2: (10pts)

Calcluate DTS_j for each year j in the training set for discrete values of E_a, varying from 5 to 40kcal ($E_a = 5, 6, 7, \cdots, 40\,\text{kcal}$), and plot this DTS_j against E_a. Also calculate the average of DTS_j over the training period, and indicate it on the plot as DTS_{mean}. Pay attention to the units of **every parameter** ($T_{i,j}$, E_a, ...) in the equation for t_s.

In [ ]:

Problem 2-3: (11pts)

Using the same E_a values and calculated DTS_{mean} from 2-2, predict the bloom date BD_j for each of the training years. Find the mean squared error relative to the actual BD and plot it against E_a. Find the optimal E_a^* that minimizes that error on the training data.

In [ ]:

Problem 2-4: (4pts)

Using the D_j dates from problem 2-1, the average DTS_{mean} from 2-2, and the best-fit E_a^* from 2-3, predict the bloom-dates BD_j for the years in the test set. Determine the error between your predicted BD_j values and the actual values, and evaluate this model using the coefficient of determination (R^2 score).

In [ ]:

Problem 2-5: (extra 10pts)

Discuss any improvements you could make to the model outlined above. If you have a suggestion in particular, describe it. How much do you think the accuracy would be improved?

In [ ]:

---

3. Predicting Bloom-date via Neural Network (30pts total)

Problem 3-1: (20pts)

Build a neural network and train it on the data from the training years. Use this model to predict the bloom-dates for each year in the test set. Evaluate the error between predicted dates and actual dates using the coefficient of determination (R2 score). Only use the weather data given in `tokyo.csv` and the sakura data acquired in problem 0-1. You may use whichever framework or strategy that you like to construct the network.

In [ ]:

Problem 3-2: (10pts)

Compare the performance (via R^2 score) of the 3 implementations above: the 600 Degree Rule, the DTS method, and the neural network approach. For all methods, and each test year, plot the predicted date vs. the actual date. Discuss the accuracy and differences of these 3 models.

In [ ]:

4. Trends of the Sakura blooming phenomenon (20pts total)

Problem 4-1: (20pts)

Based on the data from the past 60 years, investigate and discuss trends in the sakura hibernation (D_j) and blooming (BD_j) phenomena in Tokyo.

In [ ]: