COMP7103 Assignment 2

Due date: Apr 19, 2018 11:55pm

Question 1 Classification

Consider a house installed with 60m^2 of solar panel that operates in two modes: high power mode and power saving mode. The solar panel produces and consumes energy in a different rate depending on the mode it is operating. At the end of each day, the house owner would like to choose the best mode of operation for the next day based on the weather of the past 5 days. You are asked to build a classification model for the decision.

To build a model that predicts the best mode of operation, you are given a data set, **`weather2016.csv`** (available on Moodle), which contains the **weather data** of Hong Kong in the year of 2016. The file **`weather.txt`** (also available on Moodle) describes the attributes in the dataset. The prediction result should either be "**High**" (high power mode), "**Low**" (power saving mode), or "**Off**" (completely turned off).

In high power mode, efficiency of the panel is 20%. At the same time, the power system of the house needs 4MJ/hr to run in this mode. Therefore, the net energy production in this mode is:

$$E_{high}\ (\text{MJ}) = Radiation\ (\text{MJ}) \times 20\% - 24\ (\text{hr}) \times 4\ (\text{MJ/hr})$$

where amount of radiation can be calculated by:

$$Radiation\ (\text{MJ}) = Total\ solar\ radiation\ (\text{MJ/m}^2) \times Size\ of\ panel\ (\text{m}^2)$$

In power saving mode, the solar panel reduces the energy needed to run the power system by 1MJ/hr except when the panel is under a bright sunshine. However, efficiency of the panel is reduced to 18%. Therefore, the net energy production in this mode is:

$$E_{low}(\text{MJ}) = Radiation\ (\text{MJ}) \times 18\% - 24\ (\text{hr}) \times 3\ (\text{MJ/hr}) - Sunshine\ (\text{hr}) \times 1\ (\text{MJ/hr})$$

If the solar panel is switched off, no energy is consumed or produced ($E_{low} = 0$)

For example, Table 1 shows the energy generation of the first three days in year 2016 and their corresponding best mode of operation.

Day	Sunshine	Radiation	Energy production	Best
1	9.3	15.24	$E_{high} = 15.24 \times 60 \times 20\% - 24 \times 4 = \mathbf{86.9}$ $E_{low} = 15.24 \times 60 \times 18\% - 24 \times 3 - 9.3 \times 1 = 83.3$ $E_{off} = 0$	High
2	0.6	7.77	$E_{high} = 7.77 \times 60 \times 20\% - 24 \times 4 = -2.8$ $E_{low} = 7.77 \times 60 \times 18\% - 24 \times 3 - 0.6 \times 1 = \mathbf{11.3}$ $E_{off} = 0$	Low
3	–	4.39	$E_{high} = 4.39 \times 60 \times 20\% - 24 \times 4 = -43.3$ $E_{low} = 4.39 \times 60 \times 18\% - 24 \times 3 - 0 \times 1 = -24.6$ $E_{off} = \mathbf{0}$	Off

Table 1

Use Weka and/or any other tools, find a **classification model** that predicts the best mode the solar panel should operate in a day, using the weather data of at most 5 previous days of the predicting day. For example, to predict the best mode on 2017-01-15, your model can only use the weather data on 2017-01-10 to 2017-01-14.

You should also **justify** that your model is a good one by comparing evaluation results of different models.

Answer the following questions.

a) What is your **final model** and the corresponding evaluation result?
b) Briefly describe **what you have done** in finding the model, including but not limited to data preprocessing, attribute selection, parameter tuning, training and testing data construction, model building, and evaluation etc. You may omit some of the above mentioned items if you have not done that. You **should highlight and explain the choice** you have made in the process.
c) Given a data set, **`weather2017.csv`** (available on Moodle), which contains the **weather data** of Hong Kong in the year of 2017. Predict the best mode of operation on the dates of 2017-02-10, 2017-05-30, 2017-08-28, and 2017-11-18 using the final model in part a. Show the result of predictions and the actual best mode of operations on these days. Comment on the predictability of your model based on the result.
d) Using the final model in part a, **predict** the best mode of operation on the date of 2018-01-01, using the weather data in **`weather2017.csv`**. **Show clearly** how a prediction can be manually generated from the model. If it is impossible to use the model manually, where tools or utilities (e.g., Weka) must be employed to generate the prediction, **explain** instead what these tools or utilities do to generate a prediction.

Question 2 Association analysis

Given the sequence database shown in Table 2. Answer the following questions.

Sensor	Timestamp	Events
S1	1	B, D
	2	C
	3	A, B
	4	A
S2	1	A, C
	2	A, B
	3	C

Sensor	Timestamp	Events
S3	1	B
	2	C, D
	3	A, B
S4	1	B
	2	C
	3	A, B, C
	4	C

Sensor	Timestamp	Events
S5	1	C
	2	A, B
	3	D
	4	A, C, D

Table 2

a) List all the 4-element subsequences contained in the data sequence of Sensor S1.

b) Find all frequent subsequences with support $\geq 60\%$ given the sequence database shown in Table 2 using GSP algorithm as discussed in Lecture Notes Chapter 7, p.65. Show the result of Candidate Generation, Candidate Pruning, and Support Counting, as well as the frequent sequences found after each pass.