Programming Homework 3

CS540

June 14, 2019

1 Instruction

Please submit your output files and code on Canvas → Assignments → P3. Please do not put code into zip files and do not submit data files. The homework can be submitted within 2 weeks after the due date on Canvas without penalty (50 percent penalty after that).
Please add a file named ”comments.txt”, and in the file, you must include the instructions on how to generate the output, for example:

- Data files required: train.csv, test.csv. Run: main.jar.
- Data folder required: data/train1.png ... data/train100.png . Compile and Run: main.java.

2 Details

All the requirements are listed on the course website. The following is only an example workflow to solve the problem.

1. Download the data. It is recommended that you remove the irrelavent columns and remove the rows with missing data programmatically, but it is acceptable to do these manually in Excel too (in this case, put a note in the comments file, but please do NOT submit the data).
2. For the revenue variable, create (double, not integer) variable y using your threshold (get it on the website using your ID). Suppose your threshold is B millions.

$$y_i = \begin{cases} 0 & \text{if revenue }_i \leqslant B \\ 1 & \text{if revenue }_i > B \end{cases}$$

3. For other features: budget, runtime, vote average, and vote count should be uniformly split into categories, say if you want K categories, find the minimum a, the maxmimum b and set the category as the following:

$$x_{ij} = \text{ floor } \left(\frac{\text{variable }_j - a}{\frac{b + \varepsilon - a}{K}} \right), \varepsilon = 0.01$$

Here, the formula tries to the divide the range $[a, b]$ into K intervals, and label the values in the first interval 1, the second interval 2, and so on. Let $\ell = \frac{b + 0.01 - a}{K}$ be the length of each interval. floor $\left(\frac{x - a}{\ell}\right)$ will map $[a, a + \ell)$ to $1, [a + \ell, a + 2\ell)$ to 2, and so on, which is exactly the desired label. Note that 0.01 is added so that the b is labelled K, not $K + 1$. Make sure you understand what this formula is doing.

4. For the genre feature, you should combine some of the genres to make the number of categories smaller and label them arbitrarily, for example, action is 0, sci-fi is 1, etc. Note: depending on the version of the file you downloaded, you may need to take out genre value before you split the string for the row. For example, if you read a row that looks like "1, 2, ["action","scifi","drama"], 3, 4" , then you should remove the substring inside the square brackets first to make it "1, 2, 3, 4" and then split it using comma because there are unknown number of commas inside the genre substring.

5. Compute the information gain using the formula given in the slides.

$$\begin{aligned} I\left(Y|X_j\right) &= H\left(Y\right) - H\left(Y|X_j\right) \\ H\left(Y\right) &= -\sum_{y=0}^{1} \frac{\#\left(Y = y\right)}{\#\left(Y\right)} \log\left(\frac{\#\left(Y = y\right)}{\#\left(Y\right)}\right) \\ H\left(Y|X_j\right) &= -\sum_{x=1}^{K}\sum_{y=0}^{1} \frac{\#\left(Y = y, X_j = x\right)}{\#\left(Y\right)} \log\left(\frac{\#\left(Y = y, X_j = x\right)}{\#\left(X_j = x\right)}\right) \end{aligned}$$

Here, $\#\left(Y = 0\right)$ is the number of instances with label 0 and $\#\left(Y = 1\right)$ is the number of the instances with label 1. And $\#\left(Y\right)$ is the total number of instances. $\#\left(Y = 0, X_j = 1\right)$ is the number of the instance with label 0 and feature j value 1 and $\#\left(Y = 0, X_j = 2\right)$ is the number of the instance with label 0 and feature j value 2 and so on ...

If the count $\#\left(X_j = x\right)$ is 0, do not include this category x in the sum. Also, use the $0 \log 0 = 0$ convention.

6. Output the information gain and the counts.