CSCI4343 Project B: Coffee Pour Over Challenge

Yifeng Gao

November 2023

1 Introduction

In this project, you will try to become a coffee data miner. Pour over techniques is a very popular techniques that used in a lot of coffee shop. But to make a delicious coffee is not very simple. For example, to achieve the best coffee brewing performance, different world champion invented different type of strategies. If you like to know, here is a incomplete listed:

- 46 Method
- Rao Method
- Five pour Receipt
- Centre Pour Method

There are many different variations of filters and the design of pour over system, V60, Kalita, Origami, each one has its own features.

With the advance of sensor techniques, now, we could monitor the pour over method via **data**! We placed a smart scale to measure the water volume increasing throughout the process. You responsibility is: **Can we predict if the cup of coffee will be delicious given a set of data**. You will provide a set of data and the class label (delicious or not judged by users)

You will receive a file named "coffee.csv" where the data and the label (whether the coffee is tasty or not). You need to do the following tasks:

2 Task 1: Preprocessing (10 pt)

Like our Katydid vs. Grasshopper example in the class, we need to perform preprocessing to obtain a set of useful features. The raw data is shown in the following figure where x axis is time stamp and y axis is the water volume. Obviously,

Task 1.1 Extract a set of features that you think will benefit your project.

Task 1.2 Understand data behavior, try to summarize and report what the characteristics of the data and understand what behavior caused the issue.

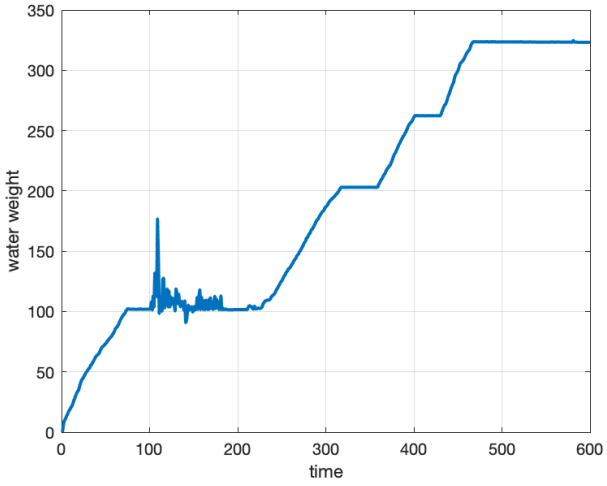

Figure 1: An Example of Brewing Data

tasty	brew data
2	0.00;5.10;9.10;10.40;12.00;13.70;15.00;17.10;18.20;19.40;20.90;21.80;23.90;26.20;27.90;30.20;31.80;33.50;34.30;36.40;39.20;41.
2	0.00;1.20;2.20;2.30;3.00;3.60;3.90;4.30;4.80;5.00;5.50;5.90;6.40;6.60;7.10;7.50;7.90;8.30;8.50;8.70;9.30;9.39;10.00;10.60;10.80;1
1	0.00;0.50;2.00;2.70;3.00;3.10;4.00;4.10;5.00;5.30;5.80;6.20;7.10;7.30;8.00;8.30;8.89;9.10;9.89;10.20;10.80;11.00;11.70;11.80;12.5
2	0.00;4.00;7.30;7.60;8.89;10.30;11.20;13.40;15.40;16.90;17.79;18.50;20.29;22.60;23.70;25.60;27.20;27.80;28.90;31.20;33.30;34.59
1	0.00;2.20;3.90;5.00;6.00;8.10;10.60;12.50;14.70;18.00;19.79;23.30;26.10;28.60;30.20;34.50;38.00;40.40;41.80;45.40;48.60;51.20;!
2	0.00;4.80;9.60;10.70;11.80;13.20;14.80;17.10;18.79;20.00;20.50;20.90;23.60;25.70;26.30;27.50;28.90;30.50;31.80;32.70;34.30;35.

Figure 2: coffee.csv

Bonus A: Feel free to share your discovered features in the discord, your feature turn out to be useful in the end and you are first mentioned it in the discord, your team will get 1 pt bonus.

Bonus B: Feel free to share your discovered data characteristics in the discord, if you finding turn out to be useful and correct and you are first mentioned it in the discord, your team will get 1 pt bonus.

3 Task 2: Model Training (10 pt)

After you extracted the feature, the next step is picking the best model. Feel free to try whatever model you want to train a classification model. You model should complete the following things:

Task 2.1 Pick one potential model we talked about in the class (Linear Classifier, MLP, SVM, Kernel SVM, Naive Bayes). And apply it to make the prediction.

Task 2.2 Apply techniques discussed in the class to conquer over-fitting issue

Task 2.3 Try to summarize the rule if found via the classifier.

4 Bonus: Entering Competition

Later in the class. we will release testing data in the Kaggle Competition platform.

Bonus A: Whoever can explain a classification rule that is explainable (persuaded your classmate) can receive additionally 1 pt bonus.

Bonus B: Top-2 team will receive additionally 1 pt bonus

Bonus Jacket Pot: There is some "generated" fake data in the testing data, if you spot it and realize how I generate these fake data, please report it in the discord. You team will receive 3 pt bonus credit. You need to clearly state how I generate the "fake" data.