Machine Learning – I

Ramakrishna Mission Vivekananda University

Assignment 3

Posted on 20 Apr 2017 | Seek clarification by 27 Apr 2017 | Submit by 4 May 2017

Your solutions for the problems in this assignment should be in a single `R` file `GroupXX_a3.R` or in a single Python file `GroupXX_a3.py`, where `XX` is your group number (i.e., `XX` = 01 to 10). The code must be properly commented. The solution file should be emailed to `sg.sourav@gmail.com`, with a copy to all your group members, *before* midnight on the submission deadline.

Properly acknowledge every source of information that you refer to, including discussions with your friends outside the group. Verbatim copy from any source is strongly discouraged, and plagiarism will be heavily penalized. You should try to write the codes completely on your own.

Problem 1 [20 points]

Handwritten Images Dataset: Digit Recognizer
`https://www.kaggle.com/c/digit-recognizer`

Background: MNIST (Modified National Institute of Standards and Technology) is the de facto "hello world" dataset of computer vision. Since its release in 1999, this classic dataset of handwritten images has served as the basis for benchmarking classification algorithms. As new machine learning techniques emerge, MNIST remains a reliable resource for researchers and learners alike. In this assignment, you will learn the application of SVM on this classic dataset.

In this challenge, your goal is to correctly identify digits from a dataset of handwritten images. In particular, you should apply Support Vector Machines to classify handwritten images to match decimal digits. You may treat it as a classic case-study of Multi-class Classification with images.

Problem statement: Take part in the basic online Kaggle Competition – "Digit Recognizer" – available at `https://www.kaggle.com/c/digit-recognizer`. Submit your predictions online.

Submission: Submit the `R` (or Python) code you wrote for the Kaggle competition as the first part of your submission file `GroupXX_a3.R`. In the commented section, please mention the name of your Kaggle team and your final score/standing on the leaderboard. You may use any public *Kernel* or *Discussion* posted on Kaggle, with proper acknowledgement for each such reference.

Restriction: Please restrict yourselves to Support Vector Machines (SVM) and Kernel Methods for the classification. You may use Principal Component Analysis (or SVD) too, if required.

Problem 2 [30 points]

Cuisine Classification and Optimal Clustering
`http://www.souravsengupta.com/ml2017/evaluation/cuisinedata.zip`

Background: Some of our strongest geographic and cultural associations are tied to a region's local foods, and the locality of a food item has close ties with its ingredients. Every country or continent has its own type of cuisine and its own list of ingredients, and quite often, the origin of a dish can be identified just from the list of ingredients mentioned in its recipe.

In the *first part* of this challenge, we ask you to predict a dish's cuisine given a list of its ingredients. In particular, we ask you to apply the tools of machine learning to classify recipes into cuisines. You are allowed to use the full dataset, along with the `"cuisine"` labels.

In the *second part* of this challenge, we ask you to find the optimal number of clusters that you can spot in the dataset, without using the `"cuisine"` labels. In particular, we ask you to apply the tools of Unsupervised Learning to find out how many "natural" types of cuisines are present in the dataset. You may only use `"id"` and `"ingredients"`, but NOT the `"cuisine"` labels.

Part 1 : In the *first part* of the challenge, use as training set the complete labeled dataset available at `http://www.souravsengupta.com/ml2017/evaluation/cuisinedata.zip` to build a robust Support Vector Machine based multi-class classifier that is capable of identifying the `"cuisine"` of a dish from the list of `"ingredients"` in its recipe.

Part 2 : In the *second part* of the challenge, use as training set the unlabelled version of the dataset available at `http://www.souravsengupta.com/ml2017/evaluation/cuisinedata.zip`, that is, after removing the `"cuisine"` labels, and find the *optimal* number of clusters in the data.

Submission: Submit the `R` (or Python) code you wrote for this challenge as the second part of your submission file `GroupXX_a3.R`. In the commented section, please mention the main JSON processing packages you used, and acknowledge any online/offline resources you have consulted. Also mention your reason (justification) for choosing the *optimal* number of clusters in the data.

Dataset: The uncompressed training set stores the observations (recipes) in JSON format, as follows, where the first element is a unique identifier (`"id"`), the second element is the cuisine label (`"cuisine"`), and the third element is the list of ingredients (`"ingredients"`) of a dish.

```
{
"id": 24717,
"cuisine": "indian",
"ingredients": [
    "tumeric",
    "vegetable stock",
    "tomatoes",
    "garam masala",
    "naan",
    "red lentils",
    "red chili peppers",
    "onions",
    "spinach",
    "sweet potatoes"
]
}
```

Cuisine dataset — `http://www.souravsengupta.com/ml2017/evaluation/cuisinedata.zip`