Computing for Data Sciences – 2017

PGDBA, First Year, First Semester, Indian Statistical Institute

Assignment 4

Posted on 5 November 2017 | Clarify by 10 November 2017 | Submit by 15 November 2017

Problem 1 [60 points]

Discovering *meaningful* Cuisine Clusters
`http://www.souravsengupta.com/cds2017/evaluation/cuisine.json`

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.

Problem statement: In this challenge, you are required to find the *optimal* number of clusters that you can spot in the dataset. In particular, you may apply the tools you know in Unsupervised Learning to find out how many "natural" types of cuisines are present in the dataset.

Submission: Submit the `R` (or Python) code you wrote to solve this problem as a single program file – `groupXXassign4prob1.R` (or `groupXXassign4prob1.py`), where `XX` is your group number. In the commented section, please mention the JSON processing packages you used, and acknowledge any online/offline resources you have consulted. Also mention, very briefly, the main reason and your justification for choosing the *optimal* number of clusters in the data.

Dataset: The dataset stores the recipes in JSON format, where the first element is a unique identifier (`"id"`) and the second element is the list of ingredients (`"ingredients"`) of a dish.

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

Cuisine dataset — `http://www.souravsengupta.com/cds2017/evaluation/cuisine.json`

Problem 2 [40 points]

Discovering *meaningful* Demographic Clusters
`https://www.kaggle.com/miroslavsabo/young-people-survey`

Background: Every individual has a unique set of preferences when it comes to music, movies, hobbies, and interests. In fact, this extends (and sometimes relates) to their health habits, phobias, personality traits, lifestyle, spending habits, and even opinions. A survey was conducted in 2013, with 1010 Slovakian young adults, aged between 15-30, to learn about their preferences vis-a-vis their demography. The dataset records the responses to the survey, over 150 attributes.

Problem statement: In this challenge, you are required to find the *optimal* number of clusters that you can spot in the dataset. In particular, you may apply the tools you know in Unsupervised Learning to find out how many "natural" types of cuisines are present in the dataset. Based on the clusters you observe, form an educated opinion regarding the *strongest* clustering parameters in this set of young adults. Do you think that gender plays a major role in clustering?

Submission: Submit the `R` (or Python) code you wrote to solve this problem as a single program file – `groupXXassign4prob2.R` (or `groupXXassign4prob2.py`), where `XX` is your group number. In the commented section, please mention your findings — number of *optimal* clusters, major clustering parameters, and whether gender is at all a deciding factor in preferences and interests. Briefly mention your justification for choosing the *optimal* number of clusters in the dataset.

Dataset: The data is a structured CSV file, consisting of 1010 rows and 150 columns (features), in which 139 variables are integer and 11 are categorical responses to the survey questionnaire.

Survey dataset on Kaggle — `https://www.kaggle.com/miroslavsabo/young-people-survey`

Your submission should be emailed to `sg.sourav@gmail.com` by midnight of 15 November 2017.

Properly acknowledge every source of information that you referred to, including discussions with other groups. Verbatim copy from any source is strongly discouraged, and plagiarism will be heavily penalized. It is strongly recommended that you write the codes completely on your own.