Home Work 2 Report - Data Mining and Machine Learning
By : R95922173 伍妮

The goal of HW : Get the experience how to generate flat file from UCI university data set transform into LIBSVM data set.

Programming language that is used to generate flat file : **Python2.5**
Step by step :

1. Download and Install python2.5
2. set path path=c:\phyton25
3. Classify attributes in UCI data set

Table 1. Number of Value of UCI data set attributes

No	Attribute Name	Number of Value
1	State	38
2	location	4
3	Control	5
4	number-of-students	5
5	male:female (ratio)	55
6	student:faculty (ratio)	23
7	sat-verbal	55
8	sat-math	59
9	expenses	4
10	percent-financial-aid	20
11	number-of-applicants	6
12	percent-admittance	19
13	percent-enrolled	19
14	academics	5
15	social	5
16	quality-of-life	5
17	academic-emphasis (target class)	117

4. Algorithm :
    a. Create function

Table 2. List of function

No	Name of Function	purpose
1	instan()	to check the first line of definition of university
2	empha()	to check value of academic emphasis and convert to numeric value
3	state()	to check value of state attribute and convert to numeric value
4	location()	to check value of location attribute and convert to numeric value
5	control()	to check value of control attribute and convert to numeric value
6	num_stu()	to check value of number of student and convert to numeric value
7	mf_ratio()	to check value of ratio male and female and convert to numeric value
8	stufa_ratio()	to check value of ratio of student and faculty and convert to numeric value

9	satver()	to check value of sat verb attribute and convert to numeric value
10	satmath()	to check value of sat math attribute and convert to numeric value
11	expense()	to check value of expenses attribute and convert to numeric value
12	percent_fin()	to check value of percent financial attribute and convert to numeric value
13	no_ap()	to check value of number of applicants attribute and convert to numeric value
14	percent_ad()	to check value of percent admittance attribute and convert to numeric value
15	percent_en()	to check value of percent enrolled attribute and convert to numeric value
16	acad()	to check value of academics attribute and convert to numeric value
17	social()	to check value of social attribute and convert to numeric value
18	quality()	to check value of quality of life attribute and convert to numeric value

Re (regular expression) **module** with search function is used to check the value of each attribute.

Before re module is used, call modul with ***import re*** syntax.

Writelines() is used to write *attribute number : value* if data match (found).

Example :

```
p=re.compile('state alabama',re.IGNORECASE)
b=p.search(a)
if b!=None:
  g.writelines('1:1 ')
```

b. Open source file in read mode with syntax :
f=open('file_name','r')
r : read mode

c. Open destination file in append mode with syntax :
g=open('filename','a')
a : append mode

d. Use **readline()** function to read line in source file and store in variable

e. Call function to check expression in source file

f. Run the program to check the result, It need around 2 minutes to run the program and get the result

Result and Evaluation :

1. Program can transform UCI data set to LIBSVM format. Result example:

```
22,20 1:25 3:2 4:1 5:21 6:8 7:25 8:9 10:12 11:1 12:13 13:7 14:2 15:2 16:2
24,41,1,43 1:2 3:5 4:5 5:36 6:13 7:9 8:14 10:1 11:6 12:15 13:11 14:3 15:4 16:5
```

2. difficulties and solve:

Problem	Solve
In windows based, university.data format can not be read	transform to txt format
Put the class target in the beginning of file	use tell() function and seek() function to go to certain position