Machine Learning Assignment 3

0416303 資工資電 楊博凱

Implementation :

在這次作業中，一開始我先在系統中以 sudo pip install matplotlib 以及 sudo apt-get install python-tk 安裝自己會用到的函示庫等等；之後便 import 一些會用到的 function 在程式中。

首先是第一題的部分，由於我已經將第一題的 test data 存入自己建立的 txt 檔中，所以要先將他 open 進來。令好 descriptive feature、target feature 以及 test data 後，將它們丟入 GaussianNB 中（ 由於 GaussianNB 自帶 smoothing 功能，所以我就沒有將他帶入自己寫的 smoothing function 之中)，並判斷出 target of test data 為 settler。

而在第二題的部分中，首先我先建立 smoothListGaussian()，以 moving average 的概念撰寫 smoothing function。在 data 的 preprocessing 中，我先將資料分成處理 data 與 attr 檔案兩部分：處理 data 部份時，先將資料以 '\r\n' 做 split，再分別以 ',' 做 split，再來將資料中的 training 與 testing data 分開來，之後刪除掉含有 '?' 的資料，成為最終拿來使用的 data。

在處理 attr 時比較複雜，首先我一樣將資料以 '\r\n' 隔開，由於有些 date 是以空白而非以逗號隔開，所以還要將 ' ' 轉成 ',' 。再來，我將所有含有 class 以及 Date 的資料提取出來，並將所有不必要的字元(空白與點點)刪除。之後先將資料放進新的 list 中，並以 class type 作為分類依據；再來將 'to' 前後的日期抓出來 append 他們之間缺少的日期；最後 remove 'to' 則完成 attr 的整理。

最後將 training 與 testing data 的日期改成相對應的 class(由於使用 if-else 判斷，所以若該日期沒有被分配到任何 class 則會屬於 class 6 (跑進 else 中))。經過 training data 的 K-Fold cross validation，測出來的準確度大致上維持於 80%~90%左右。

```
>>> # ************************************************************
... # * GaussianNB for K-Fold Validation                          *
... # ************************************************************
...
>>> for num in range(len(y_train[0])):
...     y_train[:,num] = np.array(smoothListGaussian(y_train[:,num]));
...
>>> clf = GaussianNB()
>>> clf.fit(y_train, x_train)
GaussianNB(priors=None)
>>> print confusion_matrix(x_test,clf.predict(y_test))
[[102   6   0   4   0]
 [  5  30   0   5   0]
 [  0   1   0   0   0]
 [  2   0   0  22   0]
 [  0   2   0   0   0]]
>>> print clf.score(y_test, x_test)
0.860335195531
```

Results :

```
:w !python
********** RESULT OF QUESTION ONE **********
['settler']
********** RESULT OF QUESTION TWO **********
[ 1.  5.  3.  1.  1.  1.  5.  3.  5.  5.  1.  1.  3.  5.  3.  1.  1.  1.
  1.]
```

Using library :

首先先幫系統安裝 matplotlib 以及 python-tk (曾經要看 feature 的分布，之後判斷為 gaussian 分布後，才套用 GaussianNB)

Import 的 library 有：numpy、string、random、sklearn、matplotlib.pyplot

```
# sudo pip install matplotlib
# sudo apt-get install python-tk
import numpy as np
import string
import random
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import StratifiedKFold
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import BernoulliNB
import matplotlib.pyplot as plt
```

Environment & Language :

```
kai@kai-UX305FA:~$ sudo lsb_release -a
[sudo] password for kai:
No LSB modules are available.
Distributor ID: Ubuntu
Description:    Ubuntu 16.04.2 LTS
Release:        16.04
Codename:       xenial
```

```
kai@kai-UX305FA:~$ python
Python 2.7.12 (default, Nov 19 2016, 06:48:10)
[GCC 5.4.0 20160609] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>>
```