Naïve Bayes Classifier for Spam Filter of Binary Dataset

Theory:

For this spam dataset is multivariate binary dataset, I choose Bernoulli Naïve Bayes Algorithm to predict the probability. Here is the equation:

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1 - p_{ki})^{(1-x_i)}$$

, $X_i \in (0,1)$, p_{ki} is the probability of class C_k given the value of X_i

Predict:

for test data set:

```
[0    0    0    1    0    1    1    1    1    1    1    1    0
      0    1    0    0    1    0    1    0    0    0    0    0
      0    0    0    0    0    0    0    0    0    0    0    0
      0    0    0    0    0    0    0    0    0    0    0    1
      0    1    0    0    1    1    1
]
```

$P(X|S) = p(X_1=0|S).p(X_2=0|S).p(X_3=0|S). p(X_4=1|S). p(X_5=0|S) \ldots .p(X_{57}=1|S). p(S)$

$P(X|NS) = p(X_1=0|NS).p(X_2=0|NS).p(X_3=0|NS).p(X_4=1|NS).p(X_5=0|NS) \ldots . p(X_{57}=1|NS). P(NS)$

If $p(X|S) > p(X|NS)$, it's Spam ($P_i = 1$);

If $p(X|S) < p(X|NS)$, it's not Spam ($P_i = 0$).

Accuracy:

Get the predict result of all test data P_i, and the actual result of test data T_i

If $P_i == T_i$, R+1 (R means right)

Count the total number of test data set N. So the accuracy = (R/N)*100%

For 10 folds' cross validation:

Loop the upper step 10 times, get the average accuracy of these ten times.

Result:

`After 10 times' cross validation, I get the average accuracy is` **`65.56%`**

Randomlize(shuffle) it before split to trainset and testset:

Code: data_Spam=data_Spam.sample(frac=1).reset_index(drop=True)

```
print(avgAccuracy)
```

```
0.655869565217
```

```
accuracy
```

```
array([ 0.68043478,  0.61521739,  0.65      ,  0.65217391,  0.62826087,
        0.68043478,  0.66521739,  0.66521739,  0.6326087 ,  0.68913043])
```

Not randomlize:

```
print(avgAccuracy)
```

```
0.655652173913
```

```
accuracy
```

```
array([ 0.30652174,  0.36956522,  0.35652174,  0.41304348,  0.81521739,
        0.85434783,  0.84565217,  0.87173913,  0.88043478,  0.84347826])
```