1.3

1. Report the mean and standard deviation of ten fold cross validation for the three datasets using logistic regression.

Spambase

```
              1         2         3         4         5         6         7  \
train  0.919826  0.921758  0.921275  0.922724  0.922965  0.921034  0.923448
test   0.921739  0.932609  0.919565  0.932609  0.910870  0.930435  0.895652

              8         9        10  mean accuracy  std accuracy
train  0.923690  0.918377  0.924897       0.922000      0.001949
test   0.913043  0.936957  0.902174       0.919565      0.013976
```

```
              1         2         3         4         5         6         7  \
train  0.910196  0.912465  0.912300  0.914957  0.914394  0.912215  0.914124
test   0.909944  0.925433  0.911633  0.923934  0.902167  0.921192  0.884869

              8         9        10  mean recall  std recall
train  0.915132  0.908519  0.915956     0.913026    0.002349
test   0.902463  0.931737  0.891674     0.910505    0.015332
```

```
              1         2         3         4         5         6         7  \
train  0.921496  0.922595  0.922174  0.922911  0.924000  0.922321  0.925054
test   0.924170  0.936748  0.919468  0.931433  0.907045  0.927997  0.895199

              8         9        10  mean precision  std precision
train  0.924553  0.919297  0.925799        0.923020       0.001907
test   0.913715  0.939312  0.907819        0.920291       0.014282
```

Breast Cancer

```
              1         2         3         4         5         6         7  \
train  0.976608  0.982456  0.982456  0.980507  0.980507  0.982456  0.978558
test   1.000000  0.964286  0.982143  0.928571  0.982143  0.964286  0.982143
              8         9        10  mean accuracy  std accuracy
train  0.980507  0.978558  0.980507       0.980312      0.001938
test   1.000000  0.982143  0.964286       0.975000      0.020960
```

```
              1         2         3         4         5         6         7  \
train  0.971092  0.977185  0.979567  0.976343  0.975750  0.978824  0.975464
test   1.000000  0.963542  0.977273  0.915535  0.978261  0.944444  0.976190
              8         9        10  mean recall  std recall
train  0.975646  0.972659  0.975750     0.975828    0.002532
test   1.000000  0.986842  0.956522     0.969861    0.026004
```

```
              1         2         3         4         5         6         7  \
train  0.979333  0.985189  0.982776  0.982363  0.982414  0.983907  0.978592
test   1.000000  0.963542  0.985714  0.915535  0.985294  0.975000  0.986111
              8         9        10  mean precision  std precision
train  0.982421  0.982053  0.982414        0.982146       0.001931
test   1.000000  0.973684  0.971429        0.975631       0.024203
```

Pima Indian Diabetes

```
             1         2         3         4         5         6         7  \
train  0.781792  0.790462  0.784682  0.786127  0.787572  0.771676  0.789017
test   0.789474  0.710526  0.750000  0.736842  0.750000  0.842105  0.710526
              8         9        10  mean accuracy  std accuracy
train  0.777457  0.780347  0.780347       0.782948      0.005772
test   0.815789  0.802632  0.802632       0.771053      0.045665
```

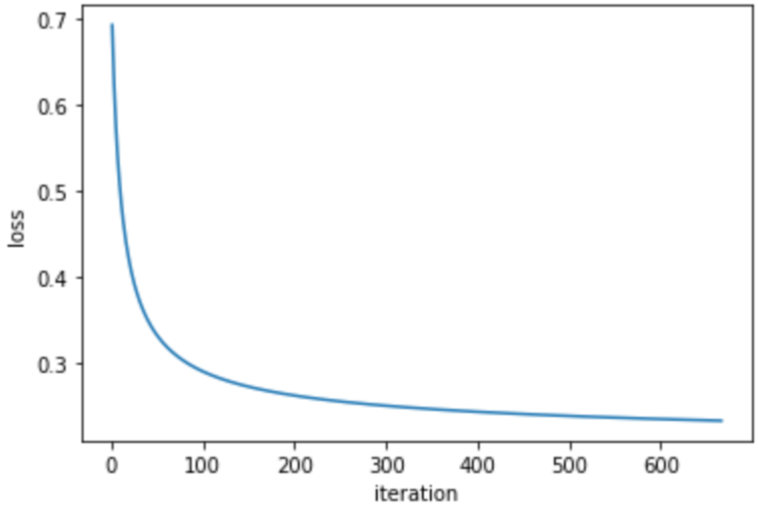
```
              1         2         3         4         5         6         7  \
train  0.736956  0.742356  0.738688  0.736526  0.743788  0.726097  0.737939
test   0.712018  0.668116  0.698826  0.717330  0.701569  0.817316  0.694856
              8         9        10  mean recall  std recall
train  0.735095  0.735180  0.735388     0.736801    0.004793
test   0.754978  0.754808  0.750980     0.727080    0.042879
```

```
              1         2         3         4         5         6         7  \
train  0.766582  0.777302  0.769188  0.770453  0.775318  0.759706  0.774335
test   0.842491  0.707143  0.759579  0.735385  0.716330  0.800656  0.708995
             8         9        10  mean precision  std precision
train  0.76499  0.767792  0.766673        0.769234       0.005313
test   0.77193  0.775325  0.785714        0.760355       0.043974
```

2.

Select any one dataset and for a particular training fold show the progression of the gradient descent algorithm by plotting the logistic loss for each iteration till convergence.

Gradient descent process of Spam dataset:

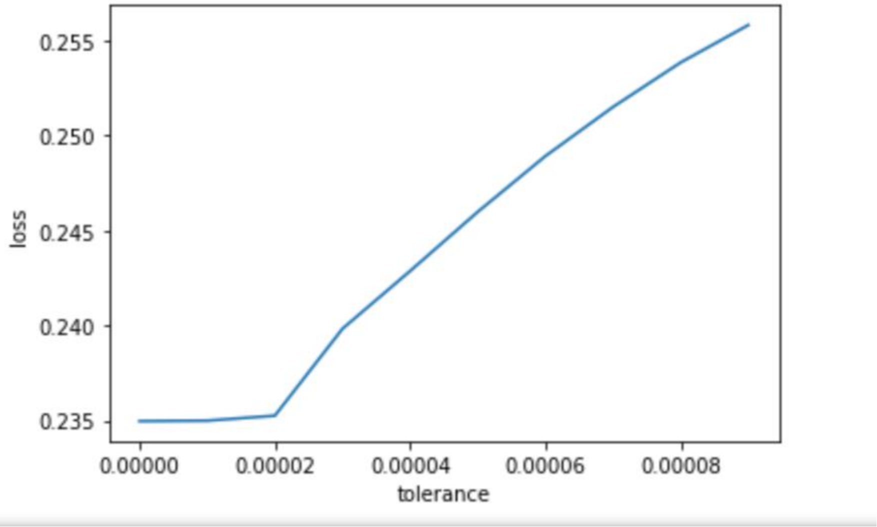


3. Explain how you chose the tolerance and maximum iterations in your implementation. If you tried different values of e, plot the training loss versus the epsilon values.

It’s better to set the maximum iterations lower than 1000 for all cases. And for different dataset, we select different epsilon values according to their loss values.

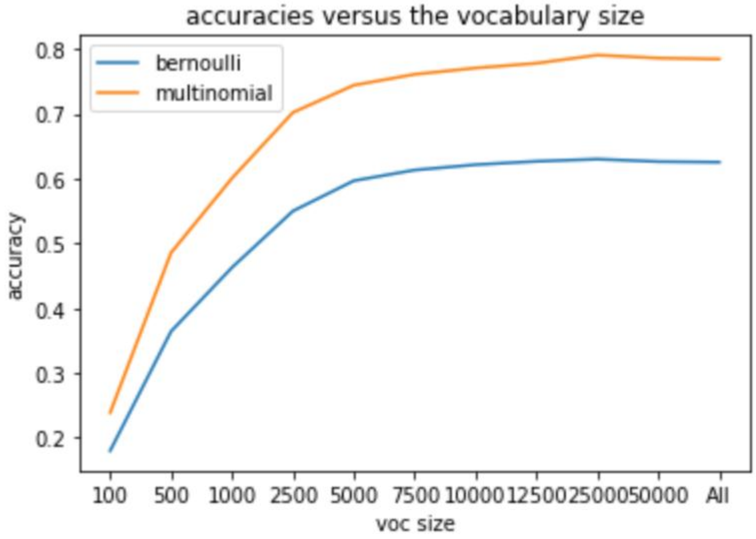
Epsilon: np.arange(1e-8, 1e-4, 1e-5)

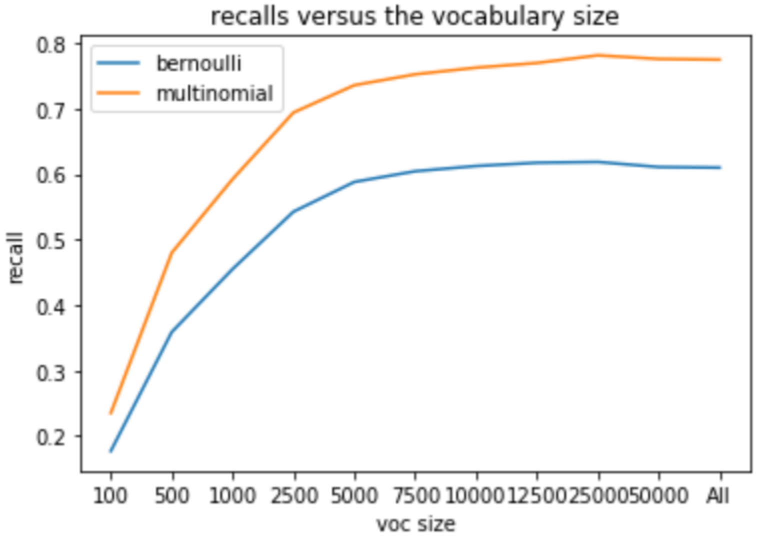
Spam dataset



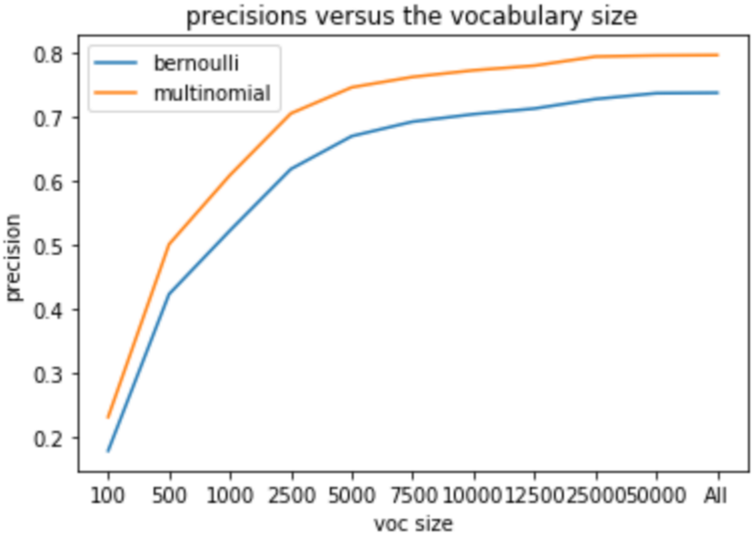
2.5

1. Plot the accuracy, recall and precision following metrics of the two models versus the vocabulary size. Create three plots for each performance metric.


accuracies versus the vocabulary size
bernoulli
multinomial
accuracy
0.8
0.7
0.6
0.5
0.4
0.3
0.2
100
500
1000
2500
5000
7500
10000
12500
25000
50000
All
voc size

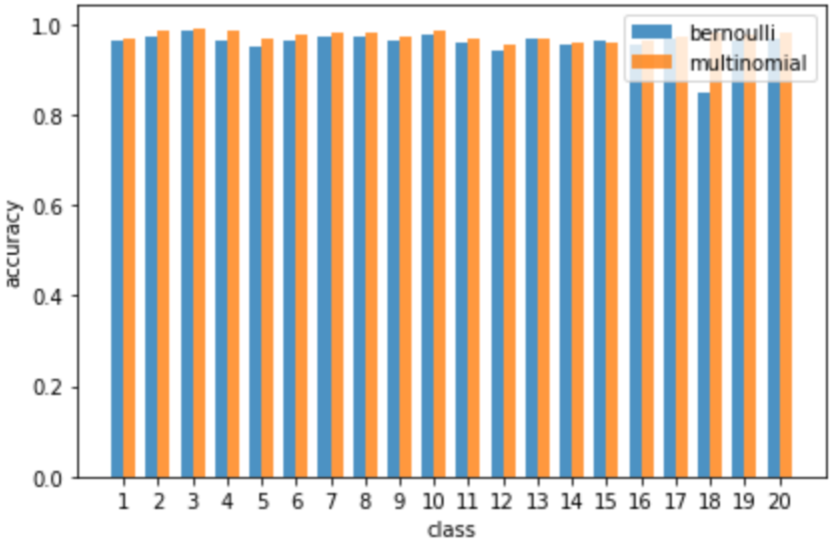

recalls versus the vocabulary size
bernoulli
multinomial
recall
0.8
0.7
0.6
0.5
0.4
0.3
0.2
100
500
1000
2500
5000
7500
10000
12500
25000
50000
All
voc size

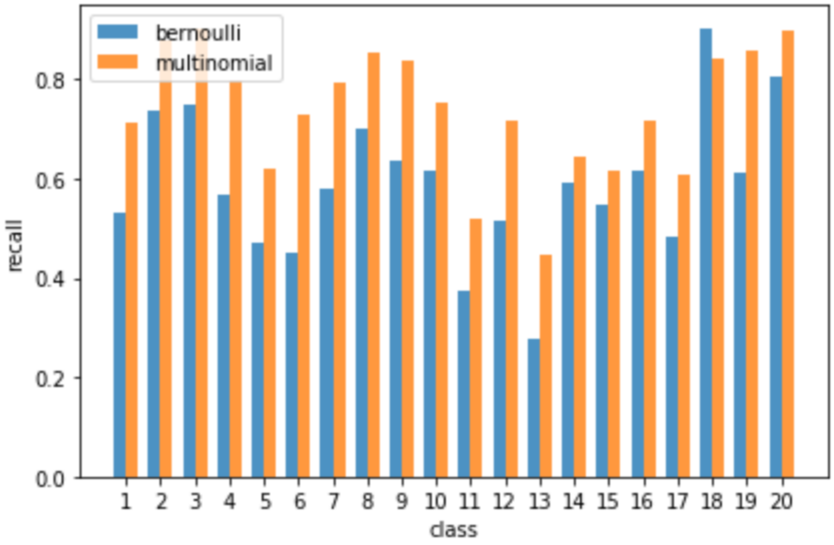
voc size

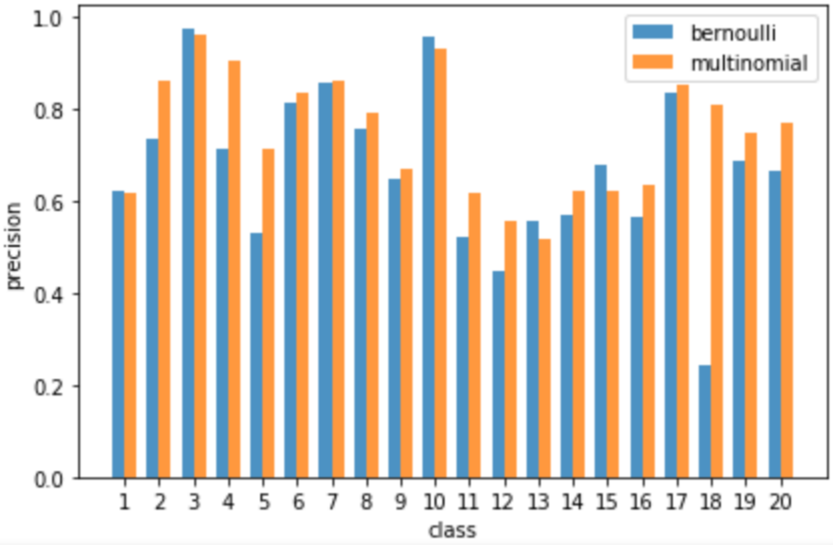


2. Create three grouped bar charts that contrast the accuracy, recall and precision of each class in the two models. A sample grouped bar chart is shown in Figure 1.

Voc size for following 3 graphs: 5000




bernoulli
multinomial
recall
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
class


bernoulli
multinomial
precision
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
class