Rajesh Sakhamuru
6/28/2020

CS6140 – Assignment 3 – Logistic Regression & Naive Bayes

1.3 - Logistic Regression Deliverables

Spambase Dataset:

Results Table:

	Fold #	Train Accuracy	Train Precision	Train Recall	Train Log-Loss	Test Accuracy	Test Precision	Test Recall	Test Log-Loss
0	1	0.926346	0.926376	0.883934	0.209763	0.928261	0.903955	0.909091	0.224714
1	2	0.924656	0.925854	0.879363	0.211758	0.928261	0.950920	0.861111	0.203707
2	3	0.924414	0.924697	0.881387	0.215282	0.945652	0.933735	0.917160	0.169919
3	4	0.925863	0.926974	0.877882	0.205683	0.917391	0.929293	0.884615	0.268313
4	5	0.927071	0.929569	0.884662	0.208717	0.902174	0.873333	0.834395	0.242330
5	6	0.927554	0.928389	0.885366	0.204582	0.919565	0.909639	0.872832	0.281934
6	7	0.928037	0.926876	0.887047	0.212621	0.932609	0.937143	0.891304	0.216123
7	8	0.926105	0.925470	0.882026	0.210741	0.923913	0.934426	0.881443	0.218184
8	9	0.928278	0.930968	0.883650	0.210785	0.919565	0.913295	0.877778	0.227300
9	10	0.927795	0.931507	0.880395	0.205091	0.917391	0.913514	0.884817	0.280316
10	Mean	0.926612	0.927668	0.882571	0.209502	0.923478	0.919925	0.881455	0.233284
11	Std Deviation	0.001368	0.002345	0.002867	0.003502	0.011448	0.022017	0.023270	0.035605

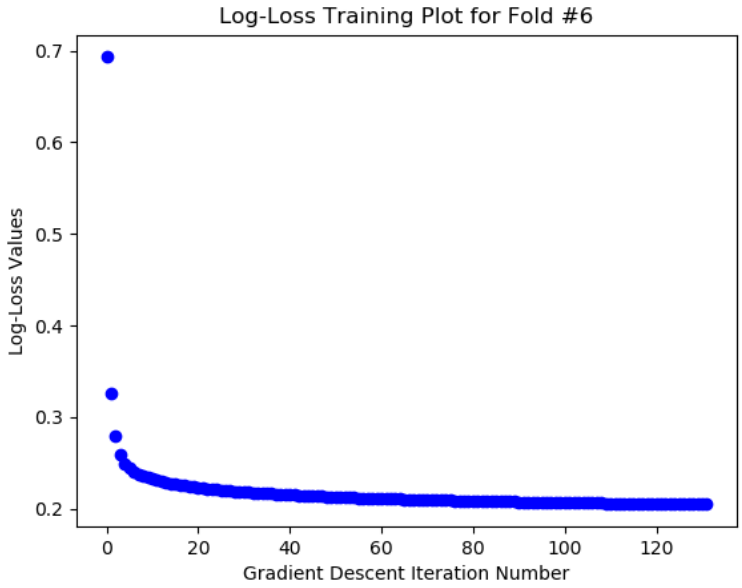
Columns are actual value, Rows are predicted value

Training Confusion Matrix:

	0	1
0	23969	1916
1	1123	14402

Testing Confusion Matrix:

	0	1
0	2650	214
1	138	1598



Breast Cancer Dataset:

Results Table:

	Fold #	Train Accuracy	Train Precision	Train Recall	Train Log-Loss	Test Accuracy	Test Precision	Test Recall	Test Log-Loss
0	1	0.986328	0.989418	0.973958	0.055813	1.000000	1.000000	1.000000	0.019315
1	2	0.992188	1.000000	0.979058	0.046202	0.947368	1.000000	0.857143	0.112315
2	3	0.988281	0.989305	0.978836	0.049124	0.982456	1.000000	0.956522	0.080132
3	4	0.986328	0.989130	0.973262	0.053831	0.982456	0.961538	1.000000	0.036905
4	5	0.982422	0.978836	0.973684	0.052815	0.982456	1.000000	0.954545	0.067035
5	6	0.988281	0.994681	0.973958	0.052720	0.982456	0.952381	1.000000	0.046828
6	7	0.988281	0.994792	0.974490	0.054317	0.982456	1.000000	0.937500	0.083382
7	8	0.988281	0.989529	0.979275	0.049711	0.964912	0.947368	0.947368	0.078132
8	9	0.986328	0.989362	0.973822	0.050782	0.982456	0.954545	1.000000	0.079369
9	10	0.988304	0.989189	0.978610	0.040813	0.982143	1.000000	0.960000	0.161585
10	Mean	0.987502	0.990424	0.975895	0.050613	0.978916	0.981583	0.961308	0.076500
11	Std Deviation	0.002471	0.005483	0.002647	0.004459	0.013830	0.024017	0.044310	0.040002

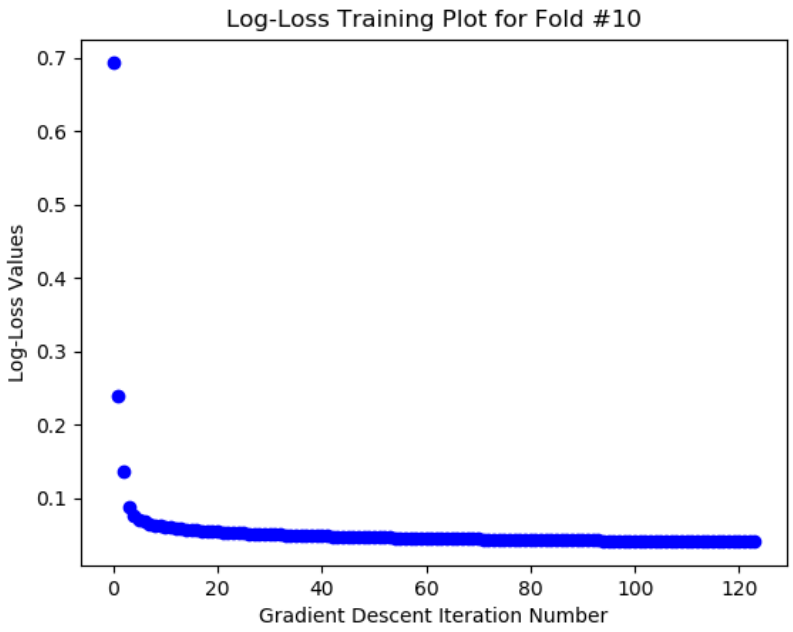
Columns are actual value, Rows are predicted value

Training Confusion Matrix:

	0	1
0	3195	46
1	18	1862

Testing Confusion Matrix:

	0	1
0	353	8
1	4	204



Diabetes Dataset:

Results Table:

	Fold #	Train Accuracy	Train Precision	Train Recall	Train Log-Loss	Test Accuracy	Test Precision	Test Recall	Test Log-Loss
0	1	0.788712	0.748634	0.578059	0.465777	0.766234	0.809524	0.548387	0.522673
1	2	0.774240	0.724868	0.568465	0.479435	0.844156	0.894737	0.629630	0.401479
2	3	0.774240	0.727273	0.564315	0.479266	0.831169	0.937500	0.555556	0.407770
3	4	0.780029	0.734375	0.582645	0.474592	0.792208	0.727273	0.615385	0.446541
4	5	0.784370	0.747423	0.591837	0.466504	0.740260	0.571429	0.521739	0.516104
5	6	0.797395	0.759358	0.599156	0.453935	0.649351	0.590909	0.419355	0.629090
6	7	0.782923	0.739796	0.594262	0.466597	0.766234	0.650000	0.541667	0.521680
7	8	0.781476	0.737113	0.588477	0.469969	0.779221	0.700000	0.560000	0.484012
8	9	0.780029	0.732620	0.573222	0.475198	0.766234	0.739130	0.586207	0.439485
9	10	0.780664	0.738220	0.580247	0.472507	0.813333	0.739130	0.680000	0.461257
10	Mean	0.782408	0.738968	0.582068	0.470378	0.774840	0.735963	0.565792	0.483009
11	Std Deviation	0.006808	0.010447	0.011428	0.007656	0.054717	0.119317	0.070240	0.067913

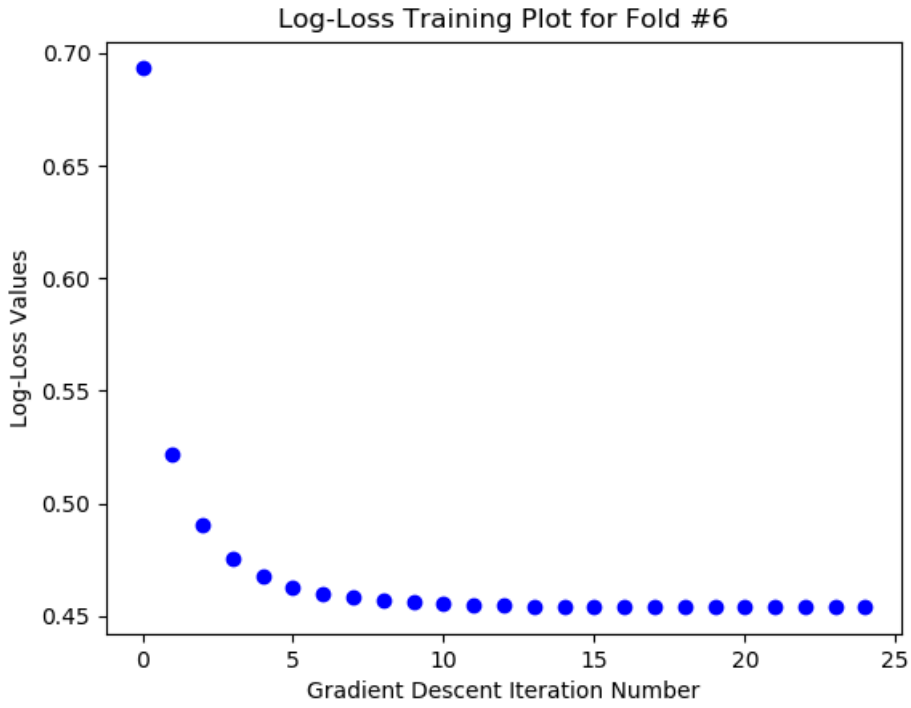
Columns are actual value, Rows are predicted value

Training Confusion Matrix:

	0	1
0	4004	1008
1	496	1404

Testing Confusion Matrix:

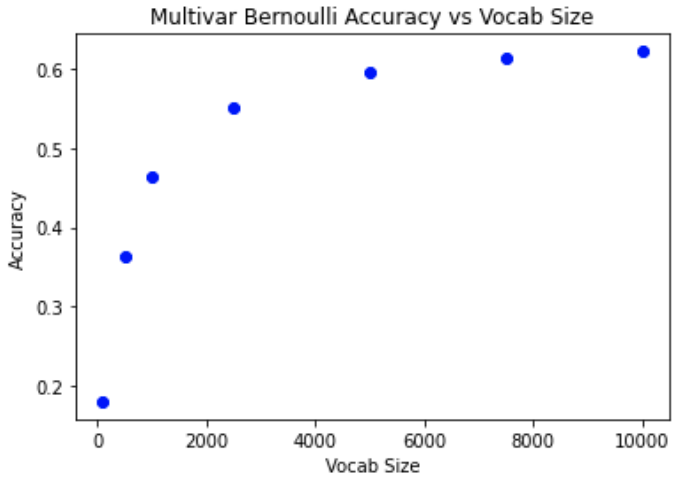
	0	1
0	444	117
1	56	151

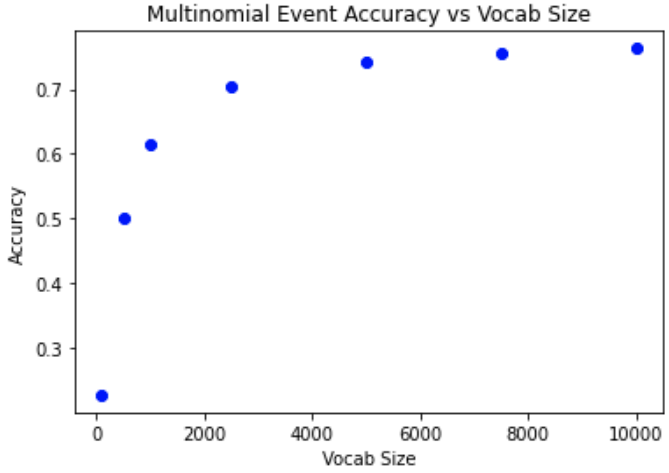


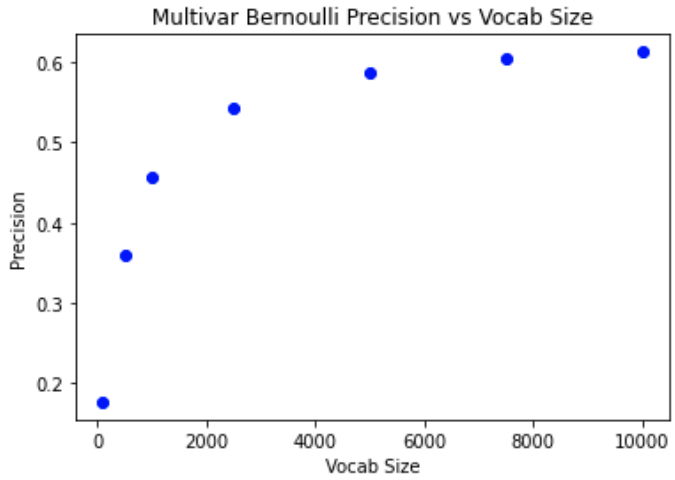
1.3.3 Tolerance and Max iterations:
I used the same tolerance value of .00005 for all three of the datasets and a maximum iterations of 1000. What I varied among the datasets is the learning rate, where I make sure that the learning rate is high enough that it always stops before reaching the maximum iterations (usually before even 300 iterations), but also doesn't make jumps that increase the error. The tolerance value is small enough that any further iterations will have diminishing returns at the point that it stops.

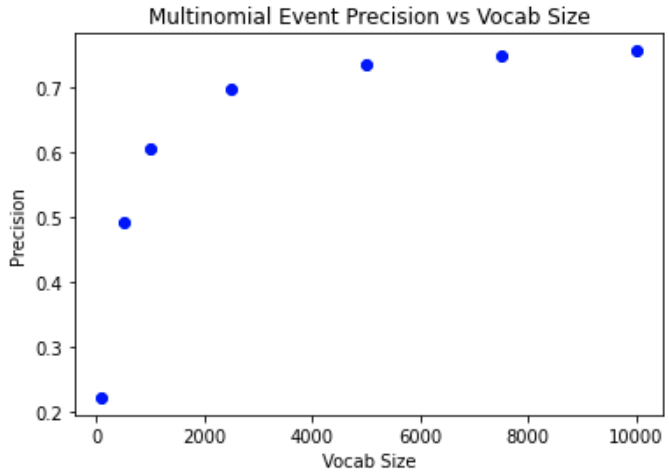
2 - Naive Bayes for Document Classification

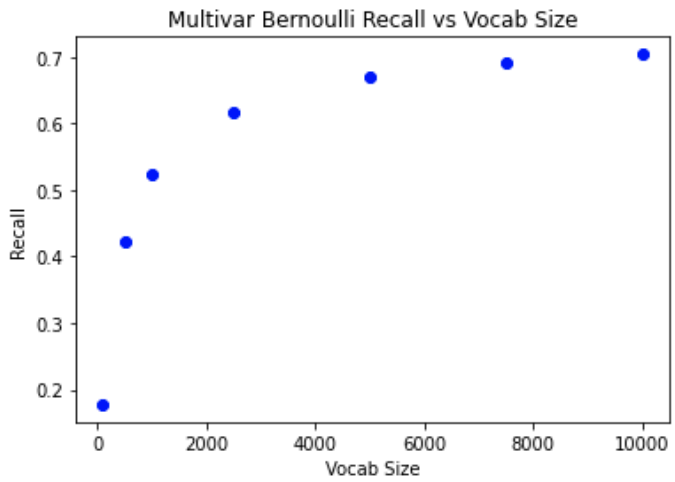
2.5 – Deliverables:
1: Accuracy, Recall, Precision vs Vocabulary Size

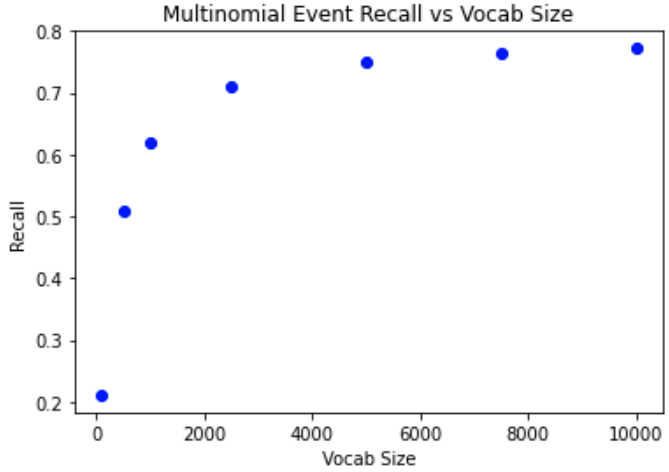








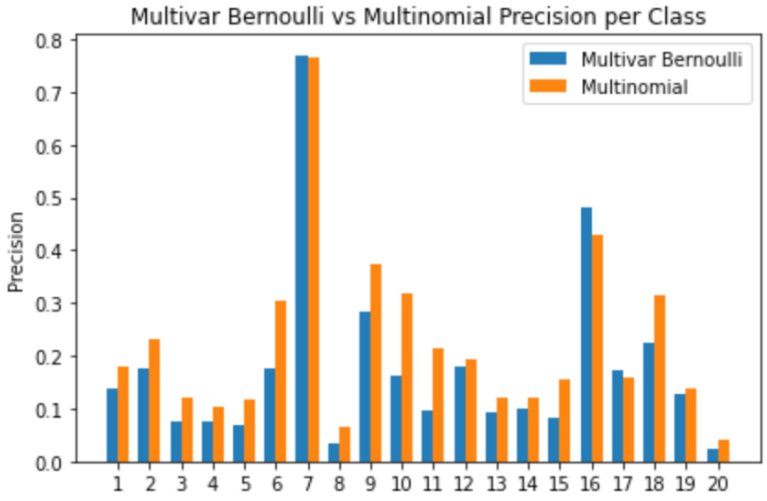


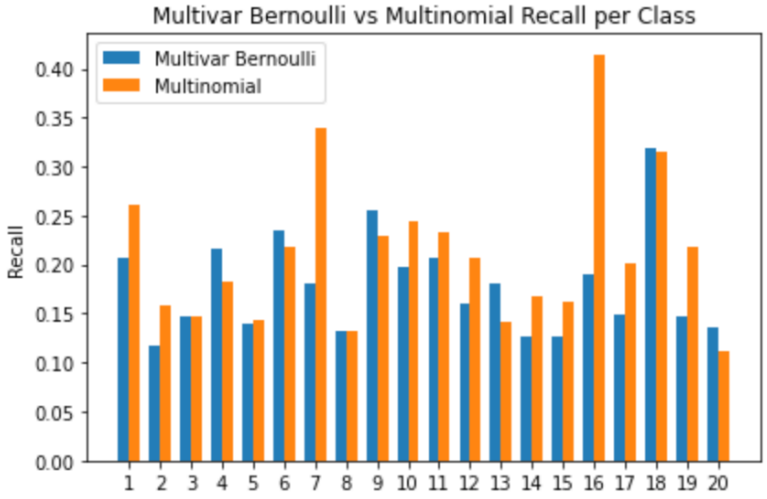


2: Grouped Bar Charts by Class
I only included graphs for Precision and Recall (and not Accuracy) at each vocabulary size level, because Accuracy per Class is the same as Precision per Class. So the Accuracy per Vocab Size is reported instead above the related graphs.
Vocab Sizes greater than 10000 are not included because it takes too long to run, but also because increasing the vocab size past 10000 has diminishing accuracy, precision and recall improvements for a far greater time penalty.

Vocab Size = 100

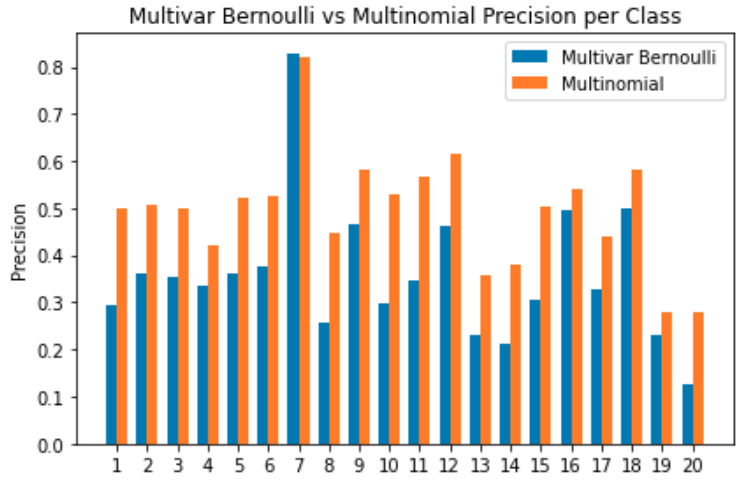
```
Multivar Bernoulli Accuracy: 0.1798800799467022
Multivar Bernoulli Avg Precision: 0.17692885354898497
Multivar Bernoulli Avg Recall: 0.178444603906789
Multinomial Accuracy: 0.22758161225849433
Multinomial Avg Precision: 0.22339743526250921
Multinomial Avg Recall: 0.21144400072285624
```

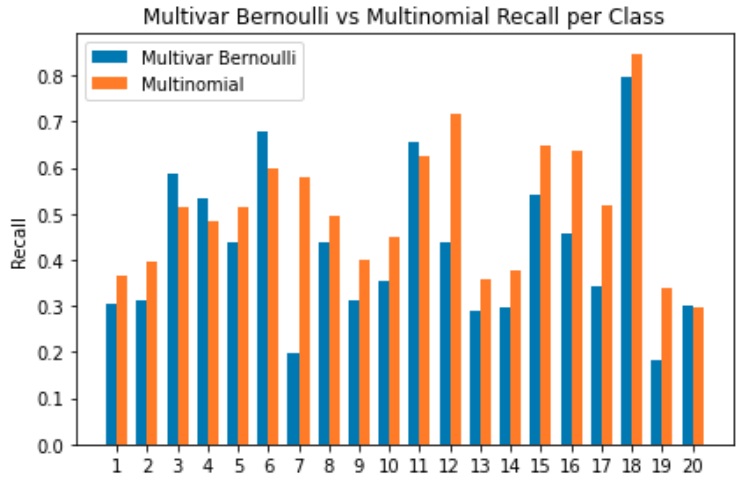




Vocab Size = 500

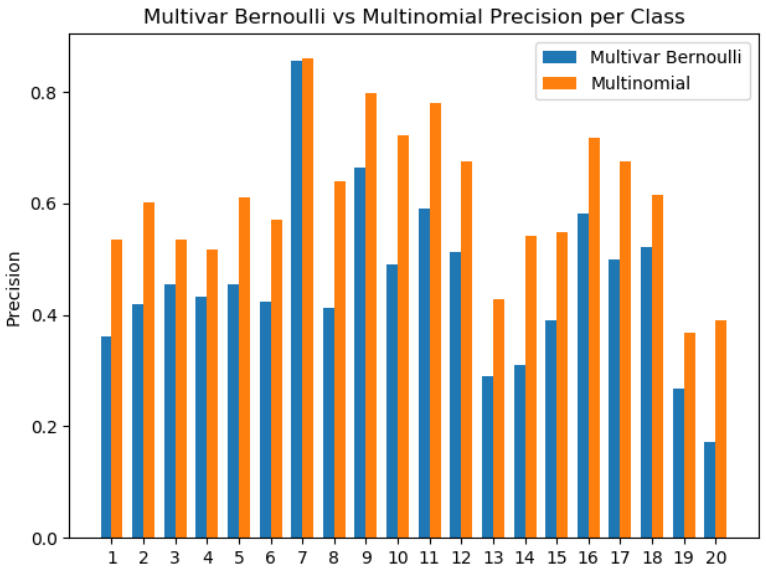
```
Multivar Bernoulli Accuracy: 0.364290473017988
Multivar Bernoulli Avg Precision: 0.35868057276327747
Multivar Bernoulli Avg Recall: 0.4228893873916669
Multinomial Accuracy: 0.5003331112591606
Multinomial Avg Precision: 0.49433108516980073
Multinomial Avg Recall: 0.5082587646508769
```

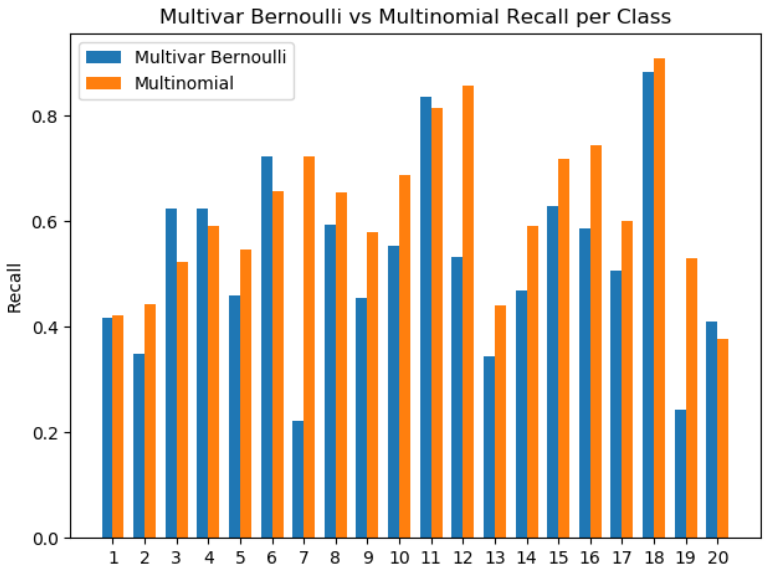




Vocab Size = 1000

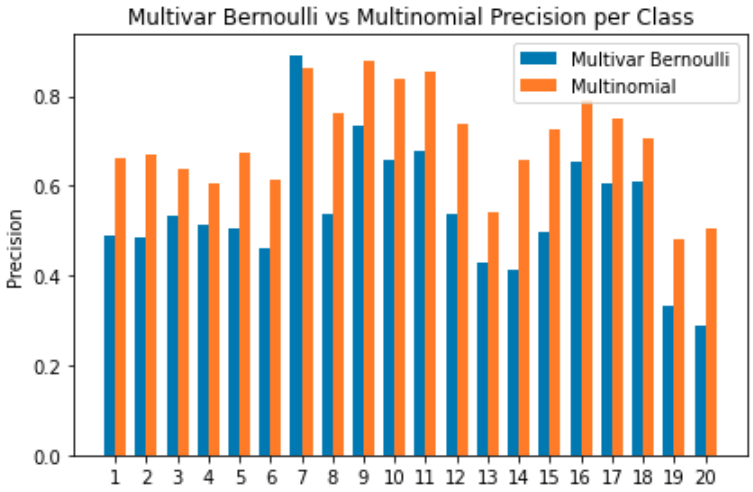
```
Multivar Bernoulli Accuracy: 0.46342438374417055
Multivar Bernoulli Avg Precision: 0.45556379426400395
Multivar Bernoulli Avg Recall: 0.5232225570818606
Multinomial Accuracy: 0.6139906728847435
Multinomial Avg Precision: 0.6067372020610088
Multinomial Avg Recall: 0.6204023502961774
```

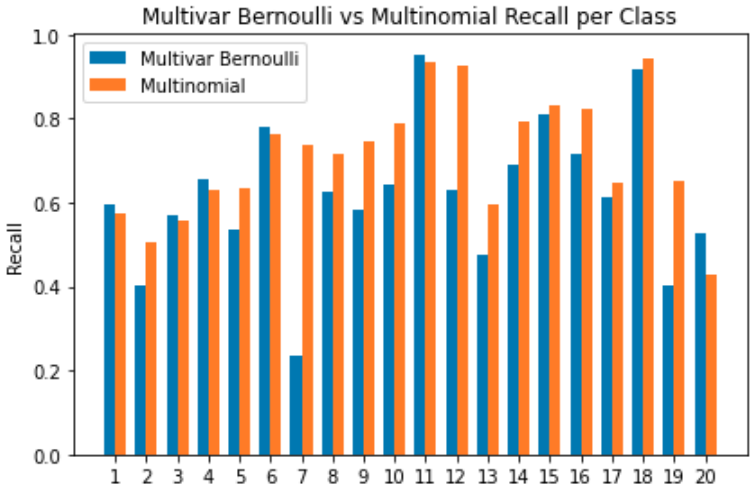




Vocab Size = 2500

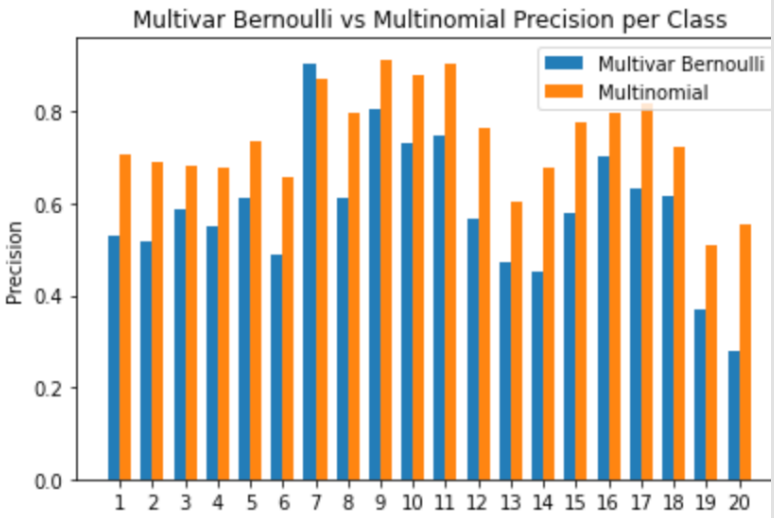
```
Multivar Bernoulli Accuracy: 0.5499000666222519
Multivar Bernoulli Avg Precision: 0.5426278196347314
Multivar Bernoulli Avg Recall: 0.6178261918673005
Multinomial Accuracy: 0.7045969353764158
Multinomial Avg Precision: 0.6981797350458281
Multinomial Avg Recall: 0.7112729486196532
```

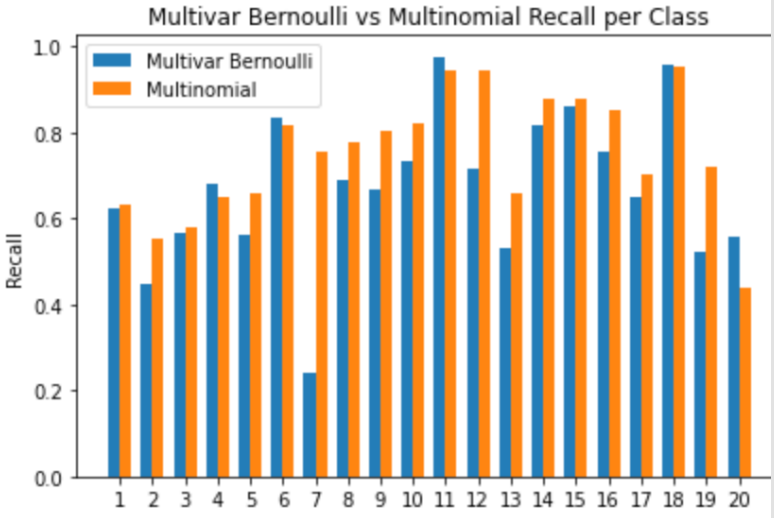




Vocab Size = 5000

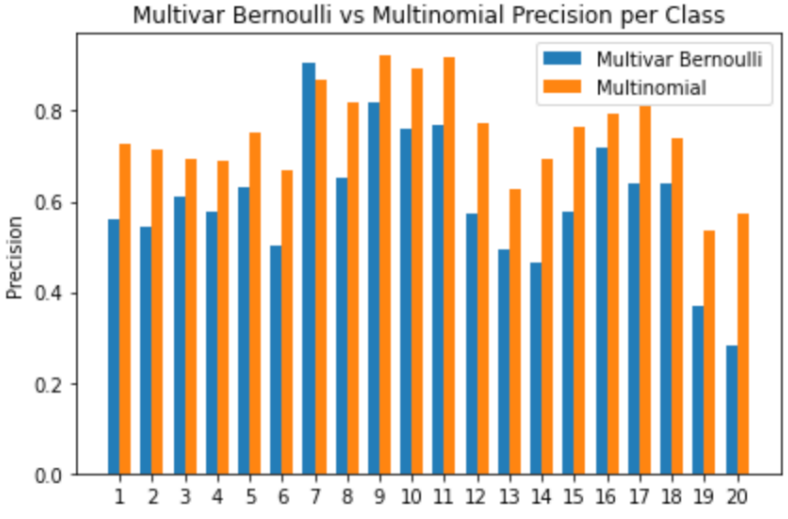
```
Multivar Bernoulli Accuracy: 0.5964023984010659
Multivar Bernoulli Avg Precision: 0.5878555273364823
Multivar Bernoulli Avg Recall: 0.6690816396033749
Multinomial Accuracy: 0.7431045969353764
Multinomial Avg Precision: 0.7368608432222242
Multinomial Avg Recall: 0.7506903269309322
```

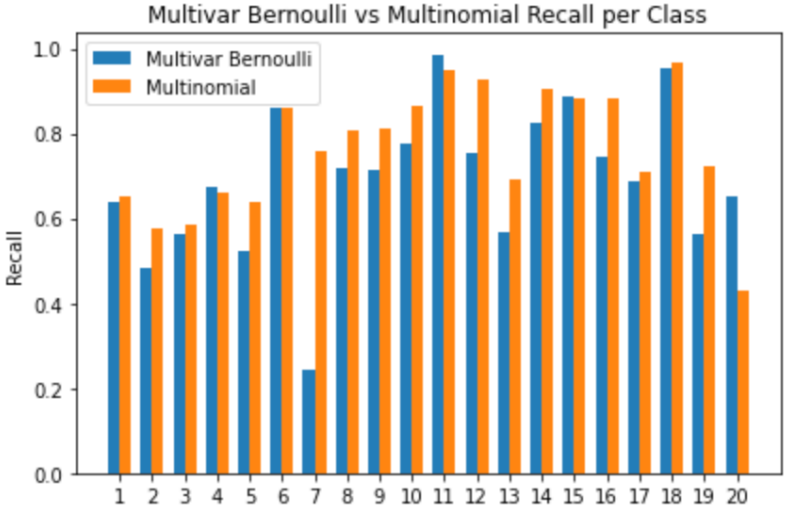




Vocab Size = 7500

```
Multivar Bernoulli Accuracy: 0.6129247168554297
Multivar Bernoulli Avg Precision: 0.6040342664650333
Multivar Bernoulli Avg Recall: 0.6908893772399785
Multinomial Accuracy: 0.7550966022651565
Multinomial Avg Precision: 0.7491520690372635
Multinomial Avg Recall: 0.7642314717595431
```





Vocab Size = 10000

```
Multivar Bernoulli Accuracy: 0.6222518321119254
Multivar Bernoulli Avg Precision: 0.6131606611636724
Multivar Bernoulli Avg Recall: 0.7041274431458724
Multinomial Accuracy: 0.7629580279813458
Multinomial Avg Precision: 0.757013543012735
Multinomial Avg Recall: 0.7722063178349752
```