1. Perceptron train loss = 0.029344175358

Perceptron test loss = 0.387546468401

MultinomialNB train loss = 0.13196040304

MultinomialNB test loss = 0.350770047796

SGDClassifier train loss = 0.0540038889871

SGDClassifier test loss = 0.426181625066

BernoulliNB baseline train loss = 0.401272759413

BernoulliNB baseline test loss = 0.542087095061

For perceptron, I used 10-fold cross validation to pick the optimal max-iteration value and check whether this method is better than Bernoulli Naïve Bayes.

For SGDClassifier, I used 10-fold cross validation to choose the optimal penalty function and check whether this method is better than Bernoulli Naïve Bayes.

For MultinomialNB, I used 10-fold cross validation to choose the optimal alpha value and check whether this method is better than Bernoulli Naïve Bayes.

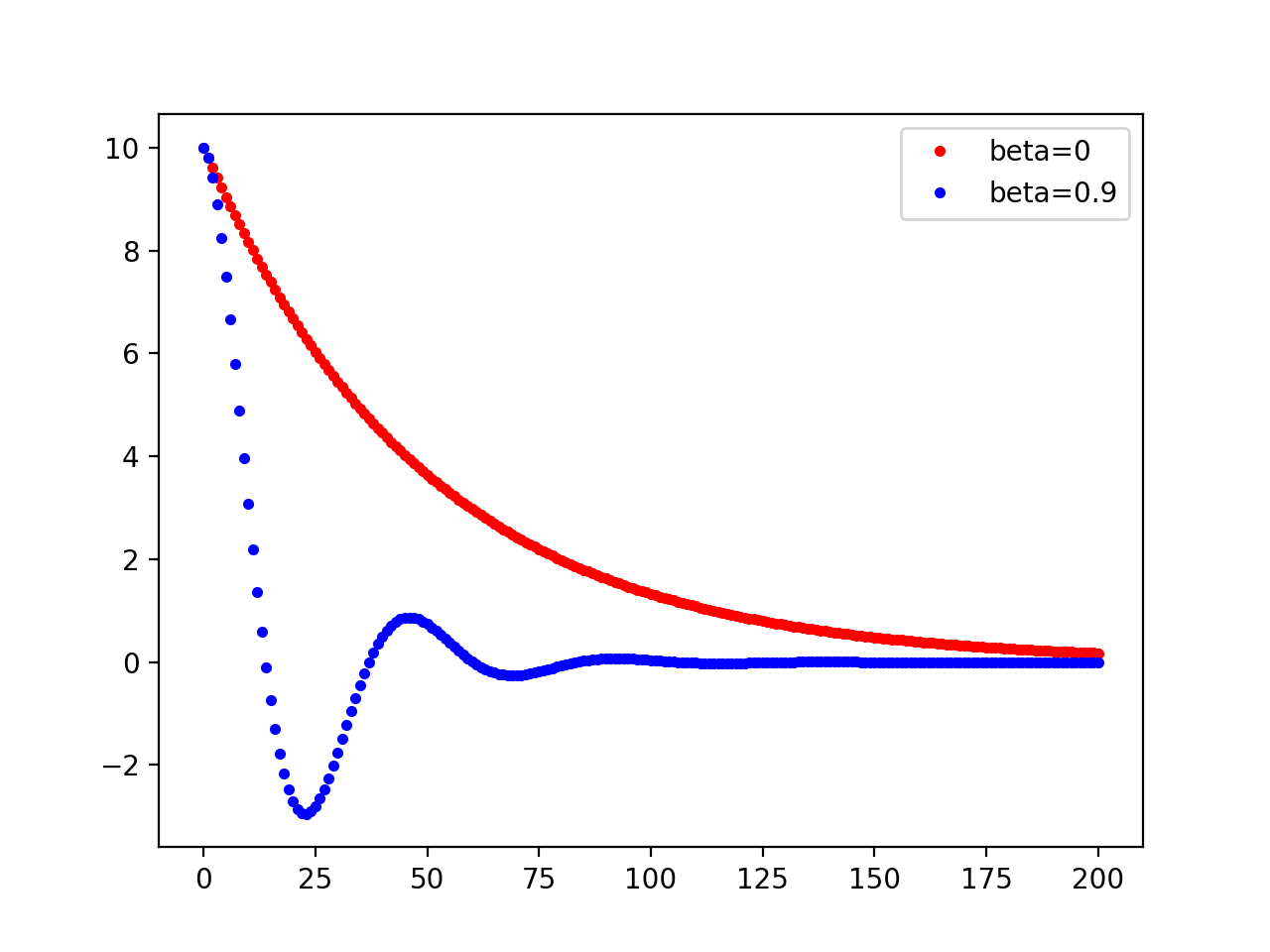
I first picked MultinomialNB since it is suitable for classification with discrete features. Then I took Perceptron as an attempt and it outperformed the baseline. Because SGDClassifier is also a linear model, I took it as the final attempt. Finally, all three models performed better than baseline and as expected, MultinomialNB generated minimal loss among them.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 172 | 4 | 6 | 0 | 1 | 0 | 0 | 1 | 5 | 8 | 6 | 4 | 2 | 7 | 11 | 22 | 15 | 21 | 26 | 50 |
| 1 | 279 | 77 | 11 | 14 | 57 | 2 | 2 | 2 | 3 | 2 | 7 | 20 | 5 | 13 | 5 | 0 | 1 | 2 | 3 |
| 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 16 | 136 | 280 | 30 | 10 | 29 | 2 | 1 | 0 | 0 | 2 | 29 | 3 | 1 | 1 | 0 | 1 | 0 | 1 |
| 0 | 22 | 44 | 50 | 273 | 12 | 29 | 1 | 2 | 1 | 0 | 6 | 18 | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| 2 | 19 | 62 | 5 | 1 | 285 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 3 | 0 | 0 |
| 0 | 2 | 2 | 7 | 6 | 3 | 272 | 8 | 3 | 3 | 0 | 0 | 5 | 2 | 0 | 0 | 2 | 0 | 1 | 0 |
| 5 | 2 | 2 | 3 | 7 | 0 | 11 | 293 | 32 | 4 | 3 | 0 | 20 | 7 | 10 | 1 | 5 | 2 | 6 | 1 |
| 3 | 3 | 4 | 0 | 3 | 3 | 8 | 27 | 295 | 6 | 6 | 6 | 5 | 7 | 6 | 0 | 4 | 3 | 5 | 3 |
| 2 | 1 | 0 | 1 | 0 | 0 | 2 | 1 | 4 | 314 | 23 | 3 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 2 |
| 9 | 5 | 15 | 7 | 14 | 5 | 10 | 25 | 13 | 20 | 328 | 16 | 11 | 14 | 18 | 14 | 11 | 6 | 7 | 7 |
| 2 | 15 | 9 | 4 | 6 | 6 | 1 | 2 | 1 | 4 | 3 | 285 | 33 | 1 | 1 | 1 | 6 | 5 | 3 | 2 |
| 1 | 5 | 5 | 21 | 19 | 2 | 6 | 10 | 10 | 2 | 1 | 8 | 222 | 7 | 6 | 0 | 3 | 0 | 1 | 4 |
| 2 | 3 | 4 | 1 | 1 | 4 | 1 | 1 | 4 | 3 | 1 | 3 | 8 | 300 | 6 | 0 | 3 | 0 | 6 | 3 |
| 6 | 6 | 9 | 0 | 4 | 1 | 8 | 3 | 0 | 2 | 2 | 7 | 10 | 4 | 277 | 1 | 3 | 0 | 6 | 4 |
| 54 | 4 | 3 | 0 | 2 | 2 | 3 | 4 | 6 | 10 | 5 | 7 | 3 | 13 | 9 | 330 | 15 | 25 | 7 | 78 |
| 9 | 0 | 0 | 1 | 1 | 1 | 3 | 4 | 9 | 6 | 1 | 15 | 2 | 4 | 2 | 2 | 226 | 8 | 84 | 17 |
| 9 | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 1 | 4 | 7 | 1 | 5 | 6 | 1 | 15 | 269 | 13 | 11 |
| 13 | 2 | 11 | 1 | 1 | 2 | 3 | 9 | 6 | 9 | 12 | 17 | 3 | 12 | 24 | 2 | 34 | 25 | 133 | 9 |
| 28 | 0 | 4 | 0 | 1 | 1 | 1 | 1 | 2 | 0 | 2 | 2 | 0 | 3 | 3 | 16 | 21 | 6 | 10 | 56 |

Following is the confusion matrix:

By calculating the proportions that test examples belonging to class j were classified as class i, I built another table and observed that class 0 and 19 are the most confused class, which are alt.atheism and talk.religion.misc, respectively.

1. The test of SGD with momentum:



For model with beta=0:

Training accuracy is 0.911111111111.

Test accuracy is 0.911498005078.

Average training hinge loss is 0.404885244108.

Average test hinge loss is 0.408611242517.



For the model with beta = 0.1:

Training accuracy is 0.903219954649.

Test accuracy is 0.904243743199.

Average training hinge loss is 0.352736680825.

Average test hinge loss is 0.340984757148.

