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.

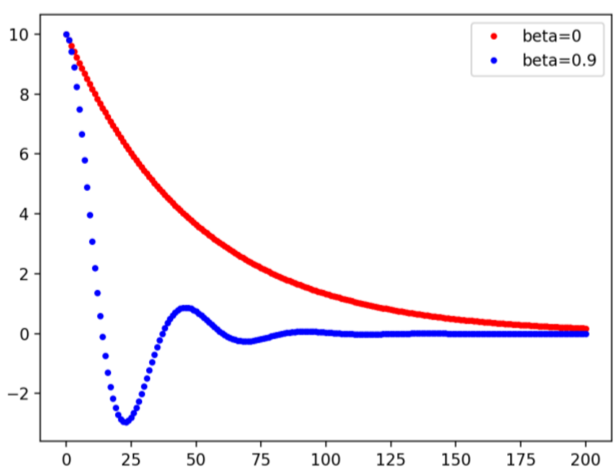
   I first picked MultinomialNB since it is suitable for classification with discrete features. Then I took Perceptron as an attempt and it outperformed the baseline. Because SGDClassifier is also a linear model, I took it as the final attempt. Finally, all three models performed better than baseline and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

Following is the confusion matrix:

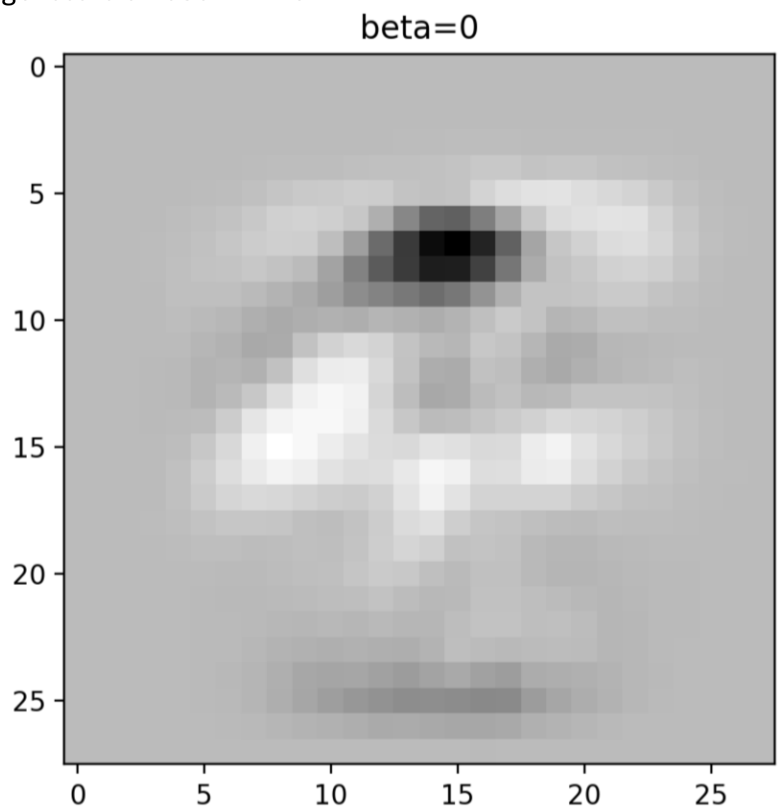
172	4	6	0	1	0	0	1	5	8	6	4	2	7	11	22	15	21	26	50
1	279	77	11	14	57	2	2	2	3	2	7	20	5	13	5	0	1	2	3
0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	16	136	280	30	10	29	2	1	0	0	2	29	3	1	1	0	1	0	1
0	22	44	50	273	12	29	1	2	1	0	6	18	2	0	1	0	0	0	0
2	19	62	5	1	285	0	0	1	1	0	1	0	0	1	0	1	3	0	0
0	2	2	7	6	3	272	8	3	3	0	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

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.

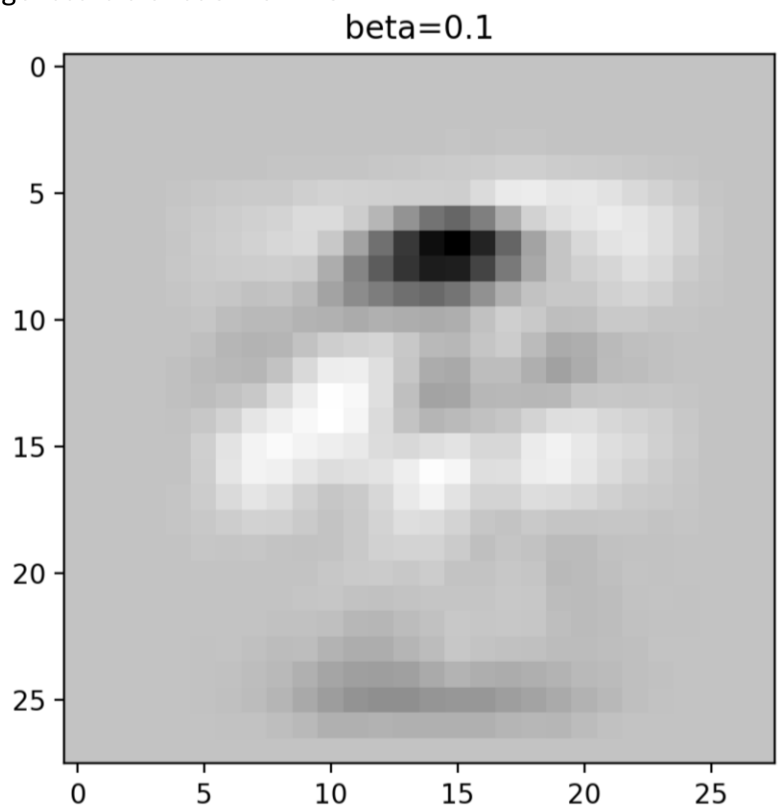
2. The test of SGD with momentum:



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.



3.1 Solution:

For K, let A be a matrix built by eigenvectors of K.

Since K is systemmetic, then we can set A to be orthonormal, i.s

$$A \cdot A^T = I \Rightarrow A^T = A^{-1}$$

and $A^{-1} K A = diag(\lambda)$, where λ is a vector of K's eigenvalues.

We can see $K = A\, diag(\lambda) A^{-1} = A\, diag(\lambda) A^T$

For any $x \in R^d$,

$$x^T K x = x^T A\, diag(\lambda) A^T x$$
$$= \sum_{i=1}^{d} \lambda_i \cdot [A^T x]_i^2$$

That is, $x^T K x \geq 0$ IFF $\lambda_i \geq 0$ for $1 \leq i \leq d$

Thus, K is positive semi-definite IFF $x^T K x \geq 0$ for all $x \in R^d$.

3.2 Solution:

1). $k(x,y) = a \Rightarrow \langle \phi(x), \phi(y) \rangle = a$

Then let embedding $\phi(x) = \begin{bmatrix} \sqrt{a} \\ \vdots \\ 0 \end{bmatrix}$, then

$\langle \phi(x), \phi(y) \rangle = a$ for $a > 0$.

Thus, $k(x,y) = a$ is a kernel for $a > 0$.

2). $k(x,y) = f(x) \cdot f(y) \Rightarrow \langle \phi(x), \phi(y) \rangle = f(x) \cdot f(y)$

For any $f: R^d \to R$, let $\phi(x) = [f(x), 0, \dots, 0]$, then

$\langle \phi(x), \phi(y) \rangle = f(x) \cdot f(y)$

Thus, $k(x,y) = f(x) \cdot f(y)$ for all $f: R^d \to R$

3). $k(x,y) = a \cdot k_1(x,y) + b \cdot k_2(x,y)$

$$K_{ij} = k(x^{(i)}, x^{(j)}) = a \cdot k_1(x^{(i)}, x^{(j)}) + b \cdot k_2(x^{(i)}, x^{(j)})$$

$$= a \cdot K_{ij}^{(1)} + b \cdot K_{ij}^{(2)}$$

and $K = a \cdot K^{(1)} + b \cdot K^{(2)}$

For any $x \in \mathbb{R}^d$, $x^T K x = x^T (aK^{(1)} + bK^{(2)}) x$

$$= a \cdot x^T K^{(1)} x + b x^T K^{(2)} x$$

$$\geq 0 \qquad \# \; x^T K^{(1)} x, \; x^T K^{(2)} x \geq 0, \; a, b > 0$$

We can see $K_{ij} = K_{ji}$ and thus K is positive semi-definite.

Therefore $k(x,y)$ is a kernel.

4). $$k(x,y) = \frac{k_1(x,y)}{\sqrt{k_1(x,x)} \cdot \sqrt{k_1(y,y)}} = \frac{\langle \phi(x), \phi(y) \rangle}{\sqrt{\langle \phi(x), \phi(x) \rangle} \cdot \sqrt{\langle \phi(y), \phi(y) \rangle}}$$

$$= \frac{\langle \phi(x), \phi(y) \rangle}{\|\phi(x)\|_2 \cdot \|\phi(y)\|_2}$$

$$= \left\langle \frac{\phi(x)}{\|\phi(x)\|_2}, \frac{\phi(y)}{\|\phi(y)\|_2} \right\rangle$$

Then $k(x,y) = \langle \varphi(x), \varphi(y) \rangle$ where $\varphi(u) = \frac{\phi(u)}{\|\phi(u)\|_2}$

and $k(x,y)$ is a kernel.