Answer to Question 1:

I tried out many different algorithms, including decision tree, PCA, conditional Gaussian, multinomial naive bayes, support vector machine and neural network. The best 3 among them are:
neural network, multinomial naive bayes and support vector machine.

BernoulliNB baseline train accuracy = 0.5987272405868835
BernoulliNB baseline test accuracy = 0.4579129049389272
neural network train accuracy = 0.9635849390136114
neural network test accuracy = 0.5967870419543282
multinomial naive bayes train accuracy = 0.8113841258617642
multinomial naive bayes test accuracy = 0.6062134891131173
support vector machine train accuracy = 0.9627894643804137
support vector machine test accuracy = 0.6846787041954329

the confusion matrix for SVM on training set is:
[20, 14, 12, 8, 16, 14, 19, 10, 17, 33, 12, 21, 17, 16, 10, 16, 14, 19, 12, 19]
[14, 19, 17, 21, 14, 23, 20, 22, 24, 38, 20, 19, 15, 20, 20, 20, 18, 19, 19, 7]
[17, 22, 18, 15, 21, 19, 14, 22, 25, 24, 13, 17, 19, 30, 25, 27, 20, 14, 22, 10]
[18, 14, 24, 27, 21, 21, 17, 16, 17, 36, 20, 15, 18, 22, 16, 24, 22, 13, 19, 12]
[13, 15, 21, 26, 17, 20, 15, 15, 23, 32, 14, 19, 24, 21, 22, 20, 12, 28, 21, 7]
[16, 21, 18, 23, 22, 23, 17, 15, 13, 25, 27, 23, 16, 28, 17, 21, 24, 20, 13, 13]
[17, 22, 20, 14, 21, 27, 26, 14, 17, 26, 27, 10, 28, 21, 23, 25, 18, 18, 11, 5]
[19, 18, 21, 23, 25, 15, 26, 21, 32, 27, 26, 14, 16, 18, 21, 23, 17, 14, 12, 8]
[14, 21, 13, 25, 22, 16, 25, 18, 21, 36, 24, 18, 29, 15, 19, 16, 20, 16, 17, 13]
[15, 24, 25, 16, 18, 27, 17, 19, 17, 33, 16, 22, 11, 17, 17, 23, 28, 25, 19, 8]
[17, 24, 19, 23, 20, 19, 24, 21, 20, 32, 17, 17, 20, 15, 19, 25, 19, 22, 12, 14]
[9, 15, 12, 26, 25, 22, 28, 20, 21, 29, 19, 21, 19, 21, 16, 28, 19, 20, 17, 9]
[14, 20, 22, 25, 18, 22, 25, 21, 22, 28, 21, 21, 13, 17, 15, 25, 14, 22, 15, 13]
[23, 19, 21, 26, 23, 25, 18, 19, 17, 37, 18, 15, 18, 22, 17, 16, 24, 21, 6, 11]
[14, 20, 15, 17, 25, 26, 19, 16, 19, 33, 17, 21, 22, 21, 18, 19, 19, 24, 21, 8]
[9, 23, 26, 21, 16, 25, 26, 19, 20, 31, 20, 25, 21, 21, 18, 22, 18, 18, 10, 9]
[18, 19, 14, 20, 22, 17, 14, 18, 20, 38, 26, 24, 21, 11, 16, 20, 16, 12, 9, 9]
[14, 26, 20, 17, 12, 20, 19, 13, 24, 26, 17, 23, 21, 25, 21, 21, 15, 15, 14, 13]
[12, 14, 19, 14, 10, 12, 17, 20, 16, 19, 11, 24, 24, 20, 18, 11, 12, 16, 10, 11]
[7, 11, 14, 11, 10, 16, 10, 17, 11, 19, 14, 14, 18, 14, 15, 12, 9, 14, 9, 6]

the confusion matrix for SVM on test set is:

[130, 1, 1, 1, 0, 2, 3, 0, 6, 13, 6, 5, 3, 12, 17, 82, 7, 20, 4, 6]

[1, 252, 29, 12, 9, 24, 9, 2, 3, 9, 0, 15, 3, 3, 11, 3, 2, 2, 0, 0]

[3, 14, 253, 31, 13, 17, 1, 1, 2, 15, 1, 6, 2, 8, 11, 4, 3, 2, 5, 2]

[0, 14, 38, 242, 25, 5, 20, 3, 0, 9, 3, 10, 19, 0, 1, 0, 0, 2, 1, 0]

[0, 7, 7, 31, 264, 5, 21, 2, 3, 14, 4, 5, 8, 6, 6, 0, 1, 0, 0, 1]

[1, 32, 29, 5, 5, 292, 5, 0, 0, 5, 0, 6, 5, 3, 4, 1, 0, 1, 1, 0]

[0, 2, 0, 14, 8, 0, 330, 5, 2, 10, 2, 2, 4, 1, 4, 1, 2, 2, 1, 0]

[1, 1, 3, 0, 2, 2, 15, 273, 18, 29, 3, 5, 17, 2, 4, 2, 6, 9, 3, 1]

[1, 0, 2, 3, 2, 0, 13, 17, 288, 20, 5, 4, 5, 9, 9, 5, 7, 5, 3, 0]

[0, 2, 3, 0, 1, 2, 8, 1, 3, 329, 34, 2, 1, 1, 0, 6, 0, 0, 4, 0]

[0, 1, 1, 0, 0, 1, 0, 1, 1, 14, 366, 1, 0, 2, 2, 5, 3, 1, 0, 0]

[1, 6, 4, 2, 4, 5, 1, 2, 1, 20, 1, 302, 4, 6, 5, 5, 13, 7, 5, 2]

[1, 12, 8, 23, 8, 11, 21, 10, 10, 15, 5, 43, 190, 17, 15, 1, 2, 1, 0, 0]

[4, 5, 1, 0, 2, 3, 9, 2, 3, 14, 7, 4, 3, 312, 2, 8, 3, 6, 7, 1]

[1, 4, 3, 0, 1, 2, 5, 5, 4, 20, 2, 2, 6, 12, 305, 6, 4, 7, 4, 1]

[12, 4, 2, 1, 0, 0, 1, 1, 0, 15, 2, 1, 3, 8, 2, 334, 0, 6, 3, 3]

[3, 2, 2, 0, 1, 3, 4, 3, 4, 16, 2, 15, 3, 8, 7, 13, 250, 14, 10, 4]

[10, 2, 1, 0, 0, 2, 1, 1, 5, 12, 2, 7, 2, 3, 1, 10, 6, 305, 5, 1]

[9, 0, 1, 1, 1, 1, 0, 5, 3, 9, 9, 14, 3, 8, 8, 10, 88, 24, 113, 3]

[34, 4, 2, 0, 1, 1, 5, 4, 2, 7, 7, 1, 0, 13, 5, 89, 27, 14, 8, 27]

from above, we can see that 89 is the greatest number in the test-set confusion matrix ignoring the diagonal. So, the classifier is likely to mis-class examples to class 20 (talk.religion.misc) whose class is actually 16 (soc.religion.christian). That make sense, because they are the most similar topic in this 20newsgroup, even human being like me feel confused about them!

I use for-loop validation, cross validation and grid search (which choose hyperparameters implicitly) to search for the best hyperparameters. I initialize the hyperparameters with offical website's hyperparameters and with respect to the consideration of my dataset's size and feature (for example, there are 20 classes, so the number of my neural network's internal node should be around 20)

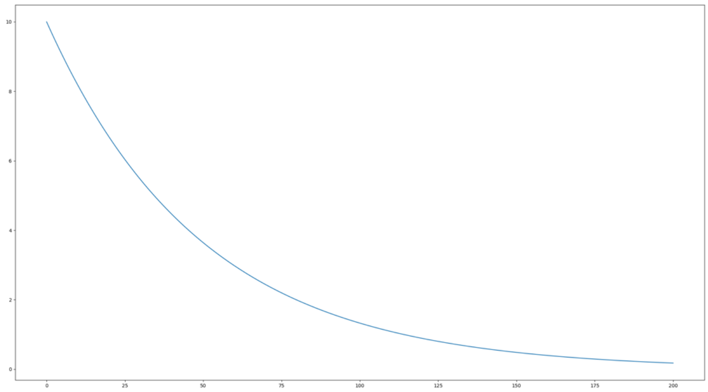
I tried plenty of methods taught in class, and picked those with the best test accuracy.
Binomial naive bayes classifier works the worst, since it makes a naive assumption, which is highly likely to be false in our dataset.
Neural network works just fine. If I have more processor and more time, I would try to adjust those hyperparameters more freely but with current resources, I think 45 layers with 15 internal nodes produces acceptable result. Neural network works well when chose the best hyperparameters, and it will generally works just as expected with generally good hyperparameters.
SVM gives the best result, since it is the common classifier for NLP classification problem, which is expected to give the best result for this 20newsgroup classification

problem. I tried grid search for SVM in an range, which gives the best result as expected, though takes a large amount of time.

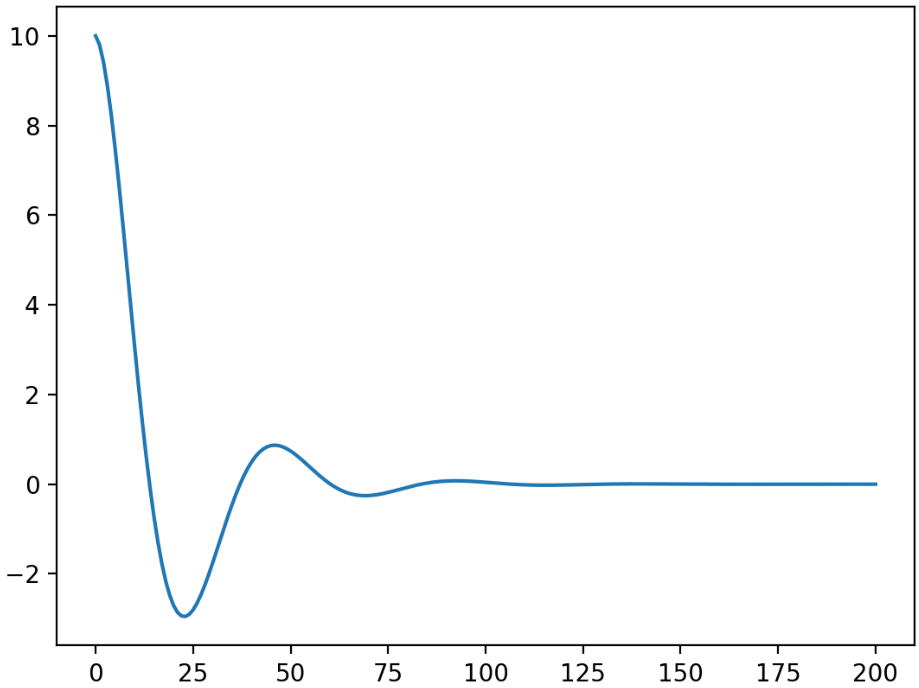
Answer to Question 2:

2.1:

for beta = 0.0

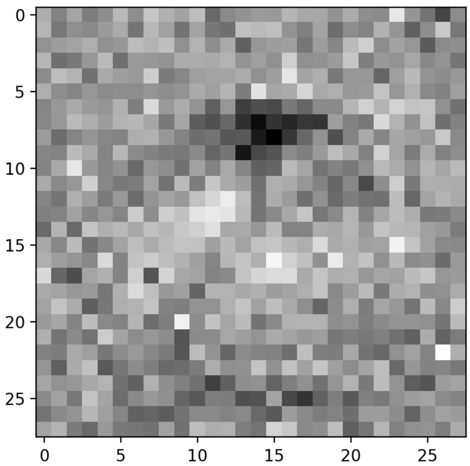


for beta = 0.9

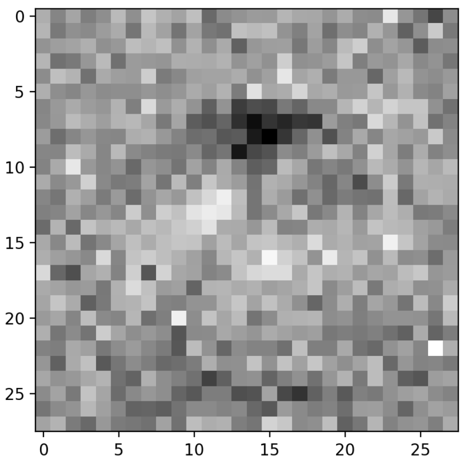


2.3:

when beta = 0 :
The surrogate hinge loss for training set is: 0.101859166915
The classification accuracy on the training set is: 0.964716553288
The surrogate hinge loss for test set is: 0.104682590814
The classification accuracy on the test set is: 0.96227783823
The grayscale image for w is:



when beta = 0.1 :
The surrogate hinge loss for training set is: 0.101651628259
The classification accuracy on the training set is: 0.963537414966
The surrogate hinge loss for test set is: 0.107272921434
The classification accuracy on the test set is: 0.961552412042
The grayscale image for w is:



Answer to Question 3:

3.1. Since K is symmetric, we can write $K = UDU^T$ for U is an orthogonal matrix and D is an diagonal matrix.

claim ①: K is positive semidefinite $\Rightarrow$ $\forall x \in \mathbb{R}^d$ we have $x^T K x \geq 0$.

Proof: Since K ~~is non~~ has no negative eigenvalues, and the entries on the digonal of D are eigenvalues of K, So we can write $D = C^2$ ~~for~~ where C is also an diagonal matrix $(C_{ii} = \sqrt{D_{ii}})$. Therefore, $K = UCC^TU^T = UC(UC)^T$

~~Therefore, $X^T K x = x^T U C C^T U^T x^T =$~~ Let $P^T = UC$, $P = (UC)^T$.

Therefore, $x^T K x = x^T P^T P x = (Px)^T (Px) \geq 0$.

Claim ②: $\forall x \in \mathbb{R}^d$ we have $x^T K x \geq 0 \Rightarrow K$ is positive semidefinite.

Proof:

Suppose v is an eigenvector of A with eigenvalue λ. $v^T A v = \lambda v^T v \geq 0$.

Since $v^T v \geq 0$, we have $\lambda \geq 0$. So, A has no negative eigenvalue. And since A is symmetric, we can conclude that A is positive semidefinite.

So, by claim ① and ②, K is positive semidefinite $\Leftrightarrow$ $\forall x \in \mathbb{R}^d$, $x^T A x \geq 0$. QED.

Q3.2.

1. let $\phi(x)=\sqrt{a}$ and $\phi(y)=\sqrt{a}$. Therefore $k(x,y)=<\phi(x),\phi(y)>=$ $<\sqrt{a},\sqrt{a}>=\sqrt{a}\cdot\sqrt{a}=a$, which is well-defined. Embedding function found, therefore $k(x,y)$ is a kernel for $a>0$.

2. Let ~~$\phi(x)=f(x)$~~ the embedding function $\phi(x)=f(x)$.
So, $k(x,y)=<\phi(x),\phi(y)>=<f(x),f(y)>=f(x)\cdot f(y)$ (since $f(x)$ and $f(y)$ are scalars), which is well-defined.
therefore $k(x,y)=f(x)\cdot f(y)$ is a kernel for all $f:\mathbb{R}^d\to\mathbb{R}$.

3. ~~claim ①:~~
let the embedding function for k_1 be ϕ^1, for k_2 be ϕ^2, the gram matrix for k_1 be $\mathbb{K}_1$, for k_2 be $\mathbb{K}_2$.
claim ① $k'(x,y)=a k_1(x,y)$ is a kernel for $a>0$.
proof: let the embedding function for k' be $\sqrt{a}\,\phi^1$.
So, $k'(x,y)=<\sqrt{a}\,\phi^1(x),\sqrt{a}\,\phi^1(y)>=\sqrt{a}<\phi^1(x),\sqrt{a}\,\phi^1(y)>$
$=\sqrt{a}\cdot\sqrt{a}<\phi^1(x),\phi^1(y)>=a\,k_1(x,y)$, which is well-defined.

claim ②. $k''(x,y)=k_1(x,y)+k_2(x,y)$ is a kernel.
let the matrix $\mathbb{K}=\mathbb{K}_1+\mathbb{K}_2$. $\mathbb{K}$ is symmetric since $\mathbb{K}_1$ and $\mathbb{K}_2$ are symmetric. ~~since $\forall x\in\mathbb{R}^d$~~ Assume $\mathbb{K}_1\in\mathbb{R}^{d\times d}$. Therefore,
$\forall\, x\in\mathbb{R}^d$, $x^T\mathbb{K}_1x\geq0$, $x^T\mathbb{K}_2x\geq0$. So, $x^T\mathbb{K}_1x+x^T\mathbb{K}_2x=$
$x^T(\mathbb{K}_1+\mathbb{K}_2)x\geq0$. $\Longrightarrow$ $x^T\mathbb{K}x\geq0$. Therefore $\mathbb{K}$ as k'''s gram matrix, is positive semidefinite by Q3.1.

Since K_1, K_2 and $a>0$ are arbitrarily chosen, we can conclude from claim ① and ② that, If K_1 and K_2 are kernels and $a,b>0$, $K(x,y)=aK_1(x,y)+bK_2(x,y)$ is a kernel.

QED.

4. let K_1's gram matrix be in $\mathbb{R}^{d\times d}$. ~~Then, $\forall x\in\mathbb{R}^d$,~~
let ϕ' be K_1's embedding function. Then, $\forall x\in\mathbb{R}^d$,
$\sqrt{K_1(x,x)}=\sqrt{\langle\phi'(x),\phi'(x)\rangle}=\|\phi'(x)\|$ by the definition of inner product and Euclidean modular.

Let $\phi(x)=\dfrac{\phi'(x)}{\|\phi'(x)\|}$ for all $x\in\mathbb{R}^d$.

therefore, $\langle\phi(x),\phi(y)\rangle=\left\langle\dfrac{\phi'(x)}{\|\phi'(x)\|},\dfrac{\phi'(y)}{\|\phi'(y)\|}\right\rangle=\dfrac{1}{\|\phi'(x)\|\|\phi'(y)\|}\langle\phi'(x),\phi'(y)\rangle$

$$=\frac{\langle\phi'(x),\phi'(y)\rangle}{\sqrt{K_1(x,x)}\sqrt{K_1(y,y)}}=\frac{K_1(x,y)}{\sqrt{K_1(x,x)}\sqrt{K_1(y,y)}}=K(x,y).$$

Therefore, $\phi(x)=\dfrac{\phi'(x)}{\|\phi'(x)\|}$ is $K(x,y)$'s embedding function.

Therefore it is a kernel. QED.