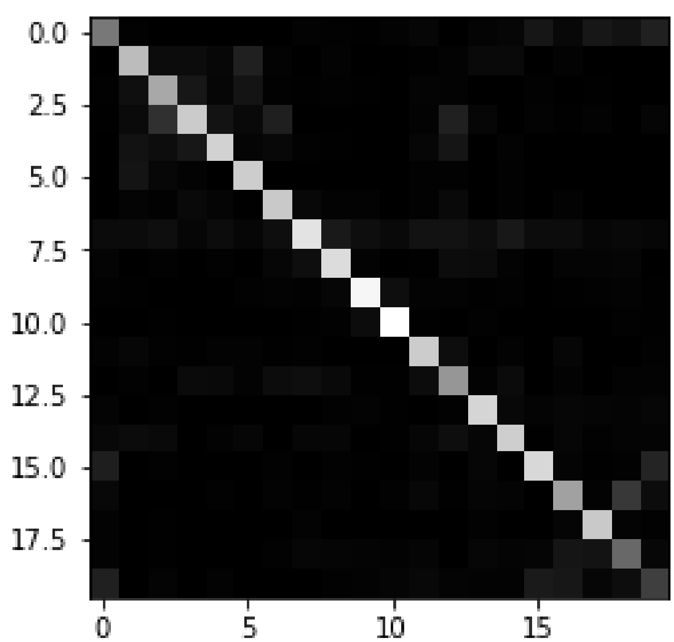
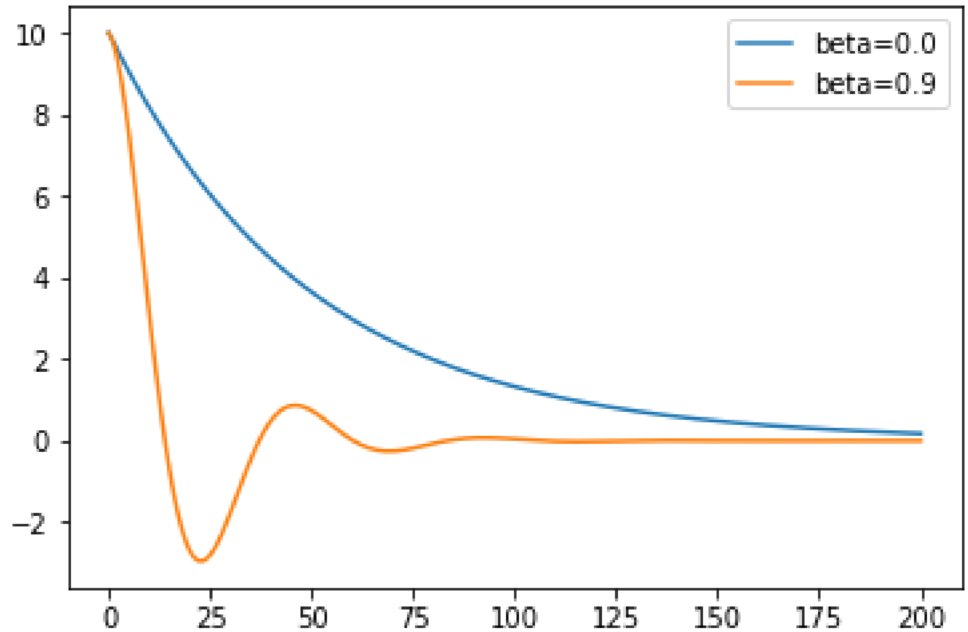
CSC411 A3
Guanxiong Liu
1002077726 liuguanx
Q1.

a. I choose SVM(linear), Neural Network (with 25 layers) and Logistic Regression to train the data. Also, I use tf_idf data because the most frequent word should have the least weight. Intuitively, the rare words will be the most helpful for classification. I did not train Neural Network with too many layers because it will run for a long time. That means the result should be better with more hidden layers.
b. For SVM, train accuracy = 0.954127629485593 and test accuracy = 0.6631704726500266
   For Neural Network, train accuracy = 0. 9747215838783808 and test accuracy = 0. 6975570897503983
   For Logistic Regression, train accuracy = 0.8957044369807319 and test accuracy = 0.6775092936802974
   For Bernoulli Naïve Bayes baseline, train accuracy = 0.5987272405868835 and test accuracy = 0.4579129049389272
c. I use sklearn builtin function GridSearchCV (cross validation) with parameters = {'kernel':('linear', 'rbf'), 'C': [1, 10]} to find the best hyper parameters for SVM. The best parameter is C=1, kernel = linear.
d. I run through all the algorithms we discussed in the class and find these three methods have the top accuracy. I thought KNN would have a decent performance, but it does not work as well as I thought. The methods work as same as my expectation. The discriminative models defeat generative models (ex. Bnb, Multinomial NB).
e. For the computation of the confusion matrix for Neural Network. Please refer to the code.
f. Neural Network is most confused about class 16 and 18. The confusion matrix graph shown below.



CSC411 A3
Guanxiong Liu
1002077726 liuguanx
Q2.1

a. Please see the attached code.
b.

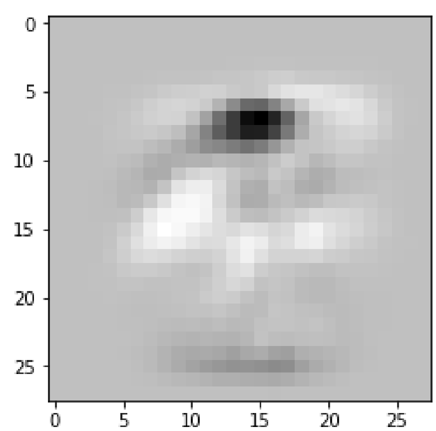


Q2.2

a. Please see the attached code.

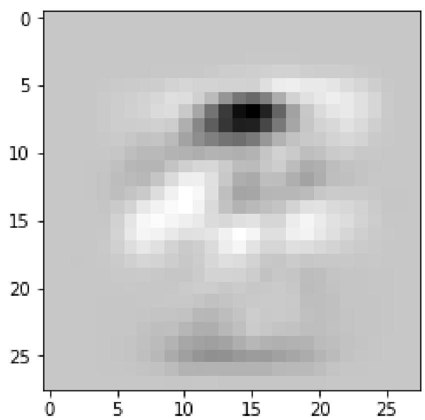
Q2.3

a. For beta = 0.1,
   train loss = 0.5835369453114629
   average train hinge loss = 0.3694244002919029
   test loss = 0.5842337924527631
   average test hinge loss = 0.37012124743320307
   train accuracy = 0.9266213151927437
   test accuracy = 0.9245556764599202



CSC411 A3
Guanxiong Liu
1002077726 liuguanx

b. For beta = 0.0
   train loss = 0.5567972513778472
   average train hinge loss = 0.3488265166912148
   test loss = 0.5463558995332269
   average test hinge loss = 0.3383851648465945
   train accuracy = 0.9118367346938776
   test accuracy = 0.9122234312658687



CSC411 A3
Guanxiong Liu
1002077726 liuguanx

3.1 1. Prove for all vectors $x \in \mathbb{R}^d$ we have $x^T k x \geq 0 \Rightarrow$ a symmetric matrix $K \in \mathbb{R}^{d \times d}$ is positive semidefinite

λ is an eigenvalue of K. Then there exist eigenvector $x \in \mathbb{R}^d$ s.t. $Kx = \lambda x$. So $0 \leq x^T k x = \lambda x^T x$. Since $x^T x$ is positive for all x, implies λ is non-negative. Therefore, a symmetric matrix $K \in \mathbb{R}^{d \times d}$ is positive semidefinite.

Prove a symmetric matrix $K \in \mathbb{R}^{d \times d}$ is positive semidefinit $\Rightarrow$ for all vectors $x \in \mathbb{R}^d$ we have $x^T k x \geq 0$

$K = ADA^T$, A is orthogonal matrix and D is diagonal matrix, $D = B^2$ where B is a diagonal matrix then $K = ABBA^T = (AB)(AB)^T = CC^T = E^T E \geq 0$ D is non-negative since All eigenvalue of K is non negative. So, for all vectors $x \in \mathbb{R}^d$ we have $x^T K x = X^T E^T E X = (EX)^T EX \geq 0$

3.2 1. since $K_1(x,y) = 0$ is a positive semidefinite kernel.

$\therefore K(x,y) = \alpha + K_1(x,y) = \alpha + 0 = \alpha$ is also a valid kernel.

Proof: Let ϕ_1 denote a feature map of K_1, using the feature map $\phi: x \rightarrow [\phi_1(x), \sqrt{\alpha}]^T$, we have.

$$\langle \phi(x), \phi(y) \rangle = \langle \phi_1(x), \phi_1(y) \rangle + \alpha = K_1(x,y) + \alpha = 0 + \alpha = \alpha$$
$$= K(x,y)$$

$\therefore K(x,y)$ is a valid kernel.

2. Using the feature map $\phi: x \rightarrow f(x),\ y \rightarrow f(y)$, we have

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

$\therefore K(x,y)$ is a valid Kernel function.

3. K_1 has feature map ϕ_1 and K_2 has its feature map ϕ_2

$\therefore aK_1(x,y) = \langle \sqrt{a}\phi_1(x), \sqrt{a}\phi_1(y) \rangle$ $\quad bK_2(x,y) = \langle \sqrt{b}\phi_2(x), \sqrt{b}\phi_2(x) \rangle$

$$K(x,y) = \langle \sqrt{a}\phi_1(x), \sqrt{a}\phi_1(y) \rangle + \langle \sqrt{b}\phi_2(x), \sqrt{b}\phi_2(x) \rangle$$
$$= \langle [\sqrt{a}\phi_1(x), \sqrt{b}\phi_2(x)], [\sqrt{a}\phi_1(y), \sqrt{b}\phi_2(y)] \rangle$$

So, It is valid kernel

CSC411 A3
Guanxiong Liu
1002077726 liuguanx

Therefore $K(x, y)$ is a valid Kernel.

4. Let ϕ be a feature map for K. Define $\phi(x) = \frac{\phi_1(x)}{\|\phi_1(x)\|}$. Then

$$K(x,y) = \frac{K_1(x,y)}{\sqrt{K_1(x,x)}\sqrt{K_1(y,y)}} = \frac{\phi_1(x) \cdot \phi_1(y)}{\|\phi_1(x)\|\|\phi_1(y)\|}$$

$$= \frac{\phi_1(x)}{\|\phi_1(x)\|} \frac{\phi_1(y)}{\|\phi_1(y)\|} = \langle \phi(x), \phi(y) \rangle$$

So, $K(x,y)$ is a valid kernel.