CSC411 A3 Report

Tianshu Zhu

Q1

BernoulliNB baseline train accuracy = 0.5987272405868835
BernoulliNB baseline test accuracy = 0.4579129049389272

Logistic regression train accuracy = 0.9567792115962525
Logistic regression test accuracy = 0.6178969729155602

SVM train accuracy = 0.9744564256673148
SVM test accuracy = 0.5720924057355284

Decision tree train accuracy = 0.9747215838783808
Decision tree test accuracy = 0.41688794476898566

I used the default hyper parameters without further tuning.

I picked the Logistic Regression and SVM because I am most familiar with them and they are closely related. I picked Decision Tree because I have never used it before.

They did not work as I expected, I expected that SVM should do better than Logistic Regression because just by theory SVM may generalize better than Logistic Regression. But it turns out that Logistic Regression have better test accuracy.

Confusion matrix for Logistic Regression:

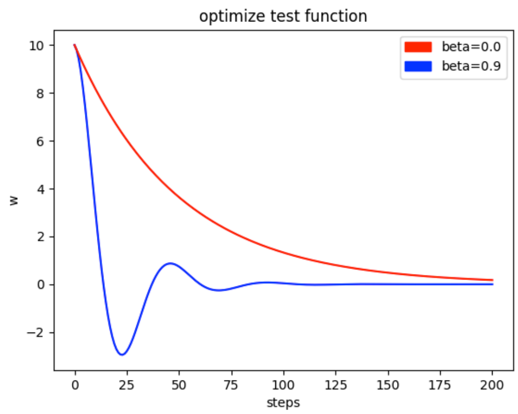
```
[[141.   2.   4.   1.   0.   0.   0.   6.   3.   6.   6.   3.   6.   5.   4.  28.   8.  30.  14.  34.]
 [  2. 252.  21.  21.  10.  50.   4.   7.   1.   4.   3.   8.  20.  12.  17.   5.   3.   1.   1.   7.]
 [  1.  22. 223.  41.  11.  32.   6.   0.   1.   0.   2.   7.  12.   2.   5.   1.   4.   0.   0.   1.]
 [  2.   7.  46. 222.  36.  12.  12.   3.   2.   0.   0.   4.  26.   3.   6.   2.   2.   2.   3.   2.]
 [  2.   7.  19.  30. 240.   7.  15.   5.   4.   2.   0.   6.  18.   2.   4.   2.   3.   0.   1.   3.]
 [  4.  21.  15.   5.   2. 246.   1.   0.   1.   1.   0.   2.   8.   2.   1.   0.   2.   1.   2.   2.]
 [  1.  10.   3.  15.  10.   4. 301.  10.   5.   6.   2.   3.  16.   8.   2.   3.   2.   0.   1.   2.]
 [  5.   3.   4.   3.  11.   1.   7. 253.  32.   6.   4.   5.  17.  16.   9.   2.  10.   7.   7.   6.]
 [ 16.  10.  21.  10.  24.   9.  20.  47. 288.  29.  17.  26.  22.  29.  24.  20.  26.  17.  16.  12.]
 [  6.   7.   5.   0.   2.   8.   4.   9.  10. 287.  19.   8.   4.   4.   6.   2.  10.   9.   5.   3.]
 [  3.   2.   1.   1.   4.   0.   0.   2.   0.  25. 324.   2.   1.   3.   2.   2.   3.   0.   3.   2.]
 [  2.   4.   6.   4.   3.   6.   1.   2.   4.   0.   1. 261.  14.   3.   4.   1.  15.   4.   6.   1.]
 [ 11.  13.   2.  33.  24.   4.   9.  17.  14.   2.   1.  15. 194.  10.  13.   3.   2.   3.   3.   1.]
 [ 10.   2.   6.   1.   1.   1.   1.   2.   5.   3.   1.   5.  11. 256.   8.  10.   4.   4.   5.  12.]
 [ 15.  12.   8.   3.   5.   5.   5.   8.  10.   5.   5.   8.  15.   7. 265.   5.   9.   5.   9.   7.]
 [ 33.   1.   0.   1.   1.   4.   1.   4.   1.   5.   2.   2.   2.  12.   5. 269.   8.  12.   9.  42.]
 [  8.   1.   2.   0.   1.   2.   1.   4.   5.   3.   2.  11.   1.   8.   7.   2. 199.  10.  81.  20.]
 [ 12.   4.   2.   0.   0.   4.   2.   3.   3.   5.   1.   6.   1.   4.   2.   9.  10. 244.  13.  11.]
 [  6.   7.   5.   1.   0.   0.   0.   9.   6.   7.   4.   8.   4.   7.   7.   4.  27.  18. 118.  15.]
 [ 39.   2.   1.   0.   0.   0.   0.   5.   3.   1.   5.   6.   1.   3.   3.  28.  17.   9.  13.  68.]]
```

Most confused 2 classes:
class 16 and class 18.
Note class index starts with 0, these two entry are highlighted in the confusion matrix above.

Q2

2.1):

Plot w for 200 steps



2.3):

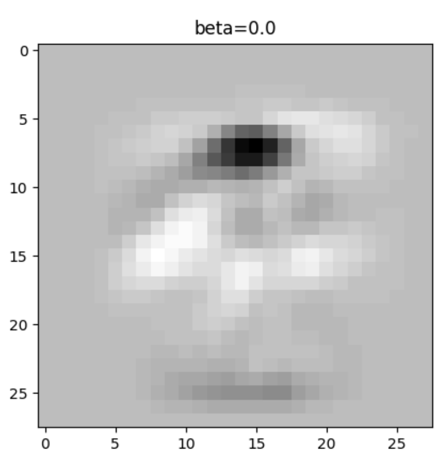
Used average of hinge loss as loss
train loss with beta=0.0: 0.397240485900029
train loss with beta=0.1: 0.35458218099670796

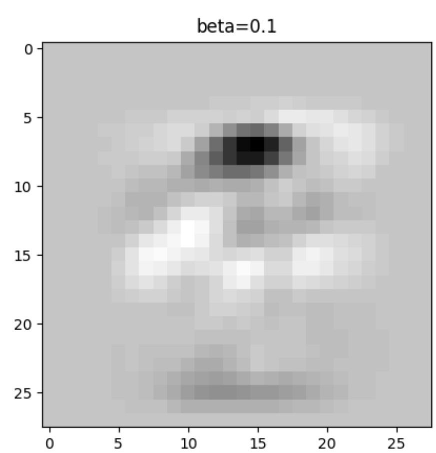
test loss with beta=0.0: 0.40062763119052897
test loss with beta=0.1: 0.34278258535832506

train accuracy with beta=0.0: 0.9126530612244897
train accuracy with beta=0.1: 0.9057596371882086

test accuracy with beta=0.0: 0.9147624229234675
test accuracy with beta=0.1: 0.9038810301051868

Plot w as a 28x28 image:





Q3.1

- show symmetric matrix $K \in \mathbb{R}^d \times \mathbb{R}^d$ is positive semidefinite
  $\Leftrightarrow \forall x \in \mathbb{R}^d \quad x^T K x \geq 0$

- $\Rightarrow$
  Assume $K \in \mathbb{R}^{d \times d}$ is a symmetric matrix
  Assume K is positive semidefinite
  Assume $x \in \mathbb{R}^d$
  Let $u_1, \cdots, u_d$ be orthogonal eigenvectors of K with eigenvalues $\lambda_1, \cdots, \lambda_d$
  (by spectral theorem they exist)
  Then $\lambda_1, \cdots, \lambda_d \geq 0$ since K is positive-semidefinite
  Let $x = c_1 u_1 + \cdots + c_d u_d \quad (c_1, \cdots, c_d \in \mathbb{R})$
  Then $x^T K x$
  $= (c_1 u_1 + \cdots + c_d u_d)^T (c_1 \lambda_1 u_1 + \cdots + c_d \lambda_d u_d)$
  $= \lambda_1 (c_1 u_1)^2 + \cdots \lambda_d (c_d u_d)^2$ since $u_1, \ldots, u_d$ orthogonal
  ≥ 0
  Then For symetric matrix $K \in \mathbb{R}^{d \times d}$, if K is positive semi definite
  then for all vectors $x \in \mathbb{R}^d$ we have $x^T K x \geq 0$

- $\Leftarrow$
  Assume $K \in \mathbb{R}^{d \times d}$ is a symmetric matrix
  Assume $\forall x \in \mathbb{R}^d . \ x^T K x \geq 0$
  Assume $v \in \mathbb{R}^d$, v is eigenvector of K with eigenvalue λ
  Then $v^T K v = v^T \lambda v = v^T v \lambda \geq 0$
  Since $v^T v \geq 0$ always true
  Then $\lambda \geq 0$
  Then all eigenvalue of $K \geq 0$
  Then K is positive-semidefinite
  Then For symetric matrix $K \in \mathbb{R}^{d \times d}$, if for all vectors $x \in \mathbb{R}^d$ we have
  $x^T K x \geq 0$, then K is positive semi definite

3.2

1): Assume $K(x,y) = \alpha$; $\alpha > 0$
Let $\phi(x) = \sqrt{\alpha}$
Then $K(x,y) = \alpha = \sqrt{\alpha} \cdot \sqrt{\alpha} = \phi(x) \cdot \phi(y)$
Then $K(,)$ is a kernel

2): Assume $f: \mathbb{R}^d \to \mathbb{R}$
Let $\phi(x) = f(x)$
Then $K(x,y) = f(x) \cdot f(y) = \phi(x) \cdot \phi(y)$
Then $K(,)$ is a kernel

3): Assume $k_1(x,y)$, $k_2(x,y)$ are kernels
Assume $k(x,y) = a \cdot k_1(x,y) + b k_2(x,y)$; $a, b > 0$
Let K_1, K_2 be gram matrix of k_1, k_2
Let K be gram matrix of k
Assume x is arbitrary
Then $x^T K x$

$$= x^T (a K_1 + b K_2) x$$
$$= a x^T K_1 x + b x^T K_2 x$$
$$\geq 0$$

Then K is positive semidefinite
Then $k(,)$ is kernel

4): Assume $k_1(x,y)$ is a kernel, $x, y \in \mathbb{R}^n$
Assume $k(x,y) = k_1(x,y) / \sqrt{k_1(x,x)} \cdot \sqrt{k_1(y,y)}$
Then $\exists\, \phi^{(1)}$ $k_1(x,y) = \phi^{(1)}(x) \cdot \phi^{(1)}(y)$
Let $\phi(x)$ s.t $\phi_i(x) = \phi_i^{(1)}(x) / \|\phi^{(1)}(x)\|$
Then $k(x,y)$

$$= k_1(x,y) / \sqrt{k_1(x,x)} \sqrt{k_1(y,y)}$$
$$= \phi^{(1)}(x) \cdot \phi^{(1)}(y) / \sqrt{\|\phi^{(1)}(x)\|^2} \sqrt{\|\phi^{(1)}(y)\|^2}$$
$$= \phi^{(1)}(x) \cdot \phi^{(1)}(y) / \|\phi^{(1)}(x)\| \, \|\phi^{(1)}(y)\|$$
$$= [\phi_1^{(1)}(x) \phi_1^{(1)}(y) + \cdots + \phi_n^{(1)}(x) \phi_n^{(1)}(y)] / \|\phi^{(1)}(x)\| \, \|\phi^{(1)}(y)\|$$
$$= \frac{\phi_1^{(1)}(x)}{\|\phi^{(1)}(x)\|} \frac{\phi_1^{(1)}(y)}{\|\phi^{(1)}(y)\|} + \cdots + \frac{\phi_n^{(1)}(x)}{\|\phi^{(1)}(x)\|} \frac{\phi_n^{(1)}(y)}{\|\phi^{(1)}(y)\|}$$
$$= \phi_1(x) \phi_1(y) + \cdots + \phi_n(x) \phi_n(y)$$
$$= \phi(x) \cdot \phi(y)$$

Then $K(,)$ is a kernel