Cogs185 HW1 Report

Jiawen Wang

Question 2

Question 2.1

The mathematical form of the gradient of the loss function

Loss function is:

$$L(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \max(0, 1 - y_i\langle \mathbf{x}_i, \mathbf{w}\rangle)$$

The gradient w.r.t $\mathbf{w}$ of the loss function:

$$L'(\mathbf{w}) = \frac{dL(\mathbf{w})}{d\mathbf{w}} = \mathbf{w} + C\sum_i \begin{cases} -y_i\mathbf{x}_i & \text{if } y_i\langle \mathbf{x}_i, \mathbf{w}\rangle \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Question 2.2

The optimal w∗ = arg minw L(w) as the minimizer

The optimal w* value is [[0.15983746 6.39176826 -4.40728501]
[0.2970873 -0.31933683 -0.92796169]
[0.45801751 -2.25023377 -1.96619639]
[-1.0526293 1.6890044 2.00050391]
[-0.60438444 -3.30471002 3.70373197]]

The C value associated is 10.0. This w is optimal because compared to all the other w coming from C values, {0.5, 2.0, 5.0},
Since all the testing accuracies are the same- 100% for different C values, we will prioritize the one with greatest training accuracy.
This w gives the greatest training accuracy of 95.83333333333334 %, makes our training model the most fittable.

Question 2.3

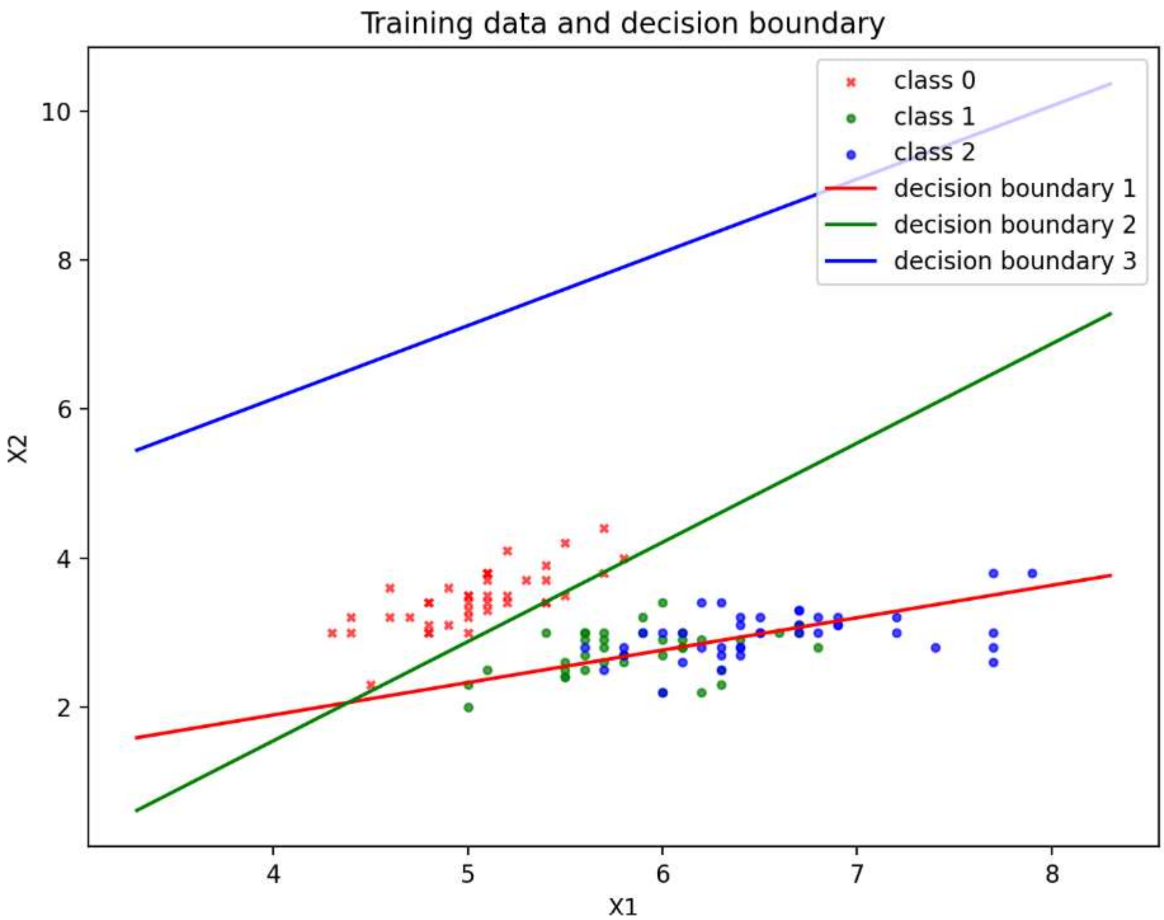
Training accuracy and test accuracy with C = 0.5, 2.0, 5.0, 10.0.

C = 0.5 Total training accuracy: 94.16666666666667 %.
Total test accuracy: 100.0 %.
C = 2.0
Total training accuracy: 91.66666666666666 %.
Total test accuracy: 100.0 %.
C = 5.0
Total training accuracy: 94.16666666666667 %.

Total test accuracy: 100.0 %.
C = 10.0
Total training accuracy: 95.83333333333334 %.
Total test accuracy: 100.0 %.

Question 2.4

Plot training data along with decision boundaries (w∗1, . . . , w∗K), K = 3, using the first two dimensions of the features for x.



Question 3

Question 3.1

The mathematical form of the gradient of the loss function.

Loss function is:

$$L(\mathbf{w}_1, \ldots, \mathbf{w}_K) = \frac{1}{2} \sum_{k=1}^{K} \|\mathbf{w}_k\|^2 + C \sum_i \sum_{k=1, k \neq y_i}^{K} max\Big(0, 1 - (< \mathbf{w}_{y_i}, \mathbf{x}_i > - < \mathbf{w}_k, \mathbf{x}_i$$

The gradient w.r.t $\mathbf{w}$ of the loss function:

$$L'(\mathbf{w}_b) = \frac{dL(\mathbf{w}_b)}{d\mathbf{w}_b} = \mathbf{w}_b + C \sum_i \sum_{k=1, k \neq y_i}^{K} \begin{cases} \mathbf{x}_i & , \texttt{ if } (b = k) \wedge (b \neq y_i) \wedge (< \mathbf{w}_{y_i}, \mathbf{x}_i > \\ -\mathbf{x}_i & , \texttt{ if } (b \neq k) \wedge (b = y_i) \wedge (< \mathbf{w}_{y_i}, \mathbf{x}_i > \\ 0 & , \texttt{ otherwise} \end{cases}$$

Question 3.2

The optimal (w∗1, . . . , w∗K) = arg minw1,...,wK L(w1, . . . , wK) as the minimizer.

The optimal w value is [[0.24049589 0.61570382 -0.85619971]
[0.54962307 0.34553133 -0.8951544]
[0.94533576 0.151773 -1.09710876]
[-1.40369287 -0.18671239 1.59040526]
[-0.77774176 -0.88098094 1.6587227]]

The C value associated is 2.0. This w is optimal because compared to all the other w coming from C values, {0.5, 5.0, 10.0},
Since all the testing accuracies are the same - 100% for different C values, we will prioritize the one with greatest training accuracy.
This w gives the greatest training accuracy of 97.5 %, fitting our training model more. Also, notice that when C = 10.0, the training accuracy
is also 97.5%. However, since when C = 2.0, the Gradient descent converged fastest - after 2136 iterations. It is the most accurate and
computationally efficient, therefore we this w is optimal.

Question 3.3

Training accuracy and test accuracy with C = 0.5, 2.0, 5.0, 10.0

C = 0.5 Total training accuracy: 96.66666666666667 %.
Total testing accuracy: 100.0 %.

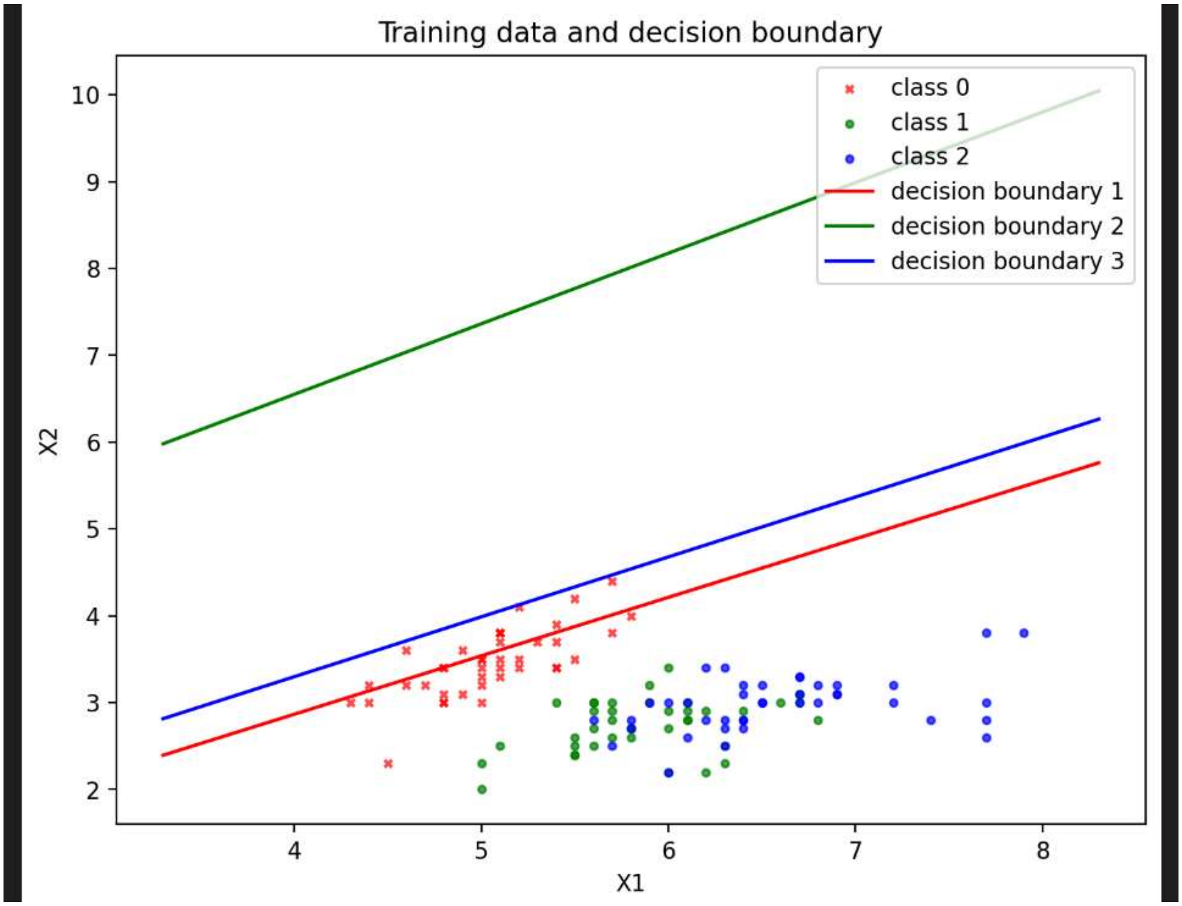
C = 2.0
Total training accuracy: 97.5 %.
Total testing accuracy: 100.0 %.

C = 5.0
Total training accuracy: 95.83333333333334 %.
Total testing accuracy: 100.0 %.

C = 10.0
Total training accuracy: 97.5 %.
Total testing accuracy: 100.0 %.

Question 3.4

Plot training data along with decision boundaries (w∗1, . . . , w∗K), K = 3, using the first two dimensions of the features for x.



Question 4

Question 4.1

The mathematical form of the gradient of the loss function.

Loss function is:

$$L(\mathbf{w}_1, \dots, \mathbf{w}_K, b_1, \dots, b_K) = -\sum_i \ln p_{y^{(i)}} + \frac{\lambda}{2} \sum_{k=1}^{K} \|\mathbf{w}_k\|^2.$$

where $p_j = p(y = j|\mathbf{x}) = \frac{e^{f_j}}{\sum_{k=1}^{K} e^{f_k}}$; $f_j = \mathbf{w}_j \cdot \mathbf{x} + b_j$

The gradient w.r.t $\mathbf{w}$ of the loss function:

$$\frac{dL(\mathbf{w}_1,\dots,\mathbf{w}_K,b_1,\dots,b_K)}{d\mathbf{w}_k} = \lambda \mathbf{w}_k + \sum_{i,y_i=k}(p(k|\mathbf{x}_i) - 1)\mathbf{x}_i + \sum_{i,y_i\neq k} p(k|\mathbf{x}_i)\mathbf{x}_i.$$

$$\frac{dL(\mathbf{w}_1,\dots,\mathbf{w}_K,b_1,\dots,b_K)}{db_k} = \sum_{i,y_i=k}(p(k|\mathbf{x}_i) - 1) + \sum_{i,y_i\neq k} p(k|\mathbf{x}_i).$$

Question 4.2

The optimal (w∗1, . . . , w∗K, b∗1, . . . , b∗K) = arg minw1,...,wK,b1,...,bK L(w1, . . . , wK, b1, . . . , bK) as the minimizer.

The optimal w value is: [[0.32541817 0.66260006 1.66815353 -2.3413803 -1.0726503]
[0.47169721 0.50819969 -0.32458591 0.02619788 -0.93840975]
[-0.79711538 -1.17079975 -1.34356762 2.31518242 2.01106005]]
The optimal b value is [[0.33784895] [0.48593311] [-0.82378206]]

The C value associated is 0.1. This w is optimal because compared to all the other w coming from lambda values {0, 10^{-5}, 10^{-3}}, this w value is the smallest,
and it results in the greatest training & testing accuracy, least number of iterations to convergence.
Since all the training and testing accuracies are the same - 97.5% and 100% for different C values, we will prioritize the one with least number of iterations to convergence
and the smallest w.
When lambda = 0.1, the Gradient descent converged fastest - after 5315 iterations. It is the most computationally efficient. Also, compared to other lambdas,
the w value associated with it is the smallest. Therefore, this w is optimal.

Question 4.3

Training accuracy and test accuracy with λ = 0, 10^{-5}, 10^{-3}, 0.1.

w = 0: The training accuracy: 97.5 %.
The test accuracy: 100.0 %.

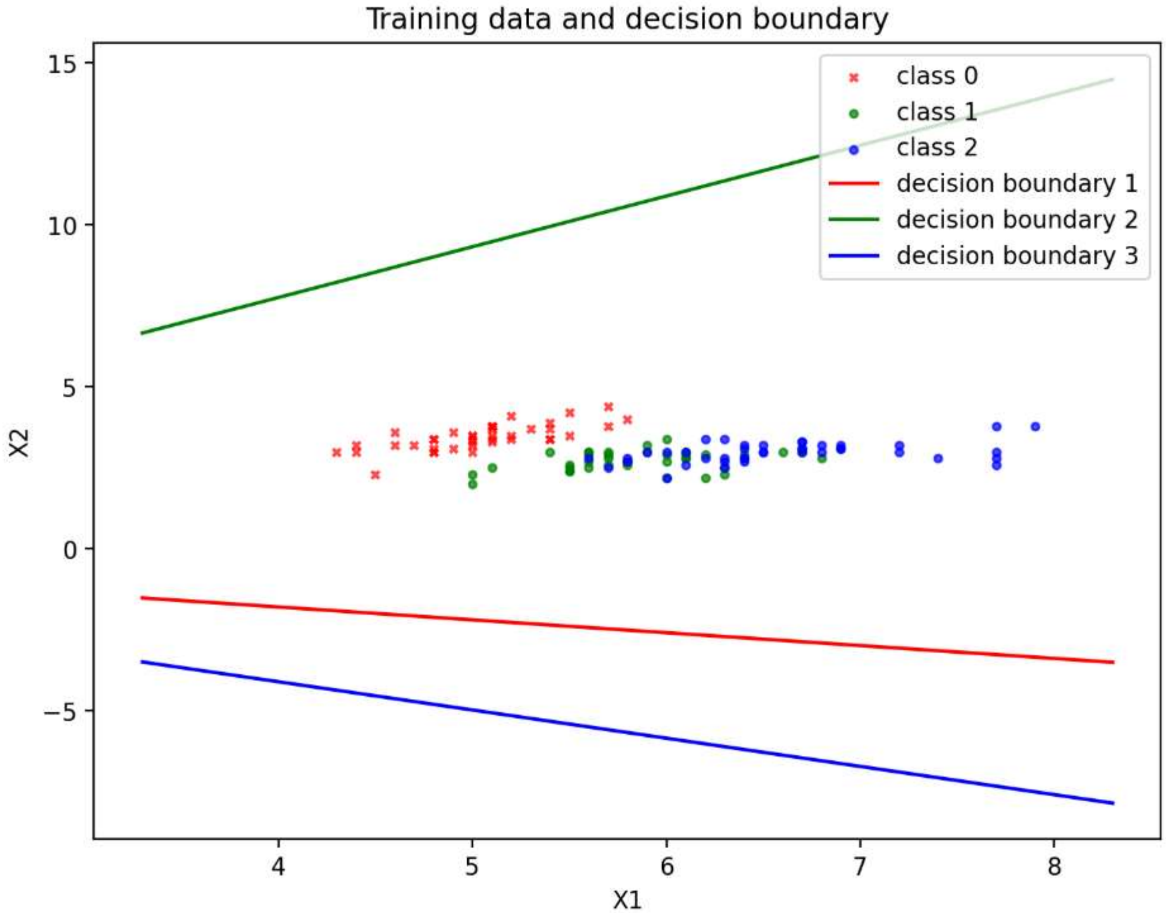
w = 10^{-5}
The training accuracy: 97.5 %.
The test accuracy: 100.0 %.

w = 10^{-3}
The training accuracy: 97.5 %.
The test accuracy: 100.0 %.

w = 0.1
The training accuracy: 97.5 %.
The test accuracy: 100.0 %.

Question 4.4

Plot training data along with decision boundaries (w∗1, . . . , w∗K, b∗1, . . . , b∗K), K =3 using the first two dimensions of the features for x.


Training data and decision boundary
class 0
class 1
class 2
decision boundary 1
decision boundary 2
decision boundary 3
X1
X2
15
10
5
0
−5
4
5
6
7
8