EE559 Homework 4 (week 5)

Jingquan Yan

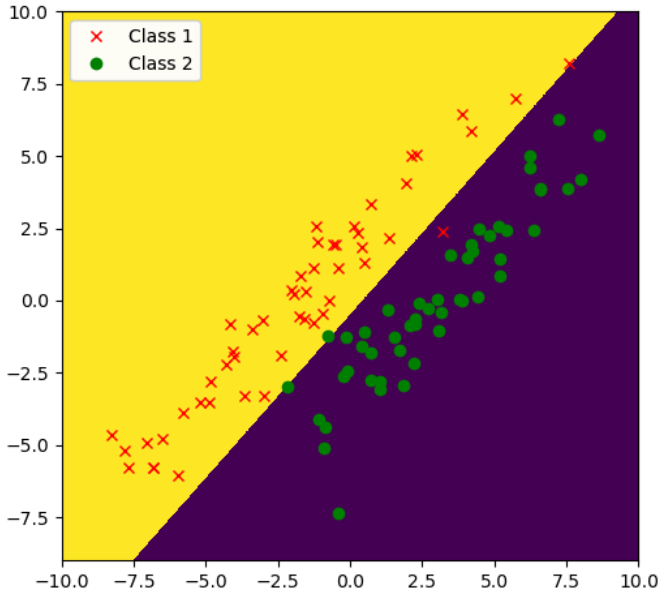
USC ID: 1071912676

Email: jingquan@usc.edu

EE559 repository: Github

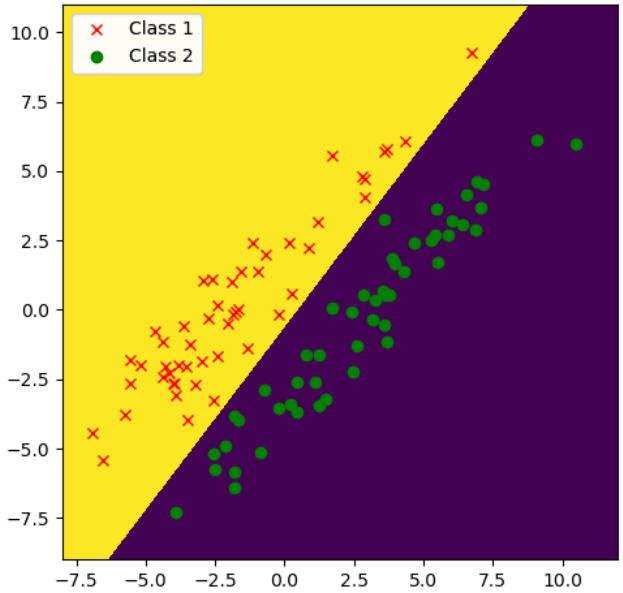
Synthetic 1 Dataset:

1. Training dataset:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW4/Problem 1.py"
The final weight vector is: [-62.0968   54.89479  26.1    ]
The error rate is : 0.02
```

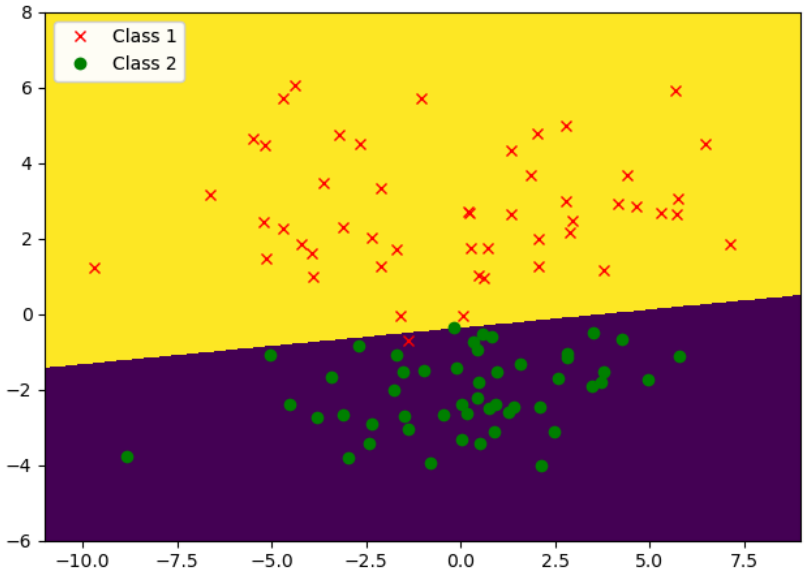
2. Test dataset:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW4/Problem 1.py"
The final weight vector is: [-17.04896  12.9296    8.1    ]
The error rate is : 0.0
```

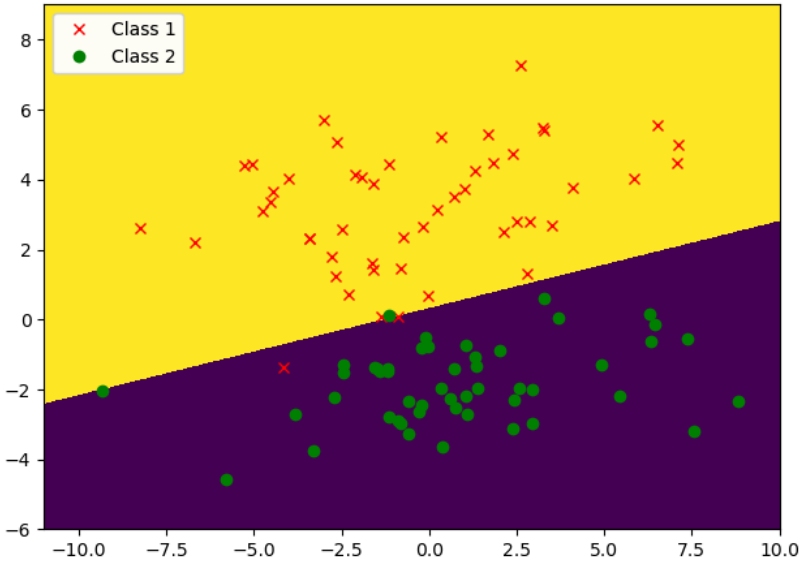
Synthetic 2 Dataset:

1. Training data:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW4/Problem 1.py"
The final weight vector is: [-1.401165 14.51406   5.1     ]
The error rate is : 0.02
```

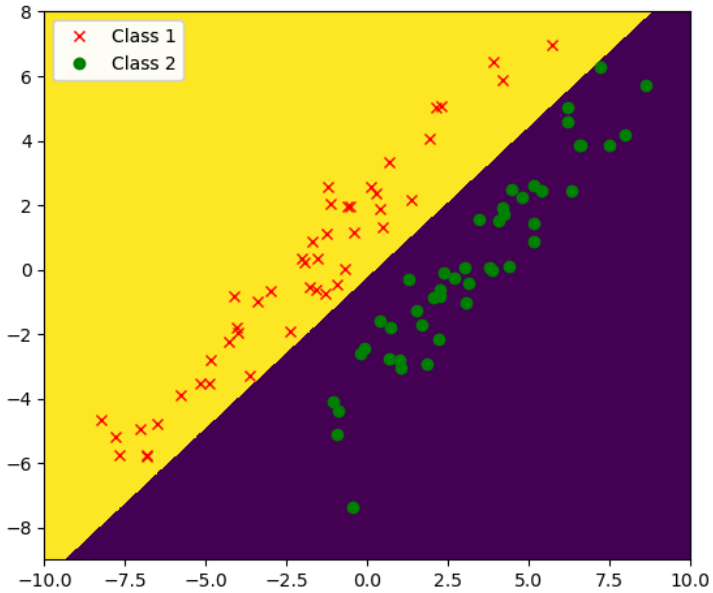
2. Test data:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW4/Problem 1.py"
The final weight vector is: [-4.0885   16.460069 -4.9     ]
The error rate is : 0.02
```

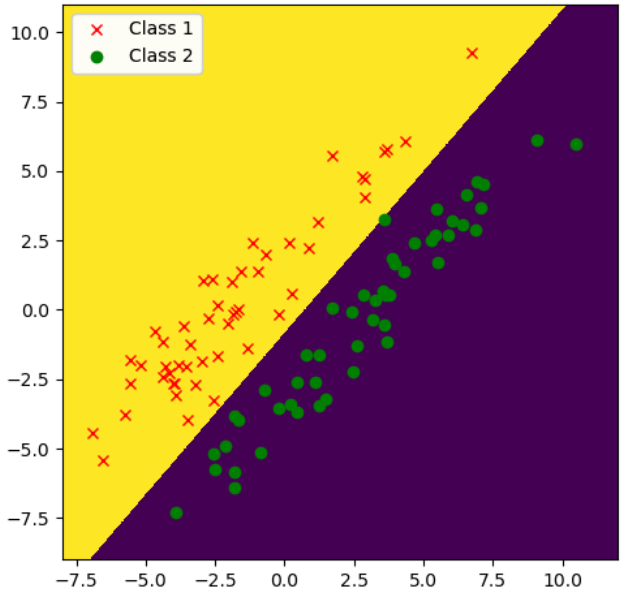
Synthetic 3 Dataset:

1. Training data:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW4/Problem 1.py"
The final weight vector is: [-14.31289  15.31443   4.1     ]
The error rate is : 0.0
```

2. Test data:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW4/Problem 1.py"
The final weight vector is: [-15.5352  13.3405  11.1   ]
The error rate is : 0.0
```

2.

(a) Since the gradient decent dimension is randomly chosen and one dimension is only chosen once in one epoch, we have

$$E\{\Delta w(i)\} \xrightarrow{LLN} \frac{\eta}{N}\sum_{i=1}^{N}\nabla_w J(w_i)$$

(b) For Batch-Gradient-Decent, we have:

$$J(w)=\sum_{i=1}^{N}J(w_i)$$

$$\text{So } \Delta W(i)=\eta\cdot\nabla J(w)=\eta\cdot\nabla\sum_{i=1}^{N}J(w_i)$$

For SGD. we have

$$E(\Delta W(i))=(\Delta W_1\cdot p_1+\dots+\Delta W_N p_N)\eta$$

Since we pick with normal distribution, we have:

$$p_1=p_2=\cdots=p_n=\frac{1}{N}$$

$$\text{So } E(\Delta W(i))=\frac{\eta}{N}\cdot(\Delta W_1(i)+\cdots+\Delta W_N(i))$$

$$=\frac{\eta}{N}\sum_{i=1}^{N}\nabla J(w_i)$$

Employ the linearity property of partial derivative: $\sum\nabla f(x)=\nabla\sum f(x)$

we have: $E(\Delta W(i))=\frac{\eta}{N}\nabla\sum_{i=1}^{N}J(w_i)=\frac{1}{N}\cdot\Delta W(i)$, Q.E.D

Conclusion: $\Delta W(i)$ of BGD is N times of the $E\{\Delta w(i)\}$ of SGD

Explaination: BGD considers the loss of the whole data set and SGD considers one at one time. Since SGD data is normally random chosen, every data has the same probability to be chosen so it is somehow equavilant with considering the whole set. Meanwhile, it's obvious that the loss of SGD should be $\frac{1}{N}$ of the BGD for its randomness.