Q1.1) given $\{1,\dots,n\}$ (x_i, y_i), $x_i \in R^d$ $y_i \in \{-1,1\}$

$\|w_{opt}\| = 1$, $\gamma > 0$, $y_i \cdot (w_{opt}^T x_i) \geq \gamma$, $\forall i \in \{1,\dots,n\}$ ③

$\|x_i\|_2 \leq R$, $R > 0$

if $y_i (w_k^T x_i) < 0 \Rightarrow w_{k+1}^T w_{opt} \geq w_k^T w_{opt} + \gamma \|w_{opt}\|$

Proof:

a. considerig perceptron Algorithm misclassifies (x_i, y_i) in k-th iteration $\Rightarrow$ we have $y_i (w_k^T x_i) < 0$ ① & updated vector of $w_{k+1} = w_k + y_i x_i$ ②

b. we are adding a dot product of w_{opt} on both sides of equesion ②:

$$w_{k+1}^T w_{opt} = (w_k + y_i x_i)^T w_{opt}$$

rules used (side Note)

$(a+b) \cdot c = a \cdot c + bc$

$a \cdot b = a^T b$

c. expanding step b:

$$w_{k+1}^T w_{opt} = w_k^T w_{opt} + y_i x_i^T w_{opt} \quad ④$$

$y_i^T = y_i$

d. given ① we can say y_i and $(w_k^T x_i)$ have different signs.

using the given ③ assumption & $y_i (x_i^T w_{opt})$ has a opposite sign to $\gamma \Rightarrow$

$$y_i (x_i^T w_{opt}) \geq \gamma \|w_{opt}\|_2 \quad ⑤$$

e. using equasions from steps c & d:

$$y_i (x_i^T w_{opt}) \geq \gamma \|w_{opt}\|_2$$

$$w_{k+1}^T w_{opt} = w_k^T w_{opt} + y_i x_i^T w_{opt} \longrightarrow$$

$$(w_{k+1}^T w_{opt}) - (w_k^T w_{opt}) = y_i x_i^T w_{opt}$$

$$(w_{k+1}^T w_{opt}) - (w_k^T w_{opt}) \geq \gamma \|w_{opt}\|_2 \xrightarrow{\text{add } (w_k^T w_{opt}) \text{ to both sides}}$$

$$\therefore w_{k+1}^T w_{opt} \geq w_k^T w_{opt} + \gamma \|w_{opt}\|_2$$

1.2)

$(a+b)^T = a^T + b^T$

1.2 According to the perceptron algorithm:

$w_{k+1} = w_k + y_i x_i$

general rule

$\|w_{k+1}\| = (w_{k+1})^T (w_{k+1})$

Calculate $\|w_{k+1}\|_2^2$

$\|w_{k+1}\|_2^2 = w_{k+1}^T w_{k+1} = (w_k + y_i x_i)^T (w_k + y_i x_i)$

$y_i^T = y_i$

$= (w_k^T + y_i x_i^T)(w_k + y_i x_i)$

$= w_k^T w_k + y_i w_k^T x_i + y_i x_i^T w_k + y_i^2 x_i^T x_i$

$\begin{cases} w_k^T w_k = \|w_k\|_2^2 \\ x_i^T x_i = \|x_i\|_2^2 \end{cases}$

Because $x_i^T w_k = w_k^T x_i$, $y_i^2 = 1$, and $\|x_i\|_2 \le R$:

$\|w_{k+1}\|_2^2 = \|w_k\|_2^2 + 2 y_i w_k^T x_i + \|x_i\|_2^2 \le \|w_k\|_2^2 + R^2$

1.3) From problem 1.1, we know that $w_{k+1}^T w_{opt} \geq w_k^T w_{opt} + \gamma \|w_{opt}\|_2$

By accumulating M mistakes, we get:

$$w_{k+1}^T w_{opt} \geq w_0^T w_{opt} + \gamma M \|w_{opt}\|_2 \quad *$$

Since w_0 is an all-zero vector:

$$w_{k+1}^T w_{opt} \geq \gamma M \|w_{opt}\|_2$$

According to the Cauchy-Schwartz inequality:

$a = w_{k+1}$, $b = w_{opt}$

$$\gamma M \|w_{opt}\|_2 \leq w_{k+1}^T w_{opt} \leq \|w_{k+1}\|_2 \|w_{opt}\|_2$$

$$\Rightarrow \gamma M \cancel{\|w_{opt}\|_2} \leq \|w_{k+1}\|_2 \cancel{\|w_{opt}\|_2} \Rightarrow \underline{\gamma M \leq \|w_{k+1}\|_2}$$

By applying the result from problem 1.2, we get $*$

$$\|w_{k+1}\|_2^2 \leq \|w_k\|_2^2 + R^2 \Rightarrow \|w_{k+1}\|_2^2 \leq \|w_0\|_2^2 + R^2 M$$

Since w_0 is an all-zero vector:

$$\|w_{k+1}\|_2^2 \leq R^2 M \Rightarrow \underline{\|w_{k+1}\|_2 \leq R\sqrt{M}}$$

By combining the results above, we can conclude that:

$$\gamma M \leq \|w_{k+1}\|_2 \leq R\sqrt{M}$$

* side note (more steps if needed)

$$w_{k+1}^T w_{opt} \geq w_k^T w_{opt} + \gamma \|w_{opt}\|_2 =$$

$$\geq (w_{k-1}^T w_{opt} + \gamma \|w_{opt}\|_2) + \gamma \|w_{opt}\|$$

$$\geq [(w_{k-2}^T w_{opt} + \gamma \|w_{opt}\|_2) + \gamma \|w_{opt}\|_2] + \gamma \|w_{opt}\|_2$$

$$\vdots$$

$$\geq w_0^T w_{opt} + \gamma M \|w_{opt}\|_2$$

* side NOTE (more steps if needed)

From 1.2 $\|w_{i+1}\|_2^2 \le \|w_i\|_2^2 + R^2$

$i = 0$ $\|w_1\|_2^2 \le \|w_0\|_2^2 + R^2$

$i = 1$ $\|w_2\|_2^2 \le \|w_1\|_2^2 + R^2 \le \left(\|w_0\|_2^2 + R^2\right) + R^2 = \|w_0\|_2^2 + 2R^2$

$i = 2$ $\|w_3\|_2^2 \le \|w_2\|_2^2 + R^2 \le \left(\|w_1\|_2^2 + R^2\right) + R^2 \le \left[\left(\|w_0\|_2^2 + R^2\right) + R^2\right] + R^2$

$= \|w_0\|_2^2 + 3R^2$

$i = M-1$ $\|w_M\|_2^2 \le \|w_0\|_2^2 + MR^2$ (M mistakes)

$i = k$ $\|w_{k+1}\|_2^2 \le \|w_0\|_2^2 + MR^2$ $\because w_{i+1} = w_i$ for $i \ge M$

Q 1.4)

a. given $\gamma\mu \leq \|w_{k+1}\|_2 \leq R\sqrt{\mu}$,

square sides

$$\gamma^2\mu^2 \leq \|w_{k+1}\|_2^2 \leq R^2\mu$$

the $\leq$ sign does not change

b. since $\gamma > 0 \Rightarrow \gamma > 0$, now we divide all sides by γ^2

$$\mu^2 \leq \frac{\|w_{k+1}\|_2^2}{\gamma^2} \leq \frac{R^2\mu}{\gamma^2}$$

c. divide by μ all sides

$$\mu \leq \frac{\|w_{k+1}\|_2^2}{\gamma^2\mu} \leq \frac{R^2}{\gamma^2}$$

we can have $\Rightarrow \therefore$

$$\mu \leq \frac{R^2}{\gamma^2}$$

2.1) Let $B(a_1, b_1, a_2, b_2)$ be the smallest rectangle returned by function f and B contains all positive examples in the training set S.

By using the realizability assumption, there exists a function f^* that correctly classifies all datapoints in the distribution D, meaning that $R(f^*) = 0$. Since the training set S is a subset of the dataset, $\hat{R}(f^*)$ is also zero. To prove f is an empirical risk minimizer, show that $\hat{R}(f) = 0$.

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(f(x_i) \neq \underbrace{y_i}_{= f^*(x_i)}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(f(x_i) \neq f^*(x_i)\right)$$

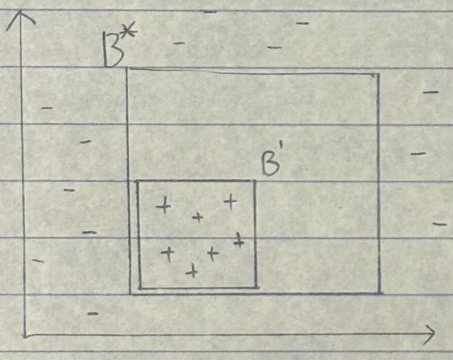
where n is the size of the training set S. Let n_{pos} be the # of mismatches for positive datapoints and n_{neg} for negative datapoints. $\because$ all positive examples are enclosed within B $\therefore n_{pos} = 0$

$$\hat{R}(f) = \frac{1}{n}\left(n_{pos}\mathbb{I}(f(x_i) \neq 1) + n_{neg}\mathbb{I}(f(x_j) \neq 0)\right)$$

$\because f^*$ exists $\therefore$ no negative example exists within $B \Rightarrow n_{neg} = 0$

In conclusion, $\hat{R}(f) = \frac{1}{n}(0+0) = 0$, meaning that f is an ERM.

2.2) Assume the area of B^* is $\text{Area}(B^*)$ and the area outside of B^* is $\text{Area}(B^*_{out})$. Next, assume each datapoint is evenly distributed in the distribution D, the probability of picking a positive example is $P_{pos} = \frac{\text{Area}(B^*)}{\text{Area}(B^*) + \text{Area}(B^*_{out})} > 0$



Let B' be the rectangle returned by $f_{S'}^{ERM}$ and $\text{Area}(B') = \alpha \text{Area}(B^*)$, where $0 < \alpha \leq 1$.

Prove $R(f_{S'}^{ERM}) \geq 0.5$ is possible

$$R(f_{S'}^{ERM}) = \frac{\text{Area}(B^*) - \text{Area}(B')}{\text{Area}(B^*) + \text{Area}(B^*_{out})} = \frac{(1-\alpha) \times \text{Area}(B^*)}{\text{Area}(B^*) + \text{Area}(B^*_{out})}$$

$$= (1-\alpha) P_{pos}$$

$\because 0 < \alpha \leq 1$ and $P_{pos} > 0$ $\therefore R(f_{S'}^{ERM}) \geq 0.5$ is possible

2.3)

Training S $\rightarrow$ f_S^{ERM} $\rightarrow$ gets a small error

if $n \geq \frac{4\log(4/\delta)}{\epsilon}$ $\rightarrow$ prob at least: $R(f_S^{ERM}) \leq \epsilon$

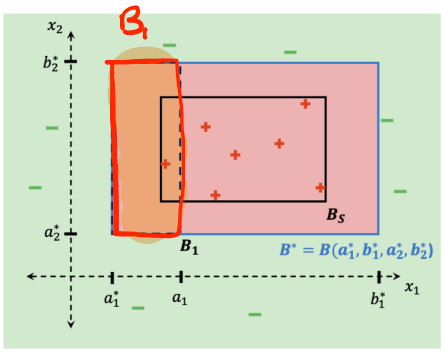
Assumption:

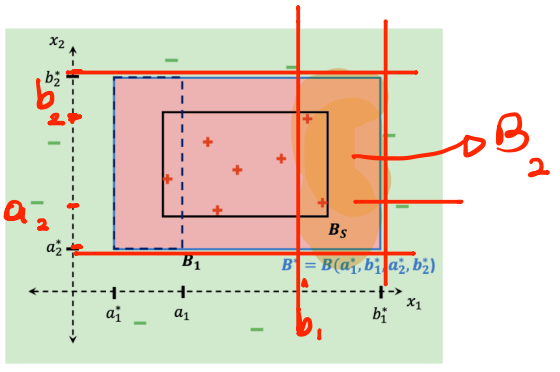
★ $a_1 \geq a_1^*$ $\rightarrow$ $P(B_1) = \frac{\epsilon}{4}$, $B_1 = B(a_1^*, a_1, a_2^*, b_2^*)$

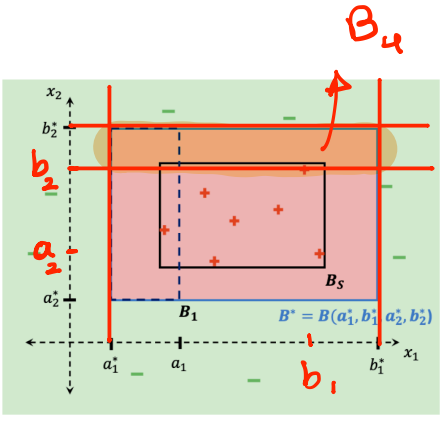
$\begin{cases} b_1 \leq b_1^* \\ a_2 \geq a_2^* \\ b_2 \leq b_2^* \end{cases}$ $\rightarrow$ $P(B_2) = \frac{\epsilon}{4}$, $B_2 = B(b_1, b_1^*, a_2^*, b_2^*)$

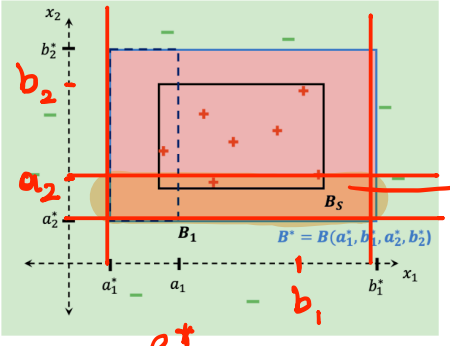
B_S $\rightarrow$ rectangle on training $\downarrow$ f_S^{ERM} ν $\rightarrow$ $P(B_3) = \frac{\epsilon}{4}$, $B_3 = B(a_1^*, b_1^*, a_2^*, a_2)$

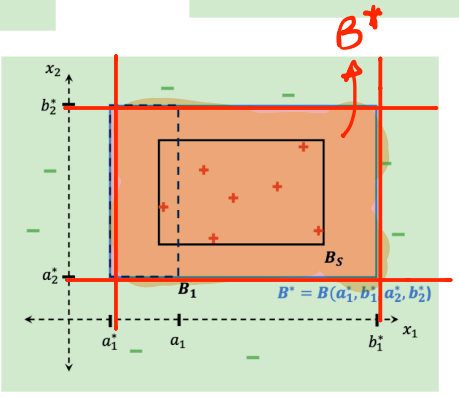
ν $\rightarrow$ $P(B_4) = \frac{\epsilon}{4}$, $B_4 = B(a_1^*, b_1^*, b_2, b_2^*)$











$B^* = B(a_1^*, b_1^*, a_2^*, b_2^*)$

for $f(a_1^*, b_1^*, a_2^*, b_2^*)$

① a) Assume we have a distribution D for (x, y) and $B^* = B(a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle for $f(a_1^*, b_1^*, a_2^*, b_2^*)$. By the realizability assumption, B^* perfectly seperates ⊕ and ⊖ cases.

b) we showed (earlier section) B_s returns all positive examples and it is the smallest rectangle.

c) According to the steps a and b which is given or proved in other sections $\Rightarrow \therefore B_s \subseteq B^*$

② a) According to given conditions (*) above all B_1, B_2, B_3, B_4 are inside B^* and we showed $B_s \subseteq B^*$

b) If S contains positive examples in all rectangles B_1, B_2, B_3, B_4, then B_s will contain all positive examples in B_1, B_2, B_3, B_4.

c) we know mass prob of each $B_\#$ is $\frac{\epsilon}{4}$.

d) from steps a, b, c $\Rightarrow R(f_i) \leq \frac{\epsilon}{4} \quad \forall i \in \{1,2,3,4\}$

$\Rightarrow R(f_s^{ERM}) \leq 4(\frac{\epsilon}{4}) \Rightarrow \therefore R(f_s^{ERM}) \leq \epsilon$ *

Probibility sample belong $B^* - B_s$ set diff

Probility of sample Belong to $B_1 \cup B_2 \cup B_3 \cup B_4$ $\sum_{i=1}^{4} \frac{\epsilon}{4} = 4(\frac{\epsilon}{4})$

(3)

a) $P(\text{sample has No Point from any } B_i) = 1 - P(\text{sample} \in B_i)$

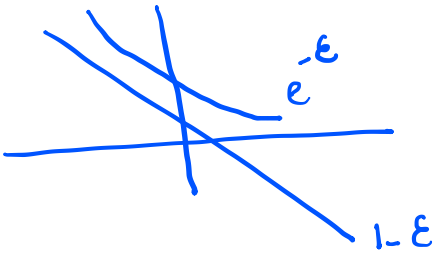
$= 1 - \left(\frac{\varepsilon}{4}\right)$

b) expanding for all rectangles B_i, $i \in \{1,2,3,4\}$

$P(\text{sample} \overset{\text{bad}}{\notin} \text{for all rectangles}) = \left(1 - \frac{\varepsilon}{4}\right)^n \leq e^{-\frac{\varepsilon}{4}n}$

we know $1 - E \leq e^{-E}$ ∴

$E = \frac{\varepsilon}{4}$



④ a) we know $R(f_s^{ERM}) \leq \varepsilon$ and know a bad sample may be found in S.

b) $p(R(f_s^{ERM}) \leq \varepsilon) = p(S \text{ is a good sample})$

$= 1 - p(S \text{ is bad sample})$

$\geq 1 - \sum_{i=1}^{4} e^{-\frac{\varepsilon}{4}n}$

$\geq 1 - 4e^{-\frac{\varepsilon}{4}n}$

side note:

on side we know $P_r[B_1 \cup B_2 \cup B_3 \cup B_4] \leq P_r \sum_{i=1}^{4} P_r(B_i)$

c) we want to show $p(R(f_s^{ERM}) < \varepsilon))$ is at least $1-\delta$

$p(R(f_s^{ERM}) \leq \varepsilon)) \geq 1-\delta$

$1 - 4(e^{\frac{-n\varepsilon}{4}}) \geq 1-\delta$ multiply both sides by (-1)

$4(e^{\frac{-n\varepsilon}{4}}) \leq \delta$

$e^{\frac{n\varepsilon}{4}} \geq \frac{4}{\delta}$ take ln both sides

$n(\frac{\varepsilon}{4}) \geq \ln(\frac{4}{\delta})$

$\therefore \quad n \geq \frac{4\ln(4/\delta)}{\varepsilon}$

Q3

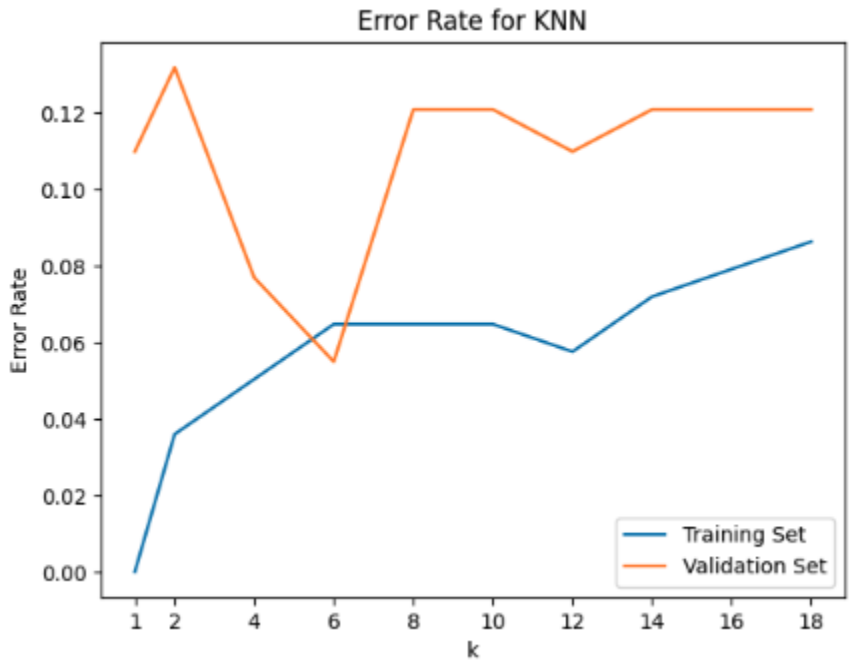
Print statements from code:

The validation error rate is 0.07692307692307693 in Problem **Set 1.1**

The validation error rate is 0.04395604395604396 in Problem Set **1.2 when using normalization**

The validation error rate is 0.04395604395604396 in Problem Set **1.2 when using minmax_scaling**

The validation error rate is 0.04395604395604396 in Problem Set **1.3, which use cosine distance**



In Problem Set 1.4, we use the best k = 6 with the best validation error rate 0.054945054945054944

Using the best k, the final test error rate is 0.07100591715976332

3.4 report:

(1) Report and draw a curve based on the error rate of your model on the training set for each k. What do you observe? (2pts)

a. When k = 1, we observe the lowest error.

b. As k increases and finding the k-nearest neighbors (knn) becomes more complex, the error rate is higher compared to the first few iterations.

c. There is a drop at k = 12. For $6 \leq k \leq 20$, k = 12 exhibits the lowest error.

d. The error line remains flat between k = 6 and k = 10.

e. As the value of k increases in KNN, the model may become less sensitive to local patterns and more influenced by global trends. This could impact the interpretability and generalization of the model.

g. When k=0, the error is the lowest equal to zero. However, this is not what we are looking for.

(2) Report and draw a curve based on the error rate of your model on the validation set for each k. What is your best k? (2pts) \n

a. The best k is 6 having the lowest Error Rate.

b. For $8 \leq k \leq 10$ and $14 \leq k \leq 18$ the error rate stays the same. Potentially, we can have an early stop after 10 if the error rate does not improve.

c. From going k=2 to k=6, the error rate improves

(3) What do you observe by comparing the difference between the two curves? (2pts) \n

a. Vladiation graph may provide better insight into the decision-making process leading to the conclusion than training graph.

b. The training curve tends to increase for 2<=k<=6 whereas the validation curve has has a drop.

d. A similar curve pattern has been observed for both training and validation curves when 8<=k<=14. However, training and validation and each k have a different error rate.

e. For k=6 (the best k - question 2) the validation error rate is lower than training error rate.

f. broadly speaking, as k increases in the training set, the error rate increrases or remains the same.

g. In the Validation graph, we have a convex function for k between 2 and 8. k=6 is the global minimum.

h. Both validation and training have a convex function for k between 10 and 14, with a local min of k=12

(4) What is the final test set error rate you get using your best-k? (1pt) \n

a. In Problem Set 1.4, we use the best k = 6 with the best validation error rate 0.054945054945054944

Using the best k, the final test error rate is 0.07100591715976332

(5) Comment on these results from the perspective of overfitting, generalization and hyper-parameter tuning. (3pts).\n

Overfitting:

When k=1, the error rate on the training set is low and higher on the validation set, this shows the model does not perform well on unseen data due to overfitting. \n

For small values of k, such as k=1 in KNN, the prediction depends solely on the nearest neighbor, which can lead to overfitting. The model becomes highly sensitive to noise in the training data, capturing local fluctuations rather than the true underlying structure of the data. \n

Generalization:

For k=6, the training and validation error rates are very close. This shows the model is generalized enough to perform the same for unseen data.

k=6 is less sensitive to a single point of data and still preserves the local pattern in the data.

k=6 provides a good balance between overfitting and underfitting in terms of generalization.

hyper-parameter tuning:

Using different Ks to find the best k is a hyper parameter tuning which we implemented in this problem. As a result, we found the best k is 6, which has the lowest error rate. \n

We learned a hyperparameter tuning is a necessary step for finding the best parameters in a classification problem.

It is crucial to have validation set in hyperparameter tuning to avoid overfitting and detect cases such as k=1, as mentioned above.

Q4

4.1

F(w_LS)= 217.48452613174004 on training data
F(w_0)= 78885.82819617869 on training data
F(w_LS)= 294.06836989399164 on testing data

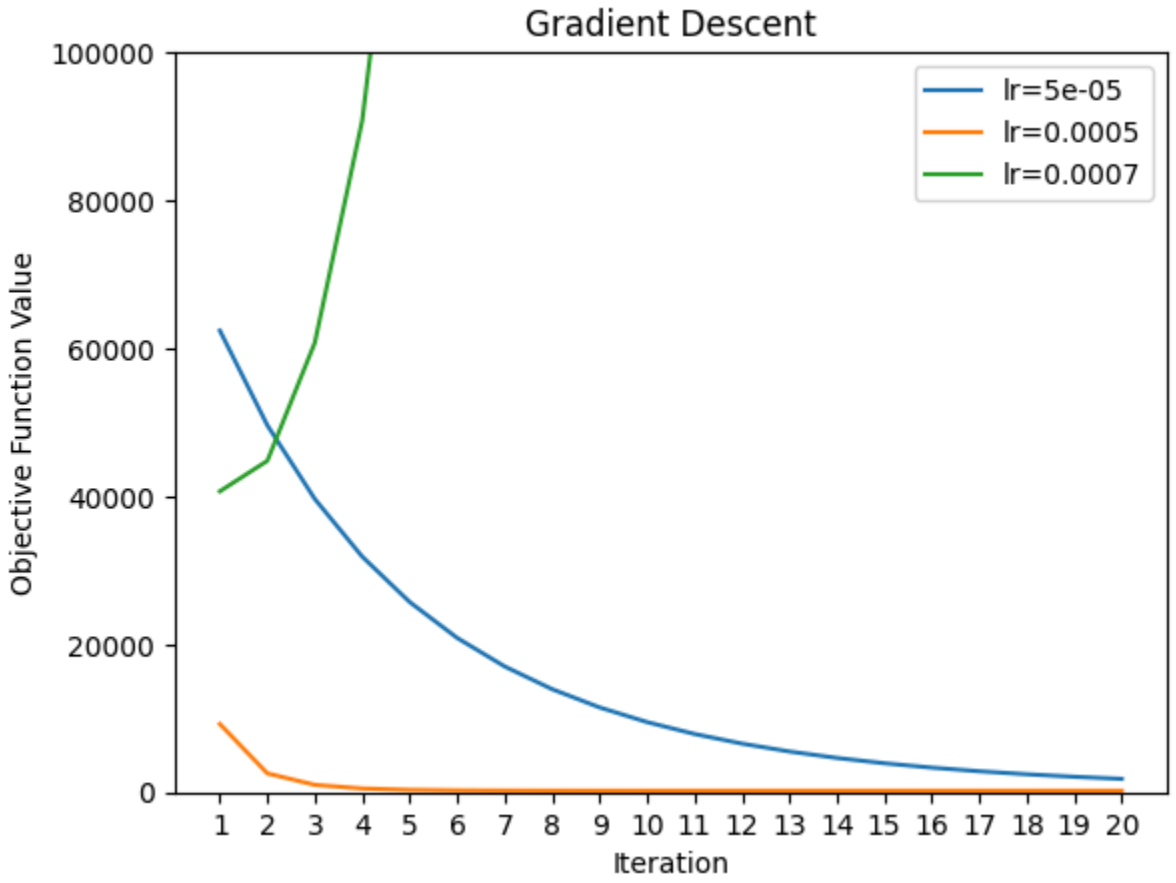
Gap: 294.068 - 217.485 = 76.583

The test objective function value with weights w_LS performed much better than setting the weights w0 to a zero vector, showing the effectiveness of w_LS in minimizing the residual sum of squares (RSS). However, there was still a gap between the training and test data, suggesting that the model with weights w_LS may be slightly overfitting the training data.

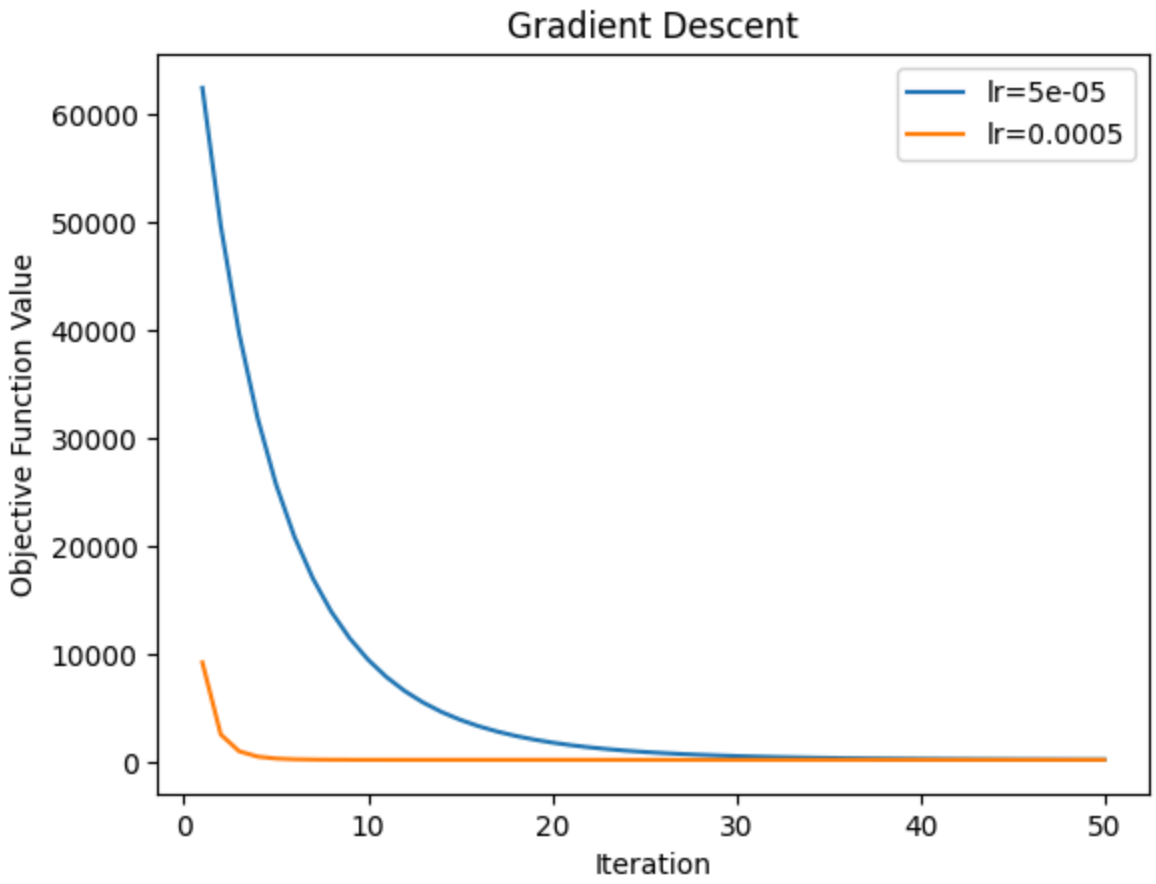
4.2

A larger step size led to faster convergence in gradient descent. For instance, with a step size of 0.0005, convergence occurred in approximately seven iterations. In contrast, when the step size was reduced to 0.00005, it took over 30 iterations to approach convergence, and the objective function value remained significantly higher than that of 0.0005. However, surpassing a step size of 0.000582 resulted in a divergence of the objective function value after 100 iterations. In conclusion, the step size should be neither too small nor too large, and increasing the number of iterations does not guarantee a lower objective function value.

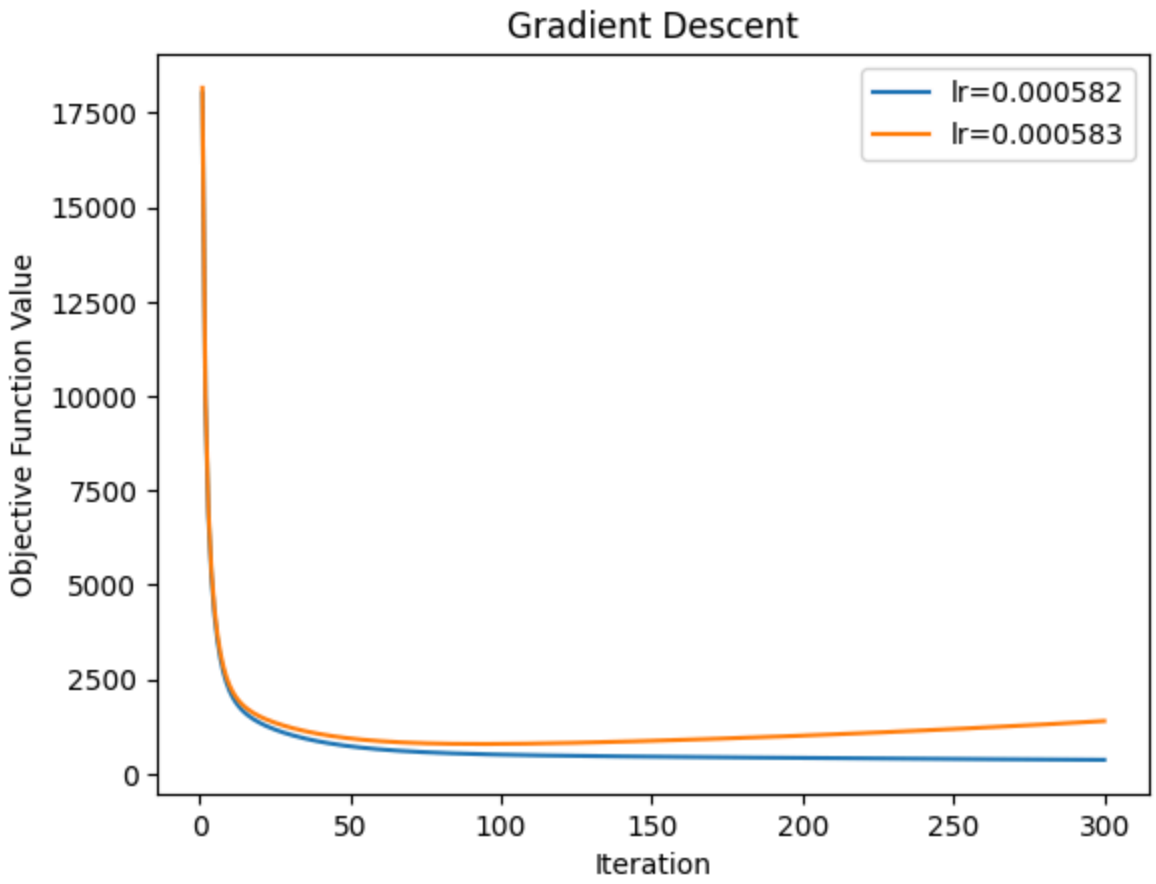
Best final objective value: 217.486, where step size=0.0005



Comparison of gradient descent convergence for various step sizes



The convergence comparison graph for step sizes of 0.00005 and 0.0005



The divergence boundary with 300 iterations

4.3

Like gradient descent, stochastic gradient descent showed faster convergence with a larger step size, and the optimal step size should be neither too small nor too large. However, stochastic gradient descent had an erratic convergence behavior because only one random data point was used in each iteration instead of the entire data set. Also, we found that SGD required more iterations than GD to converge because randomly choosing a data point for each iteration resulted in noisy gradient estimates.

Best final objective value: 443.185, where step size = 0.005