CSCI567: Machine Learning	USC, Fall 2022
Homework 1	
Instructor: Vatsal Sharan	*Due: September 14 by 2:00 pm PST*

We would like to thank previous 567 staff, and Gregory Valiant (Stanford) for kindly sharing many of the problems with us.

A reminder on collaboration policy and academic integrity: Our goal is to maintain an optimal learning environment. You can discuss the homework problems at a high level with other groups, but you should not look at any other group's solutions. Trying to find solutions online or from any other sources for any homework or project is prohibited, will result in zero grade and will be reported. To prevent any future plagiarism, uploading any material from the course (your solutions, quizzes etc.) on the internet is prohibited, and any violations will also be reported. Please be considerate, and help us help everyone get the best out of this course.

Please remember the Student Conduct Code (Section 11.00 of the USC Student Guidebook). General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty.

Instructions

We recommend that you use LaTeX to write up your homework solution. However, you can also scan handwritten notes. The homework will need to be submitted on D2L. We will announce detailed submission instructions later.

Theory-based Questions

Problem 1: Perceptron Convergence (15pts)

Recall the perceptron algorithm that we saw in class. The perceptron algorithm comes with strong theory, and you will explore some of this theory in this problem. We begin with some remarks related to notation which are valid throughout the homework. Unless stated otherwise, scalars are denoted by small letters in normal font, vectors are denoted by small letters in bold font, and matrices are denoted by capital letters in bold font.

Problem 1 asks you to show that when the two classes in a binary classification problem are linearly separable, then the perceptron algorithm will *converge*. For the sake of this problem, we define convergence as predicting the labels of all training instances perfectly. The perceptron algorithm is described in Algorithm 1. It gets access to a dataset of n instances $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. It outputs a linear classifier $\mathbf{w}$.

Algorithm 1 Perceptron

while not converged **do**
  Pick a data point $(\mathbf{x}_i, y_i)$ randomly
  Make a prediction $\hat{y} = \text{sgn}(\mathbf{w}^T\mathbf{x}_i)$ using current $\mathbf{w}$
  if $\hat{y} \neq y_i$ **then**
    $\mathbf{w} \leftarrow \mathbf{w} + y_i\mathbf{x}_i$
  end if
end while

Assume there exists an optimal hyperplane $\mathbf{w}_{\text{opt}}$, $\|\mathbf{w}_{\text{opt}}\|_2 = 1$ and some $\gamma > 0$ such that $y_i(\mathbf{w}_{\text{opt}}^T\mathbf{x}_i) \geq \gamma, \forall i \in \{1, 2, \ldots, n\}$. Additionally, assume $\|\mathbf{x}_i\|_2 \leq R, \forall i \in \{1, 2, ..., n\}$. Following the steps below, show that the perceptron algorithm makes at most $\dfrac{R^2}{\gamma^2}$ errors, and therefore the algorithm must converge.

1.1 (5pts) Show that if the algorithm makes a mistake, the update rule moves it towards the direction of the optimal weights $\mathbf{w}_{\text{opt}}$. Specifically, denoting explicitly the updating iteration index by k, the current weight vector by $\mathbf{w}_k$, and the updated weight vector by $\mathbf{w}_{k+1}$, show that, if $y_i(\mathbf{w}_k^T \mathbf{x}_i) < 0$, we have

$$\mathbf{w}_{k+1}^T \mathbf{w}_{\text{opt}} \geq \mathbf{w}_k^T \mathbf{w}_{\text{opt}} + \gamma \|\mathbf{w}_{\text{opt}}\|_2. \tag{1}$$

Ans. Because of the perceptron algorithm's update rule, we know

$$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i \tag{2}$$

Now, we take the inner product of both sides of Eq. 2 with $\mathbf{w}_{\text{opt}}$,

$$\mathbf{w}_{k+1}^T \mathbf{w}_{\text{opt}} = \mathbf{w}_k^T \mathbf{w}_{\text{opt}} + y_i \mathbf{x}_i^T \mathbf{w}_{\text{opt}} \tag{3}$$

By the assumption we have for $\mathbf{w}_{\text{opt}}$, we get

$$\begin{aligned} \mathbf{w}_{k+1}^T \mathbf{w}_{\text{opt}} &\geq \mathbf{w}_k^T \mathbf{w}_{\text{opt}} + \gamma \\ &= \mathbf{w}_k^T \mathbf{w}_{\text{opt}} + \gamma \|\mathbf{w}_{\text{opt}}\|_2 \textit{ (... Since } \|\mathbf{w}_{\text{opt}}\|_2 = 1\textit{)} \end{aligned} \tag{4}$$

Rubric: (1) Get the idea of using Eq. 2 (2pt). (2) Get the remaining proof correct (3pt).

1.2 (4pts) Show that the length of updated weights does not increase by a large amount. In particular, show that if $y_i(\mathbf{w}_k^T \mathbf{x}_i) < 0$, then

$$\|\mathbf{w}_{k+1}\|_2^2 \leq \|\mathbf{w}_k\|_2^2 + R^2. \tag{5}$$

Ans.

$$\begin{aligned} \|\mathbf{w}_{k+1}\|_2^2 &= \mathbf{w}_{k+1}^T \mathbf{w}_{k+1} \\ &= (\mathbf{w}_k + y_i \mathbf{x}_i)^T (\mathbf{w}_k + y_i \mathbf{x}_i) \textit{ (... Using Eq. 2)} \\ &= \|\mathbf{w}_k\|_2^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + y_i^2 \mathbf{x}_i^T \mathbf{x}_i \\ &\leq \|\mathbf{w}_k\|_2^2 + 2 y_i \mathbf{w}_k^T \mathbf{x}_i + R^2 \textit{ (... Since } \|\mathbf{x}_i\|_2 \leq R, \forall i \in \{1, 2, ..., n\}\textit{)} \\ &\leq \|\mathbf{w}_k\|_2^2 + R^2 \textit{ (... Since } y_i \mathbf{w}_k^T \mathbf{x}_i < 0\textit{)} \end{aligned} \tag{6}$$

Rubric: (1) Get the idea of using Eq. 2 and correctly expand the equation (1pt). (2) Correctly using $\|\mathbf{x}_i\|_2 \leq R$ (1.5pt) (3) Correctly using property of $y_i \mathbf{w}_k^T \mathbf{x}_i < 0$ (1.5pt)

1.3 (5pts) Assume that the initial weight vector $\mathbf{w}_0 = \mathbf{0}$ (an all-zero vector). Using results from the previous two parts, show that for any iteration $k+1$, with M being the total number of mistakes the algorithm has made for the first k iterations, we have

$$\gamma M \leq \|\mathbf{w}_{k+1}\|_2 \leq R\sqrt{M}. \tag{7}$$

Hint: use the Cauchy-Schwartz inequality: $\mathbf{a}^T \mathbf{b} \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$.
Ans. Here, we repeatedly apply results from **1.1** **1.2** from $k = 1$ to $k = k$. Using **1.1**, we will get:

$$\mathbf{w}_{k+1}^T \mathbf{w}_{\text{opt}} \geq \mathbf{w}_0^T \mathbf{w}_{\text{opt}} + M\gamma \|\mathbf{w}_{\text{opt}}\|_2. \tag{8}$$

Since $\mathbf{w}_0 = \mathbf{0}$, so we get:

$$\mathbf{w}_{k+1}^T \mathbf{w}_{\text{opt}} \geq M\gamma \|\mathbf{w}_{\text{opt}}\|_2. \tag{9}$$

Then, we apply Cauchy-Schwartz inequality to get:

$$M\gamma \|\mathbf{w}_{\text{opt}}\|_2 \leq \|\mathbf{w}_{k+1}\|_2 \|\mathbf{w}_{\text{opt}}\|_2. \tag{10}$$

Then, we are done with the inequality on the left-hand side.
Next, we use **1.2** and the property $\mathbf{w}_0 = \mathbf{0}$ to get,

$$\|\mathbf{w}_{k+1}\|_2^2 \leq \|\mathbf{w}_0\|_2^2 + MR^2 = MR^2. \tag{11}$$

Then, we are done with the inequality on the right-hand side.
Rubric: (1) Inequality on the left-hand side (3pt). (2) Inequality on the right-hand side (2pt).

1.4 (1pts) Use **1.3** to conclude that $M \leq R^2/\gamma^2$. Therefore the algorithm can make at most R^2/γ^2 mistakes (note that there is no direct dependence on the dimensionality of the datapoints).
From **1.3**, we know:

$$\gamma M \leq R\sqrt{M}. \tag{12}$$

We divide $\sqrt{M}$ for both side, and get: $R/\gamma \geq \sqrt{M}$, hence get $R^2/\gamma^2 \geq M$
Rubric: 1pt for this question.

Additional Note:
We can also use another perspective to prove the convergence. The basic idea is that the angle between w_k and w_{opt} will become smaller and smaller when k becomes larger. Additionally, the largest angle we need to "correct" is no larger than 90 degrees. Hence, there's a finite number of updates. To get the angle is become smaller and smaller, we can first calculate the bound of the dot product between w_k and w_{opt} (1.1). Then we use 1.2 to get the squared norm of w_k increases at most linearly in the number of total updates. By combining the two, we can show that the cosine of the angle between w_k and w_{opt} has to decrease by a finite increment due to each update.
Following this idea, we can calculate the cosine between w_k and w_{opt} as:

$$cos(\mathbf{w}_k, \mathbf{w}_{\text{opt}}) = \frac{\mathbf{w}_{k+1}^T \mathbf{w}_{\text{opt}}}{\|\mathbf{w}_{k+1}\|_2 \|\mathbf{w}_{\text{opt}}\|_2} \geq \frac{\mathbf{w}_k^T \mathbf{w}_{\text{opt}} + \gamma \|\mathbf{w}_{\text{opt}}\|_2}{\sqrt{\|\mathbf{w}_k\|_2^2 + R^2} \|\mathbf{w}_{\text{opt}}\|_2} \geq \frac{\mathbf{w}_k^T \mathbf{w}_{\text{opt}} + \gamma \|\mathbf{w}_{\text{opt}}\|_2}{\sqrt{kR^2} \|\mathbf{w}_{\text{opt}}\|_2} \tag{13}$$

Iterating k from $k = 0$ for the numerator, we get:

$$cos(\mathbf{w}_k, \mathbf{w}_{\text{opt}}) \geq \frac{M\gamma \|\mathbf{w}_{\text{opt}}\|_2}{\sqrt{MR^2} \|\mathbf{w}_{\text{opt}}\|_2} \tag{14}$$

Since cosine is bounded by 1, so we can get the same conclusion.

Problem 2: Logistic Regression (10pts)

This problems explores some properties of the logistic function to get you more comfortable with it. Consider the following univariate function first:

$$F(x; A, k, b) = \frac{A}{1 + e^{-k(x-b)}} \tag{15}$$

2.1 (3pts) Describe with words how A, k and b affect or are related to the shape of $F(x; A, k, b)$ by using the plot at `https://www.geogebra.org/graphing/mxw7wp8b`. Note: We are trying to learn the parameters of a logistic function that can fit the dataset. For example, see the figure below and guess which of the four functions fit well.

Ans. A controls the range/height of the function i.e $F \in (0, A)$. k controls the steepness of transition from 0 to A. This is most prominent at the point of inflection $(b, A/2)$. b controls the point of inflection and shifts the graph along the x-axis.
Rubric: One point each explaining roughly the effects of A,b and k.

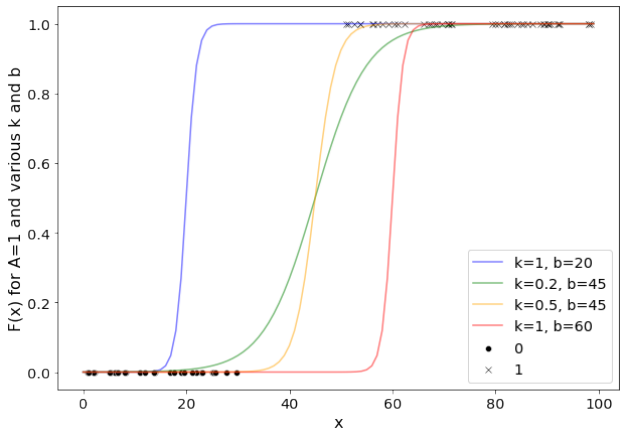

Figure 1: Learning the parameters of logistic function for a dataset.

2.2 (1pt) For what values of A, k and b is $F(x; A, k, b)$ a cumulative distribution function (CDF)?

Ans. For $A = 1$ and $k > 0$.
Rubric: 0.5 points for correct values of A and k.

2.3 (2pts) When $F(x; A, k, b)$ is a CDF, we can interpret $F(x; A, k, b)$ as the probability of x belonging to the class 1 (in a binary classification problem). Suppose we know $F(x; A, k, b)$ and want to predict the label of a datapoint x. We need to decide on a threshold value for $F(x; A, k, b)$, above which we will predict the label 1 and below which we will predict -1. Show that setting the threshold to be $F(x; A, k, b) \geq 1/2$ minimizes the classification error.

The value $x = x_0$ for which $F(x_0) = 0.5$ is called the decision boundary. In the univariate case, it is a point in on the x-axis, in the multivariate case (next question), it is a hyperplane. There is a rich literature on decision theory which studies how to make decisions if we know the probabilities of the underlying events (see `https://mlstory.org/pdf/prediction.pdf` to learn more if you're interested).

Ans. Expected error for a given threshold (T) for a given x is

$$\begin{aligned} \text{Expected loss given } x &= p(y = 0|x)I[F(x) \geq T] + p(y = 1|x)I[F(x) < T] \\ &= [1 - F(x)]I[F(x) \geq T] + [F(x)]I[F(x) < T] \end{aligned}$$

The above is minimized at $T = 0.5$. You can also see this more informaly as follows. Let us compare the two cases -
caseA - when threshold is, say, 0.7 and

caseB - threshold is 0.5.
For any the point with $F(x) > 0.7$, both cases predict label $\hat{y} = 1$.
The only case where things are different between the two cases is when $0.5 < F(x) < 0.7$. In those cases, the expected loss given some x using case A is more than case B since A predicts $\hat{y} = 0$ although the probability $p(y = 1|x) = 0.6$. So the expected loss given x on case A is 0.6 and on case B is 0.4. So case B is better. A similar argument can be made for caseC (threshold is, say, 0.3) i.e caseB is better than caseC.

Rubric: 1pt for correct threshold, 1 point for roughly correct explanation.

2.4 (4pts) We now consider the multivariate version. Consider the function:

$$F(\mathbf{x}; \mathbf{w}, b) = \frac{1}{1 + e^{-\mathbf{w^T x}+b}}. \tag{16}$$

You can explore a 2D version of this function at `https://www.geogebra.org/3d/g9fjtdeh` where $\mathbf{w} = (w_1, w_2)$. Match the items in the left column with those on the right. It will help if you can get the expression for the gradient of $F(\mathbf{x}; \mathbf{w}, b)$ w.r.t $\mathbf{x}$. Provide a short explanation of your answers.

1) level sets of $F(\mathbf{x}; \mathbf{w}, b)$ 2) direction of gradient at any point 3) distance of the level set $F(\mathbf{x}) = 1/2$ from the origin	a) parallel to $\mathbf{w}$ b) $\|\|\mathbf{w}\|\|/4$ c) $\|b\|/\|\|\mathbf{w}\|\|$ d) $\|b\|$ e) orthogonal to $\mathbf{w}$

Ans (1,e), (2,a), (3,c)

$$\nabla \mathbf{F}(\mathbf{x}) = \frac{e^{-\mathbf{w^T x}+b}}{(1 + e^{-\mathbf{w^T x}+b})^2}\ \mathbf{w} \tag{17}$$

Clearly the gradient $\nabla \mathbf{F}$ has the same direction as $\mathbf{w}$. We also know that the level sets are orthogonal to the gradient since otherwise we could move along the projection of gradient on the level set and still change the function violating definition of level sets. For 3),

$$\begin{aligned} F(x) = \frac{1}{1 + e^{-\mathbf{w^T x}+b}} &= 1/2 \\ \implies e^{-\mathbf{w^T x}+b} &= 1 \\ \implies \mathbf{w}^T x - b &= 0 \\ \implies \hat{\mathbf{w}}^T x - b/||\mathbf{w}|| &= 0 \end{aligned}$$

Distance of a point $\mathbf{x_0}$ from the above line L is given by

$$d(L, x_0) = |\ \mathbf{w}^T x_0 - b/||\mathbf{w}||\ |$$

Plugging $\mathbf{x_0} = \mathbf{0}$ gives us $\frac{|b|}{||\mathbf{w}||}$.
Rubrics: 0.5pt for each correct match and 0.5pt for correct explanation. 1pt bonus if all 3 correct match and correct explanation.

Problem 3: Learning rectangles (15pts)

An axis aligned rectangle classifier is a classifier than assigns the value 1 to a point if and only if it is inside a certain rectangle. Formally, given real numbers $a_1 \leq b_1, a_2 \leq b_2$, define the classifier $f_{(a_1,b_1,a_2,b_2)}$ on an input $\mathbf{x}$ with coordinates (x_1, x_2) by

$$f_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ -1 & \text{otherwise.} \end{cases}$$

The function class of all axis-aligned rectangles in the plane is defined as

$$\mathcal{F}^2_{\text{rec}} = \{f_{(a_1,b_1,a_2,b_2)}(x_1, x_2) : a_1 \leq b_1, a_2 \leq b_2\}.$$

We will assume that the true labels y of the datapoints $(\mathbf{x}, y)$ are given by some axis-aligned rectangle (this is the realizability assumption discussed in class). The goal of this question is to come up with an algorithm which gets small classification error with respect to any distribution D over $(\mathbf{x}, y)$ with good probability.

The loss function we use throughout the question is the 0-1 loss. It will be convenient to denote a rectangle marked by corners (a_1, b_1, a_2, b_2) as $B(a_1, b_1, a_2, b_2)$. Let $B^* = B(a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle corresponding to the function $f_{(a_1^*,b_1^*,a_2^*,b_2^*)}$ which labels the datapoints. Let $S = \{(\mathbf{x}_i, y_i), i \in [n]\}$ be a training set of n samples drawn i.i.d. from D. Please see Fig. 2 for an example.

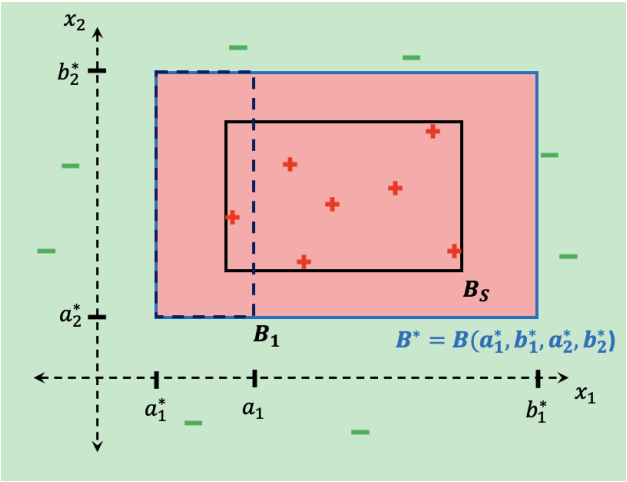

Figure 2: Learning axis-aligned rectangles in two dimensions, here $+$ (red) denotes datapoints in training set S with label 1 and $-$ (green) denotes datapoints with label -1. The true labels are given by rectangle B^* (solid blue line), with everything outside B^* being labelled negative and inside being labelled positive. B_S (solid black line) is the rectangle learned by the algorithm in Part (a). B_1 (dashed black line) is defined in Part (c).

3.1 (3pts) We will follow the general supervised learning framework from class. Given the 0-1 loss, and the function class of axis-aligned rectangles, we want to find an empirical risk minimizer. Consider the algorithm which returns the smallest rectangle enclosing all positive examples in the training set. Prove that this algorithm is an empirical risk minimizer.

Let X be the set of all points and S be the set of points used for training. Let $A(S)$ be the algorithm that finds the smallest rectangle enclosing all positive examples in S. Let $\mathcal{H}^2_{rec}$ be the function class of axis-aligned rectangles.

By realizability, there exists an $h^* \in \mathcal{H}^2_{\text{rec}}$, such that $\mathcal{R}(h^*) = 0$, and since $S \subseteq X$, $\hat{\mathcal{R}}(h^*) = 0$. By definition, $A(S) = h_{(a_1', b_1', a_2', b_2')}$, where $a_1' = \min_{\substack{(x_1,x_2)\in S \\ h^*(x)=1}} x_1, b_1' = \max_{\substack{(x_1,x_2)\in S \\ h^*(x)=1}} x_1, a_2' = \min_{\substack{(x_1,x_2)\in S \\ h^*(x)=1}} x_2, b_2' = \max_{\substack{(x_1,x_2)\in S \\ h^*(x)=1}} x_2$.

To prove that $A(S)$ is an ERM, we need to show that $\hat{\mathcal{R}}(A(S)) = 0$.

$$\hat{\mathcal{R}}(A(S)) = \frac{1}{|S|}\sum_{i=1}^{|S|} 1(A(S)(x^i) \neq h^*(x^i))$$
$$= \frac{1}{|S|}\left(\sum_{h^*(x^i)=1} 1(A(S)(x^i) \neq 1) + \sum_{h^*(x^i)=0} 1(A(S)(x^i) \neq 0)\right), \quad (18)$$

where x^i is the i^{th} sample (x_1^i, x_2^i). For the samples labelled 1, $x_1^i \in [a_1', b_1']$ and $x_2^i \in [a_2', b_2']$, and by definition, $A(S)$ labels them as 1. Similarly, for the samples labeled 0, $x_1^i \notin [a_1', b_1']$ and $x_2^i \notin [a_2', b_2']$ Hence, (18) evaluates to 0 and $A(S)$ is an ERM.

Rubrics: Full points for correct explanation (why positive points and negative points will be classified correctly). 2 points for incomplete but partially correct argument.

3.2 (2pts) Our next task is to show that the algorithm from the previous part not only does well on the training data, but also gets small classification error with respect to the distribution D. Let B_S be the rectangle returned by the algorithm in the previous part on training set S, and let f_S^{ERM} be the corresponding hypothesis. First, we will convince ourselves that generalization is inherently a probabilistic statement. Let a *bad* training set S' be a training set such that $R(f_{S'}^{ERM}) \geq 0.5$. Pick some simple distribution D and ground-truth rectangle B^*, and give a short explanation for why there is a non-zero probability of seeing a bad training set.

Let there be a uniform distribution over the unit area rectangle $[0,1] \times [0,1]$. Let B^* be the unit rectangle itself. Now, any sample S' for which area of $B_{S'} = A(S')$ is less than $\frac{1}{2}$ is a bad sample. The probability of selecting such a sample S' is finite and nonzero. For example, S' could contain n points from $[0, 0.25] \times [0, 0.25]$ with probability $(\frac{1}{16})^n$.

$$E[R(f_{S'}^{ERM})] = E[I[B_{S'}(x) \neq B^*(x)]] = P(B_{S'}(x) \neq B^*(x)) = P(x \in B^* \setminus B_{S'}) = \frac{Area(B^*) - Area(B_{S'})}{Area(B^*)}$$

Rubrics: Full points for currently describing some choice of D and B^*, and some explanation (need not be fully formal).

3.3 (5pts) We will now show that with good probability over the training dataset S, f_S^{ERM} does get small error. Show that if $n \geq \frac{4\log(4/\delta)}{\epsilon}$ then with probability at least $1-\delta$, $R(f_S^{ERM}) \leq \epsilon$.

The basic intuition here is: if we want the risk for the algorithm (the algorithm is an ERM) on your training data to be small enough with a certain confidence, we need to sample enough points. To consider the risk, we actually need to consider the boundary between B_S and B_{S*}. The more sample we get, the higher chance that the gap is small enough. And below, we will start by measuring the relationship between the risk and the "acceptable boundary" of the learned rectangle for a given ϵ. Then, we will explore the sample size n required by the learner to actually hit the "acceptable boundary" with a reasonably high probability $1-\delta$.

To prove this follow the following steps. Let $a_1 \geq a_1^*$ be such that the probability mass (with respect to D) of the rectangle $B_1 = B(a_1^*, a_1, a_2^*, b_2^*)$ is exactly $\epsilon/4$. Similarly, let b_1, a_2, b_2 be numbers such that the probability mass (with respect to D) of the rectangles $B_2 = B(b_1, b_1^*, a_2^*, b_2^*), B_3 = B(a_1^*, b_1^*, a_2^*, a_2), B_4 = B(a_1^*, b_1^*, b_2, b_2^*)$ are all exactly $\epsilon/4$.

- Show that $B_S \subseteq B^*$.
  As $a_1' = \min_{\substack{(x_1,x_2)\in S \\ h^*(x)=1}} x_1 \geq \min_{\substack{(x_1,x_2)\in X \\ h^*(x)=1}} x_1 = a_1^*$, $b_1' = \max_{\substack{(x_1,x_2)\in S \\ h^*(x)=1}} x_1 \leq \max_{\substack{(x_1,x_2)\in X \\ h^*(x)=1}} x_1 = b_1^*$, $a_2' = \min_{\substack{(x_1,x_2)\in S \\ h^*(x)=1}} x_2 \geq \min_{\substack{(x_1,x_2)\in X \\ h^*(x)=1}} x_2 = a_2^*$, $b_2' = \max_{\substack{(x_1,x_2)\in S \\ h^*(x)=1}} x_2 \leq \max_{\substack{(x_1,x_2)\in X \\ h^*(x)=1}} x_2 = b_2^*$, hence $B_S \subseteq B^*$.

- Show that if S contains (positive) examples in all of the rectangles B_1, B_2, B_3, B_4 then $R(f_S^{ERM}) \leq \epsilon$.
  If S contains positive samples in all of the rectangles $B_i, i = 1, 2, 3, 4$, then $B^* = B_1 \cup B_2 \cup B_3 \cup B_4 \cup B_S$

$$\begin{aligned} R(f_S^{ERM}) &= E[I[B_S(x) \neq B^*(x)]] = P(B_S(x) \neq B^*(x)) \\ &= P(x \in B^* \setminus B_S) \leq P(x \in B_1 \cup B_2 \cup B_3 \cup B_4) \\ &\leq \Sigma_{i=1}^4 P(x \in B_i) \\ &= 4 \cdot \frac{\epsilon}{4} = \epsilon \end{aligned}$$

  Rubrics: 1pt for a formal correct argument involving union bound.
  0.5pt for an informal but correct argument.

- For each $i \in \{1, \ldots, 4\}$ upper bound the probability that S does not contain an example from B_i.
  Let there be n samples in S. Let event E_i mean sample S contains no points from B_i. We know that for a sample point x,

$$P(E_i) = P(x \notin B_i) = 1 - \frac{\epsilon}{4}$$

  Hence,

$$\begin{aligned} P(E_i) &= P(B_S \cap B_i = \phi) = P(\forall x \in S, x \notin B_i) \\ &= (1 - \frac{\epsilon}{4})^n \\ &\leq e^{-n\epsilon/4} \\ &(\because 1 - x \leq e^{-x}) \end{aligned}$$

  Rubrics: 1pt for formal and correct proof, 0.5pt for informal and correct proof or partially correct answer

- Use the union bound to conclude the argument.
  From previous parts, we know that $R(f_S^{ERM}) \leq \epsilon$ when we have a sample S which contains points from all of B_1, B_2, B_3, B_4, i.e we have a good sample. But we cannot guarantee $R(f_S^{ERM}) \leq \epsilon$ if S does not contain a point from one or more of B_1, B_2, B_3, B_4, which means we have a bad sample.

$$\begin{aligned} P(R(f_S^{ERM}) \leq \epsilon) &= P(\text{ } S \text{ is a good sample}) \\ &= P(S \text{ contains points from all of } B_1, B_2, B_3, B_4) \\ &= 1 - P(S \text{ is a bad sample}) \\ &= 1 - P(\text{S does not contain a point from } \textit{at least one} \text{ of } B_1 \text{ or } B_2 \text{ or } B_3 \text{ or } B_4) \\ &= 1 - P(E_1 \cup E_2 \cup E_3 \cup E_4) \\ &\geq 1 - \Sigma_{i=1}^4 P(E_i) \\ &\geq 1 - 4(e^{-n\epsilon/4}) \end{aligned}$$

  But we wanted $P(R(f_S^{ERM}) < \epsilon) \geq 1 - \delta$.
  So, we must have,

$$\begin{aligned} 1 - 4(e^{-n\epsilon/4}) &\geq 1 - \delta \\ \implies \delta &\geq 4(e^{-n\epsilon/4}) \\ \implies e^{n\epsilon/4} &\geq 4/\delta \\ \implies \frac{n\epsilon}{4} &\geq \ln(4/\delta) \\ \implies n &\geq \frac{4\ln(4/\delta)}{\epsilon}. \end{aligned}$$

Rubrics: 1pt for good/bad sample argument. 1pt for arguing $P(R(f_S^{ERM})) \geq 1 - 4e^{-n\epsilon/4}$ and beyond. Partial credits for incorrect/incomplete solutions

3.4 (5pts) Repeat the previous question for the function class of axis-aligned rectangles in $\mathbb{R}^d$. To show this, try to generalize all the steps in the above question to higher dimensions, and find the number of training samples n required to guarantee that $R(f_S^{ERM}) \leq \epsilon$ with probability at least $1 - \delta$.
In d-dimensions, let us have $2d$ hyper-rectangles $B_1, B_2, ...B_{2d}$ such that the probability of selecting a sample from any of them be $\frac{\epsilon}{2d}$.

b) If S contains a point from all of the B_i, then we need to show that $R(f_S^{ERM}) \leq \epsilon$

$$\begin{aligned} R(f_S^{ERM}) = E[I[B_S(x) \neq B^*(x)]] &= P(B_S(x) \neq B^*(x)) \\ &= P(x \notin B^* \setminus B_S) \leq P(x \in \cup_{i=1}^{2d} B_i) \\ &\leq \Sigma_{i=1}^{2d} P(x \in B_i) = 2d \cdot \frac{\epsilon}{2d} = \epsilon \end{aligned}$$

c) Let x be a sample point from S.

$$P(E_i) = P(x \notin B_i) = 1 - \frac{\epsilon}{2d}$$

For x to not lie in any of the B_i, we have

$$P(E_i) = \left(1 - \frac{\epsilon}{2d}\right)^n \leq e^{-n\epsilon/2d}$$

d)

$$\begin{aligned} P(R(f_S^{ERM}) \leq \epsilon) &= P(\text{not bad sample}) \\ = 1 - P(\text{S does not contain point from any } B_i) \\ &\geq 1 - \Sigma_{i=1}^{i=2d} P(E_i) \\ &\geq 1 - \Sigma_{i=1}^{i=2d} e^{-n\epsilon/2d} \\ &= 1 - 2de^{-n\epsilon/2d} \end{aligned}$$

We want $P(R(f_S^{ERM}) \leq \epsilon)$ to be atleast $1 - \delta$. Hence,

$$\begin{aligned} 1 - 2de^{-n\epsilon/2d} &\geq 1 - \delta \\ \implies \delta &\geq 2de^{-n\epsilon/2d} \\ \implies e^{n\epsilon/2d} &\geq 2d/\delta \\ \implies \frac{n\epsilon}{2d} &\geq \ln(2d/\delta) \\ \implies n &\geq \frac{(2d)\ ln(2d/\delta)}{\epsilon} \end{aligned}$$

Rubrics: same as 3.3

Programming-based Questions

Before you start to conduct homework in this part, you need to first set up the coding environment. We use python3 (version $\geq$ 3.7) in our programming-based questions. There are multiple ways you can install python3, for example:

- You can use conda to configure a python3 environment for all programming assignments.
- Alternatively, you can also use virtualenv to configure a python3 environment for all programming assignments

After you have a python3 environment, you will need to install the following python packages:

- numpy
- matplotlib (for you plotting figures)

Note: You are **not allowed** to use other packages, such as *tensorflow, pytorch, keras, scikit-learn, scipy*, etc. to help you implement the algorithms you learned. If you have other package requests, please ask first before using them.

Problem 4: k-nearest neighbor classification and the ML pipeline (25pts)

In class, we talked about how we need to do training/test split to make sure that our model is generalizing well. We also discussed how we should not reuse a test set too much, because otherwise the test accuracy may not be an accurate measure of the accuracy on the data distribution. In reality, a ML model often has many *hyper-parameters* which need to be tuned (we will see an example in this question). We don't want to use the test set over and over to see what the best value of this hyper-parameter is. The solution is to have a third split of the data, and create a *validation set*. The idea is to use the validation set to evaluate results from the training set and tune any hyper-parameters. Then, use the test set to double-check your evaluation after the model has "passed" the validation set. Please see this nice explanation for more discussion: `https://developers.google.com/machine-learning/crash-course/validation/another-partition`.

With this final piece, we are now ready to build a real ML pipeline. It usually conducts three parts. (1) Load and pre-process the data. (2) Train a model on the training set and use the validation set to tune hyper-parameters. (3) Evaluate the final model on the test set and report the result.

In this problem, you will implement the *k-nearest neighbor (k-NN) algorithm* to conduct classification tasks. We provide the bootstrap code and you are expected to complete the classes and functions. You can find the code and data on `https://vatsalsharan.github.io/fall22/knn_hw1.zip`.

k-NN algorithm

The k-nearest neighbor (k-NN) algorithm is one of the simplest machine learning algorithms in the supervised learning paradigm. The idea behind k-NN is simple, and we explain it first for the case of $k = 1$. The 1-NN algorithm predicts the label of any new datapoint $\mathbf{x}$ by finding its closest neighbor $\mathbf{x}'$ in the training set, and then predicts the label of $\mathbf{x}$ to be the same as the label of $\mathbf{x}'$. For general k, the k-NN algorithm predicts the label by taking a majority vote on the k nearest neighbors.

We now describe the algorithm more rigorously. Given a hyper-parameter k, training instances $(\mathbf{x}_i, y_i)$ ($\mathbf{x}_i \in \mathbb{R}^d$ and y_i is the label), and a test example $\mathbf{x}$, the k-NN algorithm can be executed based on the following steps,

1. Calculate the *distances* between the test example and each of the training examples.
2. Take the k nearest neighbors based on the distances calculated in the previous step.
3. Among these k nearest neighbors, count the number of the data points in each class.
4. Predict the label $\hat{y}$ of $\mathbf{x}$ to be the most frequent class among these neighbors (we describe how to break any ties later).

You are asked to implement the missing functions in `knn.py` following each of the steps.

Part 4.1 Report 4-nearest neighbor accuracy (8pts)

Euclidean distance calculation Compute the distance between the data points based on the following equation:

$$d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2 = \sqrt{\sum_{i=1}^{d} (x_i - x_i')^2}. \tag{19}$$

Task: Fill in the code for the function `compute_l2_distances`.

k-NN classifier Implement a k-NN classifier based on the above steps. Your algorithm should output the predictions for the test set. *Note:* You do not need to worry about ties in the distance when finding the k nearest neighbor set. However, when there are ties in the majority label of the k nearest neighbor set, you should return the label with the smallest index. For example, when $k = 4$, if the labels of the 4 nearest neighbors happen to be 0, 0, 1, 1, your prediction should be the label 0.

Task: Fill in the code for the function `predict_labels`.

Report the error rate We want you to report the error rate for the classification task. The error rate is defined as:

$$\text{error rate} = \frac{\text{\# of wrongly classified examples}}{\text{\# of total examples}}. \tag{20}$$

Task: Fill in the code for the function `compute_error_rate`.

Report Item: Report the error rate of your k nearest neighbor algorithm in the validation set when $k = 4$ using Euclidean distance.

Ans. The validation error rate is 34.67% (0.3466666666666667)

Rubric: 8pt for this question.

Part 4.2 Data transformation (2+2pts)

We are going to add one more step (data transformation) in the `data_processing` part and see how it works. Data transformations and other feature engineering steps often play a crucial role to make a machine learning model work. Here, we take two different data transformation approaches. This link might be helpful: `https://en.wikipedia.org/wiki/Feature_scaling`.

Normalizing the feature vector This one is simple but sometimes may work well. Given a feature vector $\mathbf{x}$, the normalized feature vector is given by $\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. If a vector is an all-zero vector, define the normalized vector to also be the all-zero vector (in practice, a useful trick is to add a very very small value to the norm of $\mathbf{x}$, so that we do not need to worry about the division by zero error).

Min-max scaling for each feature The above normalization is independent of the rest of the data. On the other hand, min-max normalization scales each sample in a way that depends on the rest of the data. More specifically, for each feature in the training set, we normalize it linearly so that its value is between 0 and 1 across all samples, and in addition, the largest value becomes exactly 1 while the smallest becomes exactly 0. Then, we will apply the same scaling parameters we get from training data to our testing instances.

Task: Fill in the code for the function `data_processing_with_transformation`.

Report Item: Report the error rate of your k nearest neighbor algorithm in the validation set for $k = 4$ using Euclidean distance when data is using (1) Normalized featured vector, and (2) Min-max scaling featured vector.

Ans. The validation error rate is 33.33% (0.3333333333333333) when using normalization.

The validation error rate is 30.67% (0.30666666666666664) when using min-max scaling.

Rubric: 2pt for each error rate.

Part 4.3 Different distance measurement (3pts)

In this part, we will change the way that we measure distances. We will work with the original (unnormalized) data for simplicity, and continue the experiments from Part 4.1.

Cosine distance calculation Compute the distance between data points based on the following equation:

$$d(\mathbf{x}, \mathbf{x}') = \begin{cases} 1, & \text{if } \|\mathbf{x}\|_2 = 0 \text{ or } \|\mathbf{x}'\|_2 = 0 \\ 1 - \dfrac{\mathbf{x} \cdot \mathbf{x}'}{\|\mathbf{x}\|_2 \|\mathbf{x}'\|_2} & \text{otherwise.} \end{cases}$$

Similar to when we are normalizing the feature vector, a useful trick in practice is to add a very very small value to the norm of $\mathbf{x}$, so that we do not need to worry about the division by zero issues.

Task: Fill in the code for the function `compute_cosine_distances` and change distance function used in the main function in the code to get results.

Report Item: Report the error rate of your k nearest neighbor algorithm in the validation set for $k = 4$ using cosine distance for *original data*.

Ans. The validation error rate is 33.33% (0.3333333333333333)

Rubric: 3pt for getting the error rate right.

Part 4.4 Tuning the hyper-parameter k (10pts)

Again, follow Part 4.1, however, this time we conduct experiments with $k = \{1, 2, 4, 6, 8, 10, 12, 14, 16, 18\}$.

Task: Fill in the code for the function `find_best_k`.

Report Item: (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) (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) (3) What do you observe by comparing the difference between the two curves? (2pts) (4) What's the final test set error rate you get using your best-k? (1pt) (5) Comment on these results (3pts).

Ans.

(1)

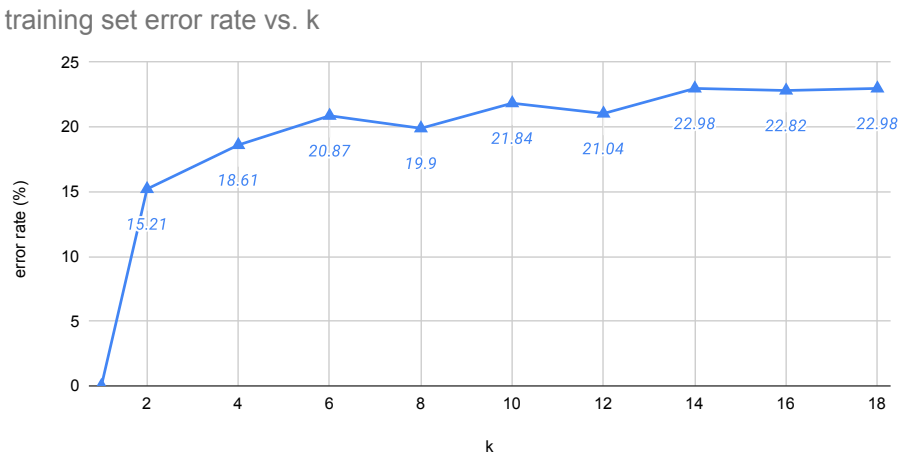

Figure 3: My kNN model's error rate on the training set using different k.

Observation: (a) Training error is zero when k equals zero since the algorithm will always find its own label. Yet, the training error will not be zero when k increases.

(2)

We observe that $k = 12$ gets the best result of a 24% error rate on the validation set.

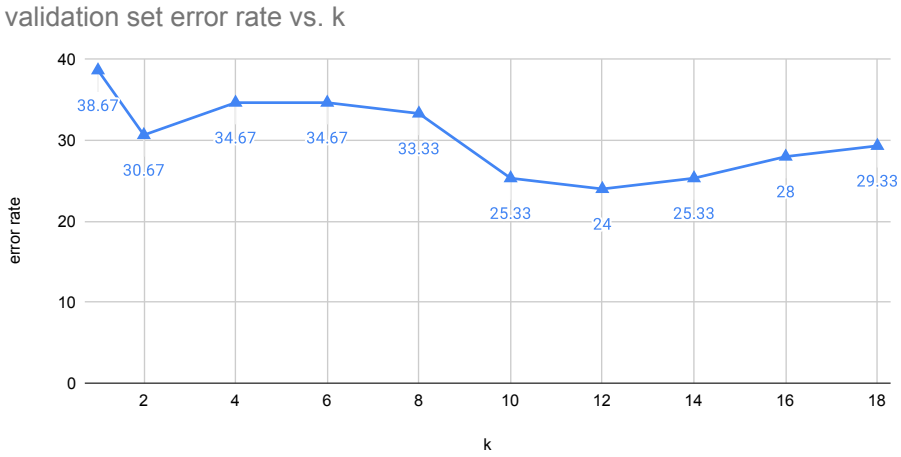

Figure 4: My kNN model's error rate on the validation set using different k.

(3) Two major difference between the training error curve versus validation error curve are: (a) In general, the training error rate is growing when k increases. This happens because you can identify an exact match if you have $k = 1$. Yet, when k grows, the influence of the "exact match" instance is getting lower and lower (its weight is amortized). (b) Different from the training error rate curve, validation error rate curve is more like a convex function with a global minimum.

(4) Using the best k, the final test error rate is 21.33% (0.21333333333333335).

(5) In this question, we learn what is "overfitting". Specifically, if we set the k that leads to the best *training* performance, we will select $k = 1$. However, as we see in the two figures. $k = 1$ actually gives the worst results on the validation set. This highlights the importance of having a separate validation set to set up the hyperparameter in your model (in our case, the hyperparameter is k).

Rubrics for grading:

- two correct figures for training/validation set error rate v.s. k (1.5pt for each figure)
- report best k (1.5pt)
- mentioned training error is 0 for $k = 0$ or training error is not 0 for $k \neq 0$ (0.5pt)
- the shape of each curve (0.5 pt for each shape)
- report test set result (1pt)
- discussing overfitting/generalization related concept (2pt) and the importance of hyperparameter tuning on the validation set (1pt)

We do not need to care too much about where these concepts are reported when grading.

Part 4.5 (0pts)

Report Item: Include your filled-in code in the submitted pdf file.

Problem 5: Linear Regression (20pts)

We will consider the problem of fitting a linear model. Given d-dimensional input data $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$ with real-valued labels $y_1, \cdots, y_n \in \mathbb{R}$, the goal is to find the coefficient vector $\mathbf{w}$ that minimizes the sum of the squared errors. The total squared error of $\mathbf{w}$ can be written as $F(\mathbf{w}) = \sum_{i=1}^{n} f_i(\mathbf{w})$, where $f_i(\mathbf{w}) = (\mathbf{w}^\mathbf{T}\mathbf{x_i} - y_i)^2$ denotes the

squared error of the ith data point. We will refer to $F(\mathbf{w})$ as the *objective function* for the problem.

The data in this problem will be drawn from the following linear model. For the training data, we select n data points $\mathbf{x}_1, \cdots, \mathbf{x}_n$, each drawn independently from a d-dimensional Gaussian distribution. We then pick the "true" coefficient vector $\mathbf{w}^*$ (again from a d-dimensional Gaussian), and give each training point $\mathbf{x}_i$ a label equal to $(\mathbf{w}^*)^\top \mathbf{x}_i$ plus some noise (which is drawn from a 1-dimensional Gaussian distribution).

The following Python code will generate the data used in this problem.

```
d = 100 # dimensions of data
n = 1000 # number of data points
X = np.random.normal(0,1, size=(n,d))
w_true = np.random.normal(0,1, size=(d,1))
y = X.dot(w_true) + np.random.normal(0,0.5,size=(n,1))
```

5.1 (6pts) Least-squares regression has the closed form solution $\mathbf{w}_{LS} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$, which minimizes the squared error on the data. (Here $\mathbf{X}$ is the $n \times d$ data matrix as in the code above, with one row per data point, and $\mathbf{y}$ is the n-vector of their labels.) Solve for $\mathbf{w}_{LS}$ on the training data and report the value of the objective function $F(\mathbf{w}_{LS})$. For comparison, what is the total squared error $F(\mathbf{w})$ if you just set $\mathbf{w}$ to be the all 0's vector? Using similar Python code, draw $n = 1000$ test data points from the same distribution, and report the total squared error of $\mathbf{w}_{LS}$ on these test points. What is the gap in the training and test objective function values? Comment on the result.

Note: Computing the closed-form solution requires time $O(nd^2 + d^3)$, which is slow for large d. Although gradient descent methods will not yield an exact solution, they do give a close approximation in much less time. For the purpose of this assignment, you can use the closed-form solution as a good sanity check in the following parts.

Denote $\mathbf{w_0}$ is all zero weight vector.

1. $F(\mathbf{w}_{\text{LS}}) = 218.3018676427049$ on training data.
2. $F(\mathbf{w_0}) = 90138.99700337248$ on training data.
3. $F(\mathbf{w}_{\text{LS}}) = 289.0410748147839$ on testing data.

Then non-zero training loss is because of the added noise when generating y. Yet, given the small value (around 0.2 difference per data point), we can indicate that the model fits the training data well. The gap between training error versus testing error is around 71, which is around 32% of the training loss. This gap mainly comes from the difference between training data and testing data. However, since both training and testing data come from similar distribution (w_{true}), the gap is not big.

Rubrics: All answer in the question is just a reference since the data is generated randomly. However, Answer(1) should be close to 200, Answer(2) should be larger than 10000, and Answer(3) should be larger than 240, but not too large.

- Reasonable reported value for (1) (2) (3) (1pt for each)
- Report the generalization gap (1pt)
- Comment on training loss (1pt)
- Comment on where the generalization gap (1pt)

5.2 (7pts) In this part, you will solve the same problem via *gradient descent* on the squared-error objective function $F(\mathbf{w}) = \sum_{i=1}^{n} f_i(\mathbf{w})$. Recall that the gradient of a sum of functions is the sum of their gradients. Given a point $\mathbf{w}_t$, what is the gradient of f at $\mathbf{w}_t$?

Now use gradient descent to find a coefficient vector $\mathbf{w}$ that approximately minimizes the least squares objective function over the training data. Run gradient descent three times, once with each of the step sizes 0.00005, 0.0005, and 0.0007. You should initialize $\mathbf{w}$ to be the all-zero vector for all three runs. Plot the objective function value for 20 iterations for all 3 step sizes on the same graph. Comment in 3-4 sentences on how the step size can affect the convergence of gradient descent (feel free to experiment with other step sizes). Also report the step size that had the

best final objective function value and the corresponding objective function value.

(1) The gradient of f is $2(\mathbf{w^T x_i} - y_i)\mathbf{x_i}$. Hence, the gradient of F is $\sum_{i=1}^{n} 2(\mathbf{w^T x_i} - y_i)\mathbf{x_i}$.

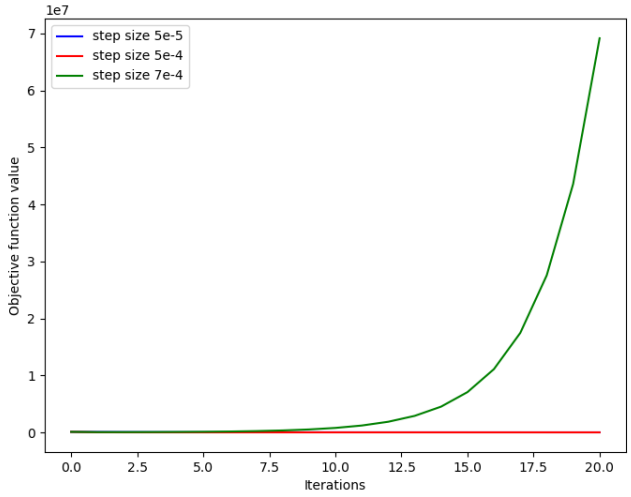

Figure 5: The objective function value for solving Question 5 using gradient descent. Note that the y-axis scale is in $1e10$.

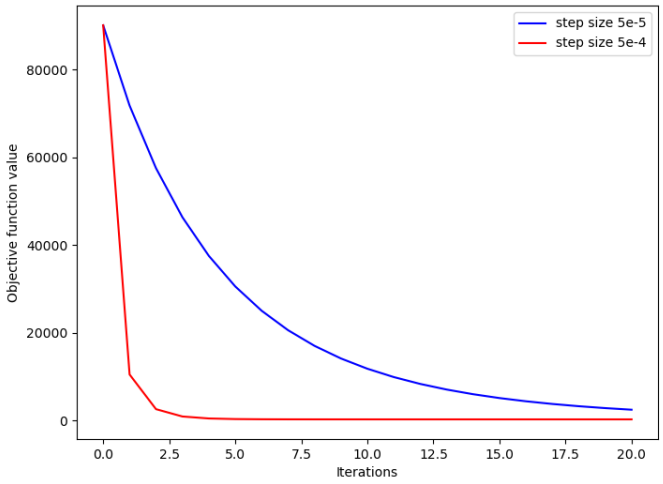

Figure 6: A closer look for Figure 5 for step size $5e-5$ and $5e-4$.

(2) Figure 5 demonstrates our experimental results for using different step sizes for our gradient descent algorithm to get the weight vector, and Figure 6 shows a closer look on the two step sizes that gets better results.

(3) From these figures, we found that: If the step size is small, it would take more iterations to converge to the minimum point. However, if the step size is too large, like the case of step size being $7e-4$, gradient decent method might oscillate and never converge. In our test, the best step size is $5e-4$ and the corresponding objective function value (F) is 218.3018924969777.

Rubrics: (1)'s should be an exact match, (3)'s best value should be very close to what they reported in 5.1. For grading this question,

- Correct gradient function (1pt) (reporting either f's gradient function or F's gradient function is fine)
- Plots (3pt)
- Comment on step size is too large (1pt)
- Comment on step size comparison of $5e-4$ and $5e-5$ (1pt)
- Comment on best value and best step size (1pt)

5.3 (7pts) In this part, you will run *stochastic gradient descent* to solve the same problem. Recall that in stochastic gradient descent, you pick one training datapoint at a time, say $(\mathbf{x}_i, y_i)$, and update your current value of $\mathbf{w}$ according to the gradient of $f_i(\mathbf{w}) = (\mathbf{w}^\mathbf{T}\mathbf{x_i} - y_i)^2$.

Run stochastic gradient descent using step sizes $\{0.0005, 0.005, 0.01\}$ and 1000 iterations. Plot the objective function value vs. the iteration number for all 3 step sizes on the same graph. Comment 3-4 sentences on how the step size can affect the convergence of stochastic gradient descent and how it compares to gradient descent. Compare the performance of the two methods. How do the best final objective function values compare? How many times does each algorithm use each data point? Also report the step size that had the best final objective function value and the corresponding objective function value.

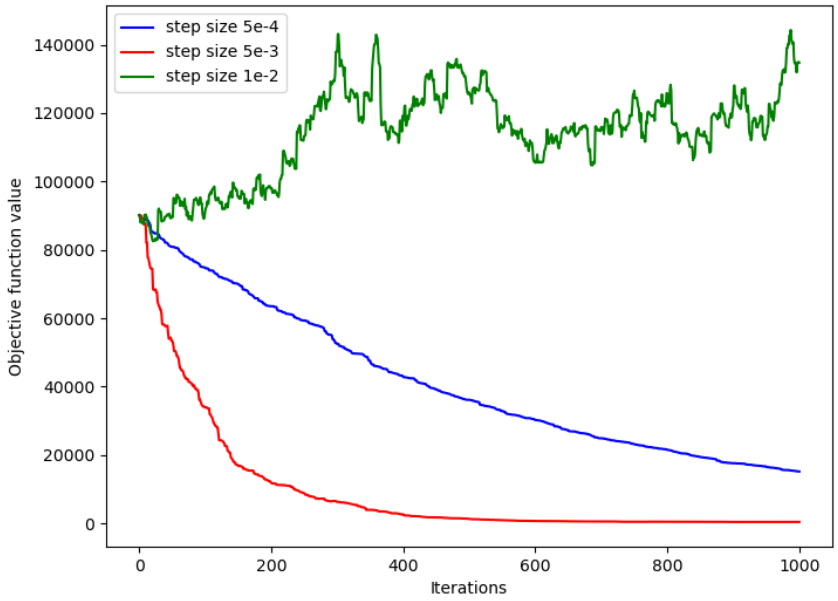

Figure 7: The objective function value for solving Question 5 using stochastic gradient descent.

Figure 7 plots our experimental results for stochastic gradient descents.

(1) Similar to gradient descent, if the step size is small, it would take more iterations to converge, for example, the step size being $5e-4$ cases is not yet converged. If the step size is too large, like in the case of step size being $1e-2$, the stochastic gradient descent method might oscillate and never converge.

(2) The best value we get is using a step size of $5e{-}3$ and we get an objective function value of 410.2370849948935. Comparing this value to what we got from 5.2, we observe that gradient descent gets better results. However, in 5.2, we use 20 times for each data point (since 20 iterations) to perform updating. Here, we only use 1 time of each data point (on average) to perform updating.

Rubrics: (2)'s the best value should be close to 500. For grading this question, use the rubrics:

- Plots (2pt)
- Comment on step size effect (2pt)
- Comment on comparison between gradient descent and stochastic gradient descent (2pt)
- Comment on best value and best step size (1pt)

5.4 (0pts) Include the code for all the previous parts in the submitted pdf file.