

CSE 5819 Assignment #3

by: Aayushi Verma (uef24001)

This assignment is organized as follows:

1. ChatGPT - pg 1–4
2. Problems - pg 5–7
3. Coding - attached as a separate .ipynb file

1 Part 1: ChatGPT Self-Learning (20pts)

You can start from the following prompts, but you need to create subsequent questions that attempt to understand the basic concepts of machine learning.

1.1 What is the definition of convex optimization?

Convex optimization is a subfield of optimization that deals with problems where the objective function is convex, and the feasible set is a convex set. Formally, a convex optimization problem can be expressed as:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in \mathcal{C}, \end{aligned}$$

where:

- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex function.
- $\mathcal{C} \subseteq \mathbb{R}^n$ is a convex set.

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if, for any two points $x_1, x_2 \in \mathbb{R}^n$ and any $\theta \in [0, 1]$, the following inequality holds:

$$f(\theta x_1 + (1 - \theta)x_2) \leq \theta f(x_1) + (1 - \theta)f(x_2).$$

A set \mathcal{C} is convex if, for any two points $x_1, x_2 \in \mathcal{C}$ and any $\theta \in [0, 1]$, the point $\theta x_1 + (1 - \theta)x_2$ is also in \mathcal{C} .

1.2 How to solve an unconstrained convex optimization problem?

To solve an unconstrained convex optimization problem, you need to minimize a convex objective function $f(x)$ over the variable $x \in \mathbb{R}^n$. The problem can be formulated as:

$$\text{minimize } f(x).$$

Here are common methods to solve this problem:

- **Gradient Descent:** This iterative method updates the solution x by moving in the direction of the negative gradient of the objective function. The update rule is:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

where α_k is the step size (or learning rate) and $\nabla f(x_k)$ is the gradient of f at x_k .

- **Newton's Method:** This method uses second-order information by incorporating the Hessian matrix $\nabla^2 f(x)$. The update rule is:

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

- **Conjugate Gradient Method:** This is suitable for large-scale problems where the Hessian is not explicitly available. It iteratively solves a sequence of linear systems and is effective for quadratic objective functions.
- **Quasi-Newton Methods:** These methods, such as BFGS, approximate the inverse Hessian matrix and update it iteratively, combining benefits of both gradient descent and Newton's method.
- **Subgradient Methods:** For nonsmooth convex functions, subgradient methods can be used, where the update rule is:

$$x_{k+1} = x_k - \alpha_k g_k,$$

where g_k is a subgradient of f at x_k .

1.3 How to solve a constrained convex optimization problem?

To solve a constrained convex optimization problem, you need to minimize a convex objective function $f(x)$ subject to convex constraints. The problem can be formulated as:

$$\begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } x \in \mathcal{C}, \end{aligned}$$

where \mathcal{C} is a convex set. Common methods to solve this type of problem include:

- **Projected Gradient Descent:** This method involves iteratively applying gradient descent followed by projection onto the feasible set \mathcal{C} . The update rule is:

$$x_{k+1} = P_{\mathcal{C}}(x_k - \alpha_k \nabla f(x_k)),$$

where $P_{\mathcal{C}}$ denotes the projection operator onto the set \mathcal{C} .

- **Lagrange Duality:** Introduce Lagrange multipliers for the constraints and solve the Lagrangian dual problem. The Lagrangian is:

$$\mathcal{L}(x, \lambda) = f(x) + \lambda^T (h(x) - b),$$

where $h(x)$ represents the constraint functions and λ are the Lagrange multipliers. Solve the dual problem to find the optimal multipliers and use them to obtain the solution to the primal problem.

- **Interior-Point Methods:** These methods solve constrained problems by transforming them into a series of unconstrained problems. They iteratively update the solution while maintaining feasibility with respect to the constraints.
- **Penalty Methods:** Convert the constrained problem into an unconstrained problem by adding a penalty term for constraint violations. The modified objective function is:

$$f(x) + \rho \sum_i \max(0, g_i(x))^2,$$

where $g_i(x) \leq 0$ are the inequality constraints and ρ is a penalty parameter.

- **Augmented Lagrangian Methods:** Combine ideas from Lagrange multipliers and penalty methods. The augmented Lagrangian function is:

$$\mathcal{L}_{\rho}(x, \lambda) = f(x) + \lambda^T (h(x) - b) + \frac{\rho}{2} \|h(x) - b\|^2,$$

where ρ is a penalty parameter and λ are the multipliers.

1.4 Would gradient descent converge if it is used to solve a non-convex optimization problem?

Gradient descent can be used to solve non-convex optimization problems, but its convergence properties are different from those in the convex case. Here are key points to consider:

- **Local Minima:** In non-convex optimization, gradient descent may converge to a local minimum rather than the global minimum. The final solution depends on the initial starting point and the nature of the objective function.
- **Saddle Points:** Gradient descent can also get stuck at saddle points, where the gradient is zero but the point is neither a local minimum nor a maximum.
- **Convergence Guarantees:** There are no general guarantees of convergence to a global minimum in non-convex problems. Convergence is usually to a local minimum or a stationary point, depending on the problem and algorithm specifics.
- **Choice of Step Size:** The choice of step size (learning rate) is crucial. A step size that is too large can lead to divergence or oscillations, while a step size that is too small can result in slow convergence.

- **Stochastic Gradient Descent (SGD):** In practice, techniques like stochastic gradient descent (SGD) and its variants (e.g., Adam, RMSprop) are often used. These methods can help escape local minima and saddle points by adding noise to the gradient updates.
- **Initialization and Restarts:** Multiple restarts with different initial points can increase the likelihood of finding a better local minimum. Proper initialization can also help avoid poor local minima.
- **Convexity Assumptions:** For guaranteed convergence to a global minimum, the objective function should be convex. In the absence of convexity, additional techniques and heuristics are required to handle the complexities of non-convex landscapes.

1.5 What loss functions will form convex functions?

A loss function $L(y, \hat{y})$ is convex in its parameters θ if for all possible parameter values θ_1 and θ_2 and for any $\lambda \in [0, 1]$, the following inequality holds:

$$L(\lambda\theta_1 + (1 - \lambda)\theta_2) \leq \lambda L(\theta_1) + (1 - \lambda)L(\theta_2).$$

Here are some commonly used convex loss functions:

- **Squared Error Loss (Mean Squared Error, MSE):**

$$L(y, \hat{y}) = (y - \hat{y})^2,$$

where y is the true value and \hat{y} is the predicted value. This loss function is convex with respect to \hat{y} .

- **Absolute Error Loss (Mean Absolute Error, MAE):**

$$L(y, \hat{y}) = |y - \hat{y}|,$$

where y is the true value and \hat{y} is the predicted value. This loss function is convex with respect to \hat{y} , but not differentiable at $\hat{y} = y$.

- **Logarithmic Loss (Log Loss or Binary Cross-Entropy):**

$$L(y, \hat{y}) = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})],$$

where $y \in \{0, 1\}$ is the binary label and \hat{y} is the predicted probability. This loss function is convex with respect to \hat{y} in the range $(0, 1)$.

- **Hinge Loss:**

$$L(y, \hat{y}) = \max(0, 1 - y\hat{y}),$$

where $y \in \{-1, 1\}$ is the true label and \hat{y} is the predicted value. This loss function is convex with respect to \hat{y} .

- **Kullback-Leibler Divergence (KL Divergence):**

$$L(P, Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)},$$

where P and Q are probability distributions. This loss function is convex in Q for a fixed P .

- **Huber Loss:**

$$L(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta, \\ \delta|y - \hat{y}| - \frac{1}{2}\delta^2 & \text{if } |y - \hat{y}| > \delta, \end{cases}$$

where δ is a threshold parameter. The Huber loss is convex with respect to \hat{y} and is used to be less sensitive to outliers compared to squared error loss.

2 Problems

[Convex Optimization, Mean/Medium/Mode, Preparation for Learning Loss Functions]

Take the training data set $D = \{(x_i, y_i): 1 \leq i \leq n\}$. Now consider the following hypothesis class (i.e., the set of constant functions)

$$H = \{h_z(x) = z, \forall x \in R\}$$

We will choose a model from this hypothesis class in our machine learning task.

2.1 Question One

1. [20 pts] Let us use L_2 vector norm to measure the discrepancy between the observed y and the model output $h_z(x)$. In other words, the squared loss is defined as

$$L_{sq}(h) = \frac{1}{n} \sum_{i=1}^n (h(x_i) - y_i)^2$$

Prove that if we set the constant z equal to the mean of y

$$m = \frac{1}{n} \sum_{i=1}^n y_i$$

this constant function $h_m(\cdot)$ minimizes the squared loss, and in other words,

$$h_m = \arg \min_{h \in H} L_{sq}(h)$$

We are given our model from the hypothesis class as:

$$H = \{h_z(x) = z, \forall x \in \mathbb{R}\} \quad (1)$$

Our squared loss function is given as:

$$\begin{aligned} L_{sq}(h) &= \frac{1}{n} \sum_{i=1}^n (h(x_i) - y_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (z - y_i)^2 \end{aligned} \quad (\text{from Eq. 1}) \quad (2)$$

This is a convex optimization problem. From the lectures, we know that the optimality condition is given as:

$$\min_{x \in \mathbb{R}^n} \Rightarrow \nabla_x f(x^*) = 0 \quad (3)$$

Therefore, we first take the derivative of our loss function, Eq. 2.

$$\begin{aligned} \frac{\partial L_{sq}(h)}{\partial z} &= \frac{\partial}{\partial z} \left(\frac{1}{n} \sum_{i=1}^n (z - y_i)^2 \right) \\ &= \frac{1}{n} \frac{\partial}{\partial z} \sum_{i=1}^n (z - y_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^n 2 \cdot (z - y_i) \\ &= \frac{2}{n} \sum_{i=1}^n (z - y_i) \end{aligned} \quad (4)$$

We then set the derivative (Eq. 4) to 0 to optimize and solve.

$$\begin{aligned}
0 &= \frac{2}{n} \sum_{i=1}^n (z - y_i) \\
0 &= \sum_{i=1}^n (z - y_i) \\
0 &= \sum_{i=1}^n z - \sum_{i=1}^n y_i \\
0 &= n \cdot z - \sum_{i=1}^n y_i \quad (z \text{ is a constant}) \\
n \cdot z &= \sum_{i=1}^n y_i \\
\Rightarrow z &= \frac{1}{n} \sum_{i=1}^n y_i \tag{5}
\end{aligned}$$

∴ We have shown in Eq. 5 that $z = m$, the mean of y , when minimizing the squared loss function, $L_{sq}(h)$ (Eq. 2).

∴ $h_m(\cdot) = h_z(x) = z = m = \frac{1}{n} \sum_{i=1}^n y_i$ minimizes the squared loss function, and hence $h_m = \arg \min_{h \in H} L_{sq}(h)$.

2.2 Question Two

2. [20 pts] Let us use the L_1 vector norm to measure the discrepancy between the observed y and the model output $h_z(x)$. In other words, the absolute loss is defined as

$$L_{abs}(h) = \frac{1}{n} \sum_{i=1}^n |h(x_i) - y_i|$$

Prove that if we set the constant z equal to the median of y

$$m = \text{median}(\{y_i, \quad 1 \leq i \leq n\})$$

this constant function $h_m(\cdot)$ minimizes the absolute loss, and in other words,

$$h_m = \arg \min_{h \in H} L_{abs}(h)$$

Our absolute loss function is given as:

$$\begin{aligned}
L_{abs}(h) &= \frac{1}{n} \sum_{i=1}^n |h(x_i) - y_i| \\
&= \frac{1}{n} \sum_{i=1}^n |z - y_i| \quad (\text{from Eq. 1}) \tag{6}
\end{aligned}$$

Following a similar procedure as the previous question, we treat this as a convex optimization problem, and take the derivative of Eq. 6 in order to find the value of z that minimizes it.

$$\begin{aligned}
\frac{\partial L_{abs}(h)}{\partial z} &= \frac{\partial}{\partial z} \left(\frac{1}{n} \sum_{i=1}^n |z - y_i| \right) \\
&= \frac{1}{n} \frac{\partial}{\partial z} \sum_{i=1}^n |z - y_i| \tag{7}
\end{aligned}$$

Let's think carefully about what's happening here. We want to find the value of z that minimizes $L_{abs}(h)$ (Eq.

6). Examining the derivative in Eq. 7 further:

$$\frac{\partial}{\partial z}|z - y_i| = \begin{cases} 1 & \text{if } z > y_i \\ 0 & \text{if } z = y_i \\ -1 & \text{if } z < y_i \end{cases}$$

This indicates that when $z = y_i$, the gradient is equal to 0, which means $L_{abs}(h)$ is minimized. Given that $L_{abs}(h)$ is summing the absolute differences between our z and every y_i , $z = y_i$ only when z is the median of y . We know that the median is the middle point of all the values in y arranged in ascending order (for example if $y = 1, 2, 4, 5, 7$, then the median of y will be 4). Therefore if our z is set to the median value of y , then the derivative of each $|z - y_i|$ will be equal to 0, which minimizes $L_{abs}(h)$ according to the optimality condition (Eq. 3).

\therefore We have shown if that $z = m$, the median of y , then the absolute loss function, $L_{abs}(h)$ (Eq. 6) is minimized.
 $\therefore h_m(\cdot) = h_z(x) = z = m = \text{median}(y)$ minimizes the absolute loss function, and hence $h_m = \arg \min_{h \in H} L_{abs}(h)$.

2.3 Question Three

3. [20 pts] Let us use the L_0 vector norm to measure the discrepancy between the observed y and the model output $h_z(x)$. In other words, the binary loss is defined as

$$L_{bin}(h) = \frac{1}{n} \sum_{i=1}^n 1\{h(x_i) \neq y_i\}$$

where $1\{h(x_i) \neq y_i\}$ returns 1 if $h(x_i) \neq y_i$; or otherwise return 0. In other words, this loss counts how many times the model output $h(x_i)$ is not equal to the observed y_i . Then normalize the sum by n .

Prove that if we set the constant z equal to the mode of y

$$m = \text{mode}(\{y_i, \quad 1 \leq i \leq n\})$$

where the mode is the most common number in the observed y 's,

this constant function $h_m(\cdot)$ minimizes the binary loss, and in other words,

$$h_m = \arg \min_{h \in H} L_{bin}(h)$$

Our binary loss function is given by:

$$\begin{aligned} L_{bin}(h) &= \frac{1}{n} \sum_{i=1}^n 1(h(x_i) \neq y_i) \\ &= \frac{1}{n} \sum_{i=1}^n 1(z \neq y_i) \end{aligned} \quad \text{(from Eq. 1)} \quad (8)$$

Once again, we want to solve for the optimality condition (Eq. 3) and therefore could take the derivative of the binary loss function. However, let's think more deeply about this. $L_{bin}(h)$ measures the normalized total number of times that $z \neq y_i$. Effectively, this is what we can re-write the binary loss function as:

$$L_{bin}(h) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1 & \text{if } z \neq y_i \\ 0 & \text{if } z = y_i \end{cases}$$

The value of $L_{bin}(h)$ will be minimized when we have more occurrences of $z = y_i$ because $\sum_{i=1}^n \Rightarrow 0 + 0 + \dots$ is lower than if we have more occurrences of $z \neq y_i$, in which case $\sum_{i=1}^n \Rightarrow 1 + 1 + \dots$, which produces a higher value for $L_{bin}(h)$. Logically then, it makes sense that $L_{bin}(h)$ is minimized when the value of z is equal to m (the mode of y , i.e. the most-frequently occurring value in y).

Therefore if our z is set to the mode value of y , then there will be more occurrences of the value 0 in the summation part of the binary loss, which inevitably leads to the lowest (therefore most optimal) value of $L_{bin}(h)$.

\therefore We have shown if that $z = m$, the mode of y , then the binary loss function, $L_{bin}(h)$ is minimized.

$\therefore h_m(\cdot) = h_z(x) = z = m = \text{mode}(y)$ minimizes the binary loss function, and hence $h_m = \arg \min_{h \in H} L_{bin}(h)$.