AIML - CS 337

Lecture 9: Convexity and Stability

DATE:- 04/09/2022

Lecturer: Abir De *Scribe: Advaid,Akshat,Utkarsh,Utkarsh Pratap Singh*

1 Convexity of Loss Function

We are given a dataset

$$D = \{(x_i, y_i)\}_{i \in [N]}$$

and we are trying to find a model such that

$$\text{X} \xrightarrow{\text{W}} \text{Y}$$

$$w^T x \approx y$$

If we change x by some value, how much W changes.
If loss function is convex, perturbation in W is also small.

$$l(w^T, x, y) \rightarrow convex$$

Examples of convex functions

- $\sum (y - w^T x)^2$
- $max(0, 1 - yw^T x)$
- $-log(1 + e^{-w^T x})$

Example of Non-convex

- $(y - (w^T x)^2 + (w^T x)^3)^2$ is not convex $w.r.t.\ w$

Deep learning systems have nice tools that finds local minimum of to search for local minima. Convexity is the requirement of single minima i.e. local minima is the global minima.
If loss function is $l(w^T, x, y)$

$$\frac{\mathrm{d}}{\mathrm{d}w}(l(w^T, x, y)) \in \mathbb{R}^{d \times 1}$$

is a vector

$$\frac{\mathrm{d}^2}{\mathrm{d}w^2}(l(w^T, x, y)) \in \mathbb{R}^{d \times d}$$

is a **Hessian** matrix. We require that all its $eigen - values \geq 0$ for $l(w^T, x, y)$ to be convex .
For example
example 1
$l(w) = \sum (y - w^T x)^2$

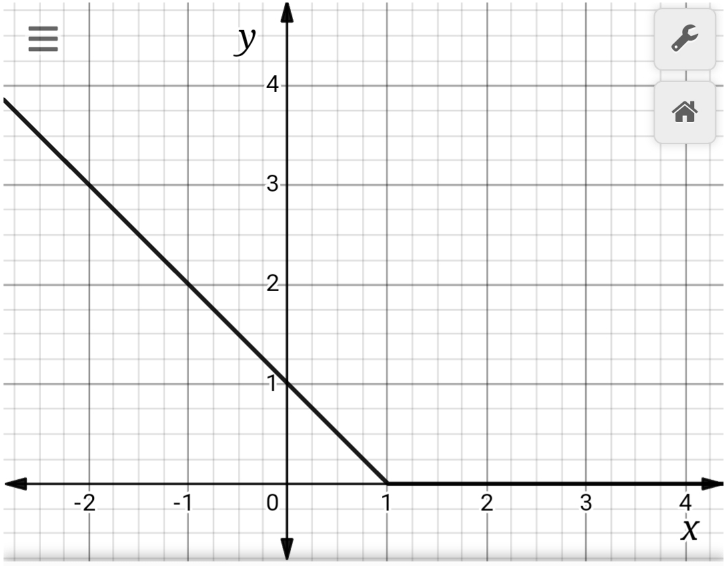
$$\frac{\mathrm{d}}{\mathrm{d}w}(l) = \sum 2(y - w^T x)w$$

$$\frac{\mathrm{d}^2}{\mathrm{d}w^2}(l) = 2XX^T$$

whose eigenvalues are positive (non negative)

Note $\frac{\mathrm{d}^2}{\mathrm{d}w^2}(w^T Aw) = A$
XX^T is a rank 1 matrix which implies it has one non zero eigenvalue
Now, $trace(XX^T) =$ sum of eigenvalues$\rightarrow$ eigen values are $\{||x||^2, 0, 0...\}$
Hence, $l(w) = \sum (y - w^T x)^2$ is a convex function.
example 2
$l(w) = max(0, 1 - yw^T x)$
It is the maximum of a convex function so it is a convex function.
Or, $\frac{\partial^2}{\partial w^2}$ is zero at almost all points, almost everywhere the $eigen - values$ are zero.

The plot of $max(0, 1 - a)$ where $a = yw^T x$



2 Regularization and Stability

Adding regularizer terms leads to a **strictly convex** loss function.

$$l_\lambda(w,x,y) = \sum (y - w^T x)^2 + \lambda||w||^2$$

$$l_\lambda(w,x,y) = l(w,x,y) + \lambda||w||^2$$

2.1 Assumptions regarding loss function

1. $l_\lambda(w,x,y) = l(w,x,y) + \lambda||w||^2$ is strictly convex, since $\frac{\mathrm{d}^2}{\mathrm{d}w^2}(l_\lambda(w,x,y)) > 0$
2. $\left\|\frac{\mathrm{d}l(w,x,y)}{\mathrm{d}w}\right\|$ is **bounded** , i.e $\left\|\frac{\mathrm{d}l(w,x,y)}{\mathrm{d}w}\right\| \leq$ B
3. $\|W(DUK) - W(D)\| = \mathcal{O}(\frac{1}{|D|})$

Let's prove the assumptions,

$$\frac{\mathrm{d}^2}{\mathrm{d}w^2}(l_\lambda(w,x,y)) = \frac{\mathrm{d}^2}{\mathrm{d}w^2}(l(w,x,y)) + 2\lambda\mathbb{I}$$

whose eigenvalues are **strictly positive**. Since $\lambda > 0$, $2\lambda\mathbb{I}$ has strictly positive eigenvalues, while $l(w,x,y)$ has non negative eigenvalue as shown earlier. Hence $l_\lambda(w,x,y))$ is strictly convex.

$$F(W,S) = \sum_{i=1}^{|S|} l_\lambda(W,x_i,y_i)$$

$$W(S) = \min_W \sum_{i=1}^{|S|} l_\lambda(W,x_i,y_i)$$

First of all , note that $F(W(S \cup K),S) - F(W(S),S)$ > 0 because W(S) is the value of W that minimizes F(W,S) .
We want to bound $F(W(S \cup K),S) - F(W(S),S)$. A simple bound on this difference using the notion of Lipschitz continuity is

$$F(W(S \cup K),S) - F(W(S),S) \leq B_\lambda |S| \, \|W(S \cup K) - W(S)\|$$

Now we attempt to form a tighter bound on $F(W(S \cup K),S) - F(W(S),S)$
We can split $F(W(S),S)$ as $F(W(S),S \cup K) - F(W(S),K)$ Note that K is a single data point .

$$\begin{aligned} F(W(S \cup K),S) - F(W(S),S) &= (F(W(S \cup K),S \cup K) - F(W(S),S \cup K)) \\ &\quad + (F(W(S),K) - F(W(S \cup K),K)) \end{aligned}$$

$F(W(S \cup K), S \cup K) - F(W(S), S \cup K) \leq 0$ because $W(S \cup K)$ is the optimal value of W which minimizes $F(W, S \cup K)$.
F is Lipschitz continuous . Hence ,

$$\begin{aligned} F(W(S), K) - F(W(S \cup K), K) &= (W(S) - W(S \cup K))^T \frac{\partial F}{\partial W}\bigg|_{W'} \\ &\leq B \left\| W(S \cup K) - W(S) \right\| \end{aligned}$$

Here we used the Assumption 2 regarding loss function that the first derivative is bounded by some constant B . Hence , we proved that

$$F(W(S \cup K), S) - F(W(S), S) \leq B \left\| W(S \cup K) - W(S) \right\| \tag{1}$$

Now we try to find a lower bound for the expression $F(W(S \cup K), S) - F(W(S), S)$ using the convexity of the loss function .

First we apply Taylor Series Expansion to $F(W(S \cup K), S)$

$$\begin{aligned} F(W(S \cup K), S) = F(W(S), S) &+ (W(S \cup K) - W(S))^T \frac{\partial F}{\partial W}\bigg|_{W=W(S)} \\ &+ \frac{1}{2!}(W(S \cup K) - W(S))^T H(W)(W(S \cup K) - W(S)) \\ &= F(W(S), S) + \frac{1}{2}(W(S \cup K) - W(S))^T H(W)(W(S \cup K) - W(S)) \end{aligned} \tag{2}$$

Where $H(W) = \frac{\partial^2 F}{\partial^2 W}$ is the Hessian Matrix corresponding to F at some W^* lying on the line between W(S) and $W(S \cup K)$. Here we used the idea that $\frac{\partial F}{\partial W}|_{W=W(S)} = 0$ because W(S) is the optimum value which minimizes F(W,S)

Note from Assumption 1 we have that all eigen values of the double derivative of $l(w, x, y)$(loss function with regularization) w.r.t. W are positive. In other words , all eigen values of $\frac{\partial^2 l(W,x,y)}{\partial^2 W}$ are positive .

$$\begin{aligned} F(W, S) &= \sum_{i=1}^{|S|} l_\lambda(W, x_i, y_i) \\ &= \sum_{i=1}^{|S|} l(W, x_i, y_i) + \sum_{i=1}^{|S|} \lambda \left\| W \right\|^2 = \sum_{i=1}^{|S|} l(W, x_i, y_i) + \lambda |S| \left\| W \right\|^2 \end{aligned}$$

$$
\begin{aligned}
H(W) &= \frac{\partial^2 F(W,S)}{\partial^2 W} \\
&= \frac{\partial^2 \sum_{i=1}^{|S|} l_\lambda(W,x_i,y_i)}{\partial^2 W} \\
&= \sum_{i=1}^{|S|} \frac{\partial^2 l(W,x_i,y_i)}{\partial^2 W} + 2\lambda|S|I \\
v^T H(W) v &= \sum_{i=1}^{|S|} v^T \frac{\partial^2 l(W,x_i,y_i)}{\partial^2 W} v + 2\lambda|S|v^T v \\
&\geq 2\lambda|S| \left\|v\right\|^2
\end{aligned}
$$

Apply this to (2) , we get

$$F(W(S\cup K),S) - F(W(S),S) \geq ||W(S\cup K) - W(S)||^2 \lambda|S| \tag{3}$$

From 1 and 3 , we have

$$||W(S\cup K) - W(S)||^2\lambda|S| \leq F(W(S\cup K),S) - F(W(S),S) \leq B\left\|W(S\cup K) - W(S)\right\|$$

$$
\begin{aligned}
||W(S\cup K) - W(S)||^2\lambda|S| &\leq B\left\|W(S\cup K) - W(S)\right\| \\
||W(S\cup K) - W(S)|| &\leq \frac{B}{\lambda|S|}
\end{aligned}
$$

Thus we have proved Assumption 3 which states that

$$\left\|W(S\cup K) - W(S)\right\| = \mathcal{O}(\frac{1}{|S|}) \tag{4}$$

3 Non Linear Regression

In nonlinear regression, we modeled a given data-set by using a function which is a nonlinear combination of the model parameters and depends on one or more independent variables.
That is why nonlinear regression shows association using a curve, making it nonlinear in the parameter.
In contrast to linear regression, we cannot use the ordinary least squares method to fit the data and estimation of the parameters are also not easy.

Let's think about the solution

$$w = (\lambda I + \sum_{i\epsilon D} X_i X^T)^{-1} \sum_{i\epsilon D} X_i Y_i \tag{5}$$

In linear case, we have a line graph. So, what do we have to change in our approach to fit the non-linear regression?

What is that $X_i X^T$ indicates?

$$X_i X^T = ||X_i||^2$$

Suppose we have two vector [1 2 3]^T and [3 6 9]^T.
Are they similar?
Although its directions are same but these are not the similar points.

If the given two points are close enough that its covariance is atleast 0.7, then we will consider the data set as comparately similar to linear regression otherwise we have to think of the non-linear regression otherwise we have to think of th non-linear regression.

Instead of term $X_i X^T$, we can put the Kernel $K(X_i X_i)$ to get some sort of good approximation to our non-linear data set.

Example
For infinite dimension vector in infinite dimension space :-
The measure of **similarity** between $\vec{X_i}$ and $\vec{Y_i}$ is approximate to

$$e^{-||X_i - X_j||^2}$$

Instead of $X_i X^T$, take the matrix formed by using the above equation for each entry (i,j).
The above equation $e^{-||X_i - X_j||^2}$ can be represented as the dot product of two vectors.

Dot product isn't the good measure of similarity
Other measures are:-

$$e^{-||X_i - X_j||^2}$$

$$\frac{1}{1 + ||X_i - X_j||^2}$$

We can replace $X^T X$ by K in the loss function to get

$$w = (\lambda I + K)^{-1} \sum_{i \epsilon D} X_i Y_i \tag{6}$$

Here, K is the Kernel matrix

$$e^{-||X_i - X_j||^2} => \phi_i^T \phi_j$$

Doing non-linearity in finite dimension to linearity in infinite dimension space

3.1 Stability

It is easy for linear regression but it is bit difficult for non linear regression.
So, we need to apply **deep learning** for those data sets which don't have linearity even in infinite dimension i.e, $\phi_i^T \phi_j$

Example

$$\{\vec{x_i}, y_i | i \epsilon D\}$$

For the above example, linear regression or kernel both didn't work.
We can use deep learning concepts to find some relation between data sets i.e, Y = F($\vec{X}$).
We can approximate any kind of functions for the given set of data that follows non linearity using functions like ReLU , Sigmoid and Tanh.
For example, try to find a linear model for y = log(x) such that x $>$ 0.