Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Monika Rathore
Roll Number: 170400
Date: October 25, 2020

QUESTION 1

Yes, the above objective function is convex. If we break the above function in two part we can see first is absolute loss function that is a convex function and second part is L1 regularizer that is also a convex function. We know that non - negative weighted sum of two convex function is also a convex function. Since weight here is $\lambda > 0$.So above function is convex.

Lets see function one by one.

1. For absolute loss, f(w) = $|y_n - \mathbf{w}^T x_n|$ let's assume $y_n - \mathbf{w}^T x_n$ = t

$\partial(f_n(\mathbf{w}))= \begin{cases} x_n & t > 0 \\ -x_n & t < 0 \\ cx_n & c \in [-1, 1], t = 0 \end{cases}$

2. For L1 regularizer

$\frac{\partial L1}{\partial w_k} = \begin{cases} 1 & w_k > 0 \\ -1 & w_k < 0 \\ c & c \in [-1, 1], w_k = 0 \end{cases}$

so, $\partial(L1(\mathbf{w}))$ is Dx1 matrix where kth element is $\frac{\partial L1}{\partial w_k}$ let's denote it by matrix $\mathbf{A}$

Expression of sub-gradient is:

$\partial L(\mathbf{w}) = \sum_{n=0}^{N} \partial f_n(\mathbf{w}) + \lambda \mathbf{A}$

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 2

Student Name: Monika Rathore
Roll Number: 170400
Date: October 25, 2020

My solution to problem 2

Our normal Loss function is:

L $= \sum_{n=1}^{N} (y_n - \mathbf{w}^T x_n)^2$

Now we have a drop out rate m$_n = Bernoulli(p)$.
So new Loss function would be:

$L_D = \sum_{n=1}^{N} (y_n - m_n\mathbf{w}^T x_n)^2$

$\frac{\partial L}{\partial w_i} = \sum_{n=1}^{N} 2(-y_n m_n x_n + w_i m_n x_{ni}^2 + \sum_{j=1,j\neq i}^{n} \mathrm{w}_j m_{ni} m_{nj} x_{ni} x_{nj})$

For normal Loss function let's say $\mathbf{w}_1$ = p$\mathbf{w}$ where p is constant. So,

L $= \sum_{n=1}^{N} (y_n - \mathbf{w}_1^T x_n)^2$

$\frac{\partial L}{\partial w_i} = \sum_{n=1}^{N} 2(-y_n p_{ni} x_{ni} + w_i p_{ni}^2 x_n i^2 + \sum_{j=1,j\neq i}^{n} \mathrm{w}_j p_{ni} p_{nj} x_{ni} x_{nj})$

Now, the expected value of loss function of drop out.

E$[\frac{\partial L_D}{\partial w_i}]$ = -y$_n p_{in} x_{in} + w_{in} Var(m_{ni}) x_{ni}^2 + \sum_{j=1,j\neq i}^{n} (w_j p_{ni} p_{nj} x_{ni} x_{nj})$

E$[\frac{\partial L_D}{\partial w_i}] = \frac{\partial L_n}{\partial w_i}$ + w$_i^2 p_i (1 - p_i) x_{in}^2$

Hence minimizing the expected value of drop out regularizer is equivalent to minimizing regularized function

We can write

L$_D = ||y - \mathbf{w}^T \mathbf{X}||^2 + p(1-p)||(\mathbf{X}^T\mathbf{X})^{0.5}\mathbf{w}||^2$

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 3

Student Name: Monika Rathore
Roll Number: 170400
Date: October 25, 2020

My solution to problem 3

Replacing $\mathbf{W}$ by $\mathbf{BS}$,

$$L(\mathbf{W}) = TRACE[(\mathbf{Y} - \mathbf{XBS})^T(\mathbf{Y} - \mathbf{XBS})] \tag{1}$$

$$= TRACE[(\mathbf{Y}^T - \mathbf{S}^T\mathbf{B}^T\mathbf{X}^T)(\mathbf{Y} - \mathbf{XBS})] \tag{2}$$

$$L(\mathbf{W}) = TRACE[\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{XBS} - \mathbf{S}^T\mathbf{B}^T\mathbf{X}^T\mathbf{Y} + \mathbf{S}^T\mathbf{B}^T\mathbf{X}^T\mathbf{XBS}] \tag{3}$$

Using the identities we get,

$\frac{\partial L(\mathbf{w})}{\partial \mathbf{S}}$ = -$(\mathbf{Y}^T\mathbf{XB})^T - \mathbf{B}^T\mathbf{X}^T\mathbf{Y} + (\mathbf{B}^T\mathbf{X}^T\mathbf{XB}) + (\mathbf{B}^T\mathbf{X}^T\mathbf{XB})^T)\mathbf{S}$

$\frac{\partial L(\mathbf{w})}{\partial \mathbf{B}}$ = -$(\mathbf{Y}^T\mathbf{XS})^T - \mathbf{S}^T\mathbf{X}^T\mathbf{Y} + (\mathbf{S}^T\mathbf{X}^T\mathbf{XS}) + (\mathbf{S}^T\mathbf{X}^T\mathbf{XS})^T)\mathbf{B}$

Now Minimizing the respective gradient function we get,

So closed form solution would be,

1. For $\mathbf{S}$:

$$(\mathbf{B}^T\mathbf{X}^T\mathbf{XB})^T)\mathbf{S} = (\mathbf{Y}^T\mathbf{XB})^T \tag{4}$$

$$\mathbf{S} = (\mathbf{B}^T\mathbf{X}^T\mathbf{XB})^T)^{-1}(\mathbf{Y}^T\mathbf{XB})^T \tag{5}$$

2. For $\mathbf{B}$:

$$(\mathbf{S}^T\mathbf{X}^T\mathbf{XS})^T)\mathbf{B} = (\mathbf{Y}^T\mathbf{XS})^T \tag{6}$$

$$\mathbf{B} = (\mathbf{S}^T\mathbf{X}^T\mathbf{XS})^T)^{-1}(\mathbf{Y}^T\mathbf{XS})^T \tag{7}$$

Now, We will use value of B and update S and then we will use value of S to update B, and will do updates untill both values converges

ALT-OPT:

Step 0 : Initialise $\mathbf{S}^0$

Now,

Step 1 : $\mathbf{B}^{t+1} = (\mathbf{S}^{tT}\mathbf{X}^T\mathbf{XS}^t)^T)^{-1}(\mathbf{Y}^T\mathbf{XS}^t)^T$
Step 2 : $\mathbf{S}^{t+1} = (\mathbf{B}^{tT}\mathbf{X}^T\mathbf{XB}^{t+1})^T)^{-1}(\mathbf{Y}^T\mathbf{XB}^{t+1})^T$
Step 3 : t = t+1 , repeat from Step 1 if both value not converges

Both the sub-problems are equally easy/difficult as they have similar expression so any calculation will require same number of steps

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Monika Rathore
Roll Number: 170400
Date: October 25, 2020

QUESTION 4

My solution to problem 4

$$L(\mathbf{w}) = \frac{1}{2}(y - \mathbf{X}\mathbf{w})^T(y - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \tag{8}$$

Learning rate for Newton's method is hessian so,

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T y + \lambda w \tag{9}$$

$$H = \frac{\partial^2 L}{\partial \mathbf{w}^2} = \mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_D \tag{10}$$

General Newton's method update equation: $\mathrm{w}^{t+1} = w^t - H^{-1}\frac{\partial L}{\partial w}$

Now by inserting value,

Now value of w after first iteration is:

$$w^1 = w^0 - (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}^D)^{-1}((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_D)w^0 - \mathbf{X}^T y) \tag{11}$$

$$w^1 = w^0 - w^0 + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_D)\mathbf{X}^T y \tag{12}$$

$$w^1 = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_D)\mathbf{X}^T y \tag{13}$$

It converges after first iteration.

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Monika Rathore
Roll Number: 170400
Date: October 25, 2020

QUESTION 5

My solution to problem 5

As here number of output are six, we can use multinomial distribution for dice roll. So,

$$p(N|\pi) = \frac{\Gamma(\sum_{n=1}^{6} N_i)}{\prod_{n=1}^{6}\Gamma(x_i)} \prod_{n=1}^{6} \pi_i^{N_i} \tag{14}$$

and distribution of probability we can take conjugate of multinomial that is dirchlet

$$p(\pi) = \frac{\Gamma(\sum_{n=1}^{6} \alpha_i)}{\prod_{n=1}^{6}\Gamma(\alpha_i)} \prod_{n=1}^{6} \pi_i^{\alpha_i - 1} \tag{15}$$

Now,

$$p(\pi|N) \propto p(N|\pi)p(\pi) \tag{16}$$

$$LP(\pi) = log(p(N|\pi)) + log(p(\pi)) \tag{17}$$

$$LP(\pi) = k - \sum_{n=1}^{6} log(N_i) + \sum_{n=1}^{6} N_i log(\pi_i) + \lambda(1 - \sum_{n=1}^{6} \pi_i) + \sum_{n=1}^{6} (\alpha_i - 1) log(\pi_i) \tag{18}$$

$$0 = \frac{N_i}{\pi_i} + \frac{\alpha_i - 1}{\pi_i} - \lambda \tag{19}$$

$$\pi_i = \frac{N_i + \alpha_i - 1}{\lambda} \tag{20}$$

We know that $\sum_{i=1}^{6} \pi_i = 1$

$$\sum_{i=1}^{6} N_i + \sum_{i=1}^{6} \alpha_i - 6 = \lambda$$

$$N + \sum_{i=1}^{6} \alpha_i - 6 = \lambda \tag{21}$$

so,

$$\pi_i = \frac{N_i + \alpha_i - 1}{N + \sum_{i=1}^{6} \alpha_i - 6} \tag{22}$$

MAP Solution is better than MLE when number of observation is small. Because for small observation MLE generally overfits whereas MAP does not.

Now Expression for fully posterior would be:

$$p(\pi|N) = \frac{p(\pi)p(N|\pi)}{p(N)}$$

$$p(\pi|N) \propto \frac{\Gamma(\sum_{n=1}^{6} \alpha_i)}{\prod_{n=1}^{6} \Gamma(\alpha_i)} \frac{\Gamma(\sum_{n=1}^{6} N_i)}{\prod_{n=1}^{6} \Gamma(N_i)} \prod_{n=1}^{6} \pi_i^{\alpha_i - 1} \pi_i^{N_i}$$

$$p(\pi|N) \propto \prod_{n=1}^{6} \pi_i^{N_i + \alpha_i - 1}$$

Fully posterior would be multinomial, so we can write it as:

$$p(\pi|N) = \frac{\Gamma(\sum_{n=1}^{6} N_i + \alpha_i)}{\prod_{n=1}^{6} \Gamma(N_i + \alpha_i)} \prod_{n=1}^{6} \pi_i^{N_i + \alpha_i + 1}$$

MAP can be calculated by optimizing the mod value of fully posterior and MLE will be obtain form MAP when parameters of Dirichlet that is α is equal to 1.