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

Student Name: Abhas Kumar
Roll Number: 20111001
Date: October 30, 2020

QUESTION 1

Let, $\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} |y_n - \mathbf{w}^T \mathbf{x}_n|$ and $\mathcal{R}(\mathbf{w}) = \sum_{d=1}^{D} |w_d|$

Then objective function

$$\mathcal{O}(\mathbf{w}) = \arg\min_{\mathbf{w}} \Big\{ \mathcal{L}(\mathbf{w}) + \lambda \mathcal{R}(\mathbf{w}) \Big\} = \arg\min_{\mathbf{w}} \Big\{ \sum_{n=1}^{N} |y_n - \mathbf{w}^T \mathbf{x}_n| + \lambda \sum_{d=1}^{D} |w_d| \Big\}$$

$\mathcal{L}(\mathbf{w}) = |y_1 - \mathbf{w}^T \mathbf{x}_1| + |y_2 - \mathbf{w}^T \mathbf{x}_2| + ... + |y_N - \mathbf{w}^T \mathbf{x}_N|$ where each of $|y_i - \mathbf{w}^T \mathbf{x}_i|$ is a convex function and since sum of convex functions is also convex, $\mathcal{L}(\mathbf{w})$ is convex.

Similarly, $\mathcal{R}(\mathbf{w}) = |w_1| + |w_1| + ... + |w_D|$ where each of $|w_i|$ is a convex function and sum of convex functions is also convex, $\mathcal{L}(\mathbf{w})$ is convex, so is $\lambda \mathcal{R}(\mathbf{w})$ for $\lambda > 0$.

As $\mathcal{L}(\mathbf{w})$ and $\lambda \mathcal{R}(\mathbf{w})$ are convex , so objective function $\mathcal{O}(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \lambda \mathcal{R}(\mathbf{w})$ is convex.

Let, $t = y_n - \mathbf{w}^T \mathbf{x}_n$ then, $\partial \mathcal{L}(\mathbf{w}) = -\mathbf{x}_n \partial |t|$

$$\implies \partial \mathcal{L}(\mathbf{w}) = \begin{cases} -\mathbf{x}_n \times 1 = -\mathbf{x}_n, & t > 0 \\ -\mathbf{x}_n \times -1 = \mathbf{x}_n, & t < 0 \\ -\mathbf{x}_n \times k = -k\mathbf{x}_n, k \in [-1, 1], & t = 0 \end{cases}$$

Also,

$$\partial_{w_d} \mathcal{R}(\mathbf{w}) = \begin{cases} \lambda, & w_d > 0 \\ -\lambda, & w_d < 0 \\ [-\lambda, \lambda], & w_d = 0 \end{cases}$$

$$\implies \partial_{w_d} \mathcal{R}(\mathbf{w}) = \begin{cases} \lambda \mathbf{sign}(w_d), & |w_d| > 0 \\ [-\lambda, \lambda], & w_d = 0 \end{cases}$$

Hence Optimality conditions for the objective function $\mathcal{O}(\mathbf{w})$ can be given as

$$\partial_{w_d} \mathcal{O}(\mathbf{w}) = \begin{cases} \partial_{w_d} \mathcal{L}(\mathbf{w}) + \lambda, & w_d > 0 \\ \partial_{w_d} \mathcal{L}(\mathbf{w}) - \lambda, & w_d < 0 \\ [\partial_{w_d} \mathcal{L}(\mathbf{w}) - \lambda, \partial_{w_d} \mathcal{L}(\mathbf{w}) + \lambda], & w_d = 0 \end{cases}$$

$$\implies \partial_{w_d} \mathcal{O}(\mathbf{w}) = \begin{cases} \partial_{w_d} \mathcal{L}(\mathbf{w}) + \mathbf{sign}(w_d) = 0, & |w_d| > 0 \\ |\partial_{w_d} \mathcal{L}(\mathbf{w})| \leq \lambda, & w_d = 0 \end{cases}$$

$$\implies \partial_{w_d} \mathcal{O}(\mathbf{w}) = \begin{cases} \partial_{w_d} \mathcal{L}(\mathbf{w}) + \mathbf{sign}(w_d) & |w_d| > 0 \\ \partial_{w_d} \mathcal{L}(\mathbf{w}) + \lambda, & w_d = 0, \partial_{w_d} \mathcal{L}(\mathbf{w}) > -\lambda \\ \partial_{w_d} \mathcal{L}(\mathbf{w}) - \lambda, & w_d = 0, \partial_{w_d} \mathcal{L}(\mathbf{w}) > \lambda \\ 0 & w_d = 0, -\lambda \leq \partial_{w_d} \mathcal{L}(\mathbf{w}) \leq \lambda \end{cases}$$

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

Student Name: Abhas Kumar
Roll Number: 20111001
Date: October 30, 2020

QUESTION 2

Let, the given Loss function using masked input be

$$\mathcal{L}(M) = \sum_{n=1}^{N} \left(y_n - \mathbf{w}^T \tilde{\mathbf{x}}_n\right)^2$$

$$\implies \mathcal{L}(M) = \|\mathbf{y} - \tilde{\mathbf{X}}\mathbf{w}\|^2$$

$$\implies \mathcal{L}(M) = \|\mathbf{y} - (\mathbf{R} * \mathbf{X})\mathbf{w}\|^2$$

When the input $\mathbf{X}$ is dropped out such that any input dimension is retained with probability $\mathbf{p}$, then expected value of $\mathcal{L}(M)$ i.e

$$E_{R\sim Bernoulli(n,p)}\big[\mathcal{L}(M)\big]$$

This needs to be minimised w.r.t $\mathbf{w}$, so let new objective function

$$\mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\{ E_{R\sim Bernoulli(n,p)}\big[\|\mathbf{y} - (\mathbf{R} * \mathbf{X})\mathbf{w}\|^2\big] \right\}$$

$$\implies \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\{ E_{R\sim Bernoulli(n,p)}\big[(\mathbf{y} - (\mathbf{R} * \mathbf{X})\mathbf{w})^T(\mathbf{y} - (\mathbf{R} * \mathbf{X})\mathbf{w})\big] \right\}$$

Let $\mathbf{k} = \mathbf{y} - (\mathbf{R} * \mathbf{X})\mathbf{w}$ and $\mu = E_{R\sim Bernoulli(n,p)}[\mathbf{k}] = \mathbf{y} - p\mathbf{X}\mathbf{w}$ then, we get

$$\mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\{ E_{R\sim Bernoulli(n,p)}\big[\mathbf{k}^T\mathbf{k}\big] \right\}$$

$$\implies \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\{ \mu^T\mu + TRACE\big[(\mathbf{k} - \mu)(\mathbf{k} - \mu)^T\big] \right\}$$

$$\implies \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\{ (\mathbf{y} - p\mathbf{X}\mathbf{w})^T(\mathbf{y} - p\mathbf{X}\mathbf{w}) + p(1-p)TRACE\big[(\mathbf{X}\mathbf{w})(\mathbf{X}\mathbf{w})^T\big] \right\}$$

$$\implies \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{y} - p\mathbf{X}\mathbf{w}\|^2 + p(1-p)TRACE\big[\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T\big] \right\}$$

$$\implies \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{y} - p\mathbf{X}\mathbf{w}\|^2 + p(1-p)\left\|\big(\sqrt{\text{diag}\mathbf{X}^T\mathbf{X}}\big)\mathbf{w}\right\|^2 \right\}$$

Comparing $\mathcal{L}(\mathbf{w})$ with ridge regression objective function,

$$\mathcal{L}_{ridge}(\mathbf{w}) = \arg\min_{\mathbf{w}} \left\{ (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w} \right\} = \arg\min_{\mathbf{w}} \left\{ \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2 \right\}$$

** Clearly, the objective $\mathcal{L}(\mathbf{w})$ resembles with the objective function of ridge regression, hence is equivalent to minimizing a regularized loss function where the term $p(1-p)\left\|\big(\sqrt{\text{diag}\mathbf{X}^T\mathbf{X}}\big)\mathbf{w}\right\|^2$ acts as a regularizer and $\|\mathbf{y} - p\mathbf{X}\mathbf{w}\|^2$ is like squared loss.

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

Student Name: Abhas Kumar
Roll Number: 20111001
Date: October 30, 2020

QUESTION 3

Given Loss function,

$$\mathcal{L}(\mathbf{W}) = \sum_{n=1}^{N}\sum_{m=1}^{M}\Big(y_{mn} - \mathbf{w}_m{}^T\mathbf{x}_n\Big)^2$$

Let, $\mathbf{A} = \mathbf{Y} - \mathbf{XW}$. Then trace, $\text{tr}(\mathbf{A}^T\mathbf{A})$

$$tr(\mathbf{A}^T\mathbf{A}) = \sum_{n=1}^{N}\big[\mathbf{A}^T\mathbf{A}\big]_{nn}$$

$$\implies tr(\mathbf{A}^T\mathbf{A}) = \sum_{n=1}^{N}\sum_{m=1}^{M}\big[\mathbf{A}^T\big]_{nm}\big[\mathbf{A}\big]_{mn}$$

$$\implies tr(\mathbf{A}^T\mathbf{A}) = \sum_{n=1}^{N}\sum_{m=1}^{M}\big[\mathbf{A}\big]_{mn}\big[\mathbf{A}\big]_{mn}$$

$$\implies tr(\mathbf{A}^T\mathbf{A}) = \sum_{n=1}^{N}\sum_{m=1}^{M}\Big(y_{mn} - \mathbf{w}_m{}^T\mathbf{x}_n\Big)^2 = tr[(\mathbf{Y}-\mathbf{XW})^T(\mathbf{Y}-\mathbf{XW})] = \mathcal{L}(\mathbf{W})$$

New objective Function after the transformation,

$$\mathcal{L}(B,S) = \underset{\mathbf{B},\mathbf{S}}{\arg\min}\big\{tr[(\mathbf{Y}-\mathbf{XBS})^T(\mathbf{Y}-\mathbf{XBS})]\big\}$$

$$\implies \mathcal{L}(\mathbf{B},\mathbf{S}) = \underset{\mathbf{B},\mathbf{S}}{\arg\min}\big\{tr[(\mathbf{Y}^T-\mathbf{S}^T\mathbf{B}^T\mathbf{X}^T)(\mathbf{Y}-\mathbf{XBS})]\big\}$$

$$\implies \mathcal{L}(\mathbf{B},\mathbf{S}) = \underset{\mathbf{B},\mathbf{S}}{\arg\min}\big\{tr[\mathbf{Y}^T\mathbf{Y}] - tr[\mathbf{Y}^T\mathbf{XBS}] - tr[\mathbf{S}^T\mathbf{B}^T\mathbf{X}^T\mathbf{Y}] + tr[\mathbf{S}^T\mathbf{B}^T\mathbf{X}^T\mathbf{XBS}]\big\}$$

Taking,

$$\frac{\partial\mathcal{L}(\mathbf{B},\mathbf{S})}{\partial\mathbf{S}} = 0$$

$$\implies 0 - \Big(\mathbf{Y}^T\mathbf{XB}\Big)^T - \mathbf{B}^T\mathbf{X}^T\mathbf{Y} + \Big\{\mathbf{B}^T\mathbf{X}^T\mathbf{XB} + \Big(\mathbf{B}^T\mathbf{X}^T\mathbf{XB}\Big)^T\mathbf{S}\Big\} = 0$$

$$\implies -2\mathbf{B}^T\mathbf{X}^T\mathbf{Y} + 2\mathbf{B}^T\mathbf{X}^T\mathbf{XBS} = 0$$

$$\implies \mathbf{B}^T\mathbf{X}^T\mathbf{XBS} = \mathbf{B}^T\mathbf{X}^T\mathbf{Y}$$

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

Taking,

$$\frac{\partial \mathcal{L}(\mathbf{B},\mathbf{S})}{\partial \mathbf{B}} = 0$$

$$\implies 0 - \left(\mathbf{Y}^T\mathbf{X}\right)^T\mathbf{S}^T - \mathbf{X}^T\mathbf{Y}\mathbf{S}^T + \left(\mathbf{X}^T\mathbf{X}\right)^T\mathbf{B}\mathbf{S}\mathbf{S}^T + \mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{S}\mathbf{S}^T = 0$$

$$\implies -\mathbf{X}^T\mathbf{Y}\mathbf{S}^T - \mathbf{X}^T\mathbf{Y}\mathbf{S}^T + \mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{S}\mathbf{S}^T + \mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{S}\mathbf{S}^T = 0$$

$$\implies -2\mathbf{X}^T\mathbf{Y}\mathbf{S}^T + 2\mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{S}\mathbf{S}^T = 0$$

$$\implies \mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{S}\mathbf{S}^T = \mathbf{X}^T\mathbf{Y}\mathbf{S}^T$$

$$\implies \mathbf{B}\mathbf{S}\mathbf{S}^T = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{S}^T$$

$$\therefore \mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{S}^T\left(\mathbf{S}\mathbf{S}^T\right)^{-1}$$

ALT-OPT Algorithm to Learn $\mathbf{B}$ and $\mathbf{S}$

1. Initialise $\mathbf{B} = \mathbf{B}^0$ for iteration t=0
2. Repeat until convergence

$$3.\mathbf{S}^{(t+1)} = \left(\mathbf{B}^{(t)^T}\mathbf{X}^T\mathbf{X}\mathbf{B}^{(t)}\right)^{-1}\mathbf{B}^{(t)^T}\mathbf{X}^T\mathbf{Y}$$

$$4.\mathbf{B}^{t+1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{S}^{(t+1)^T}\left(\mathbf{S}^{(t+1)}\mathbf{S}^{(t+1)^T}\right)^{-1}$$

5. t=t+1

** Solving sub-problems $\mathbf{B}$ and $\mathbf{S}$ are not equally easy because solving $\mathbf{S}$ requires to find inverse of only **1** $K \times K$ matrix i.e $\mathbf{X}^T\mathbf{X}$ whereas solution of $\mathbf{B}$ requires to find inverse of **2** $K \times K$ matrices namely $\mathbf{X}^T\mathbf{X}$ and $\mathbf{S}\mathbf{S}^T$.

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

Student Name: Abhas Kumar
Roll Number: 20111001
Date: October 30, 2020

QUESTION 4

Second order approximation of function f(x) near x=a

$$f(x) \approx f(a) + f'(x)(x-a) + \tfrac{1}{2}f''(x)(x-a)^2.$$

let,

$$\mathcal{R}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 + \frac{1}{2}\lambda\mathbf{w}^T\mathbf{w}$$

Since Newton's method uses second-order information, at each point $\mathbf{w}^t$ we need to minimise second order approximation(SOA) of $\mathcal{R}(\mathbf{w})$, hence

$\text{SOA}(\mathcal{R}(\mathbf{w}))=\mathcal{R}(\mathbf{w}^{(t)}) + \mathbf{g}^{(t)^T}(\mathbf{w}-\mathbf{w}^{(t)}) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^{(t)})^T\mathcal{H}^{(t)}(\mathbf{w}-\mathbf{w}^{(t)})$

$\text{SOA}(\mathcal{R}(\mathbf{w}))=\mathcal{R}(\mathbf{w}^{(t)}) + \mathbf{g}^{(t)^T}\mathbf{w} - \mathbf{g}^{(t)^T}\mathbf{w}^{(t)} + \frac{1}{2}[\mathbf{w}^T\mathcal{H}^{(t)}\mathbf{w} - 2\mathbf{w}^{(t)^T}\mathcal{H}^{(t)}\mathbf{w} + \mathbf{w}^{(t)^T}\mathcal{H}^{(t)}\mathbf{w}^{(t)}]$

$$\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}}[SOA(\mathcal{R}(\mathbf{w}))]$$

where $\mathbf{g}^{(t)} = \nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)})$ and $\mathcal{H}^{(t)} = \nabla^2_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)})$

$$\frac{\partial SOA(\mathcal{R}(\mathbf{w}))}{\partial \mathbf{w}} = 0 + \mathbf{g}^{(t)} - 0 + \frac{1}{2}[(\mathcal{H}^{(t)} + \mathcal{H}^{(t)^T})\mathbf{w} - 2(\mathbf{w}^{(t)^T}\mathcal{H}^{(t)})^T + 0] = 0$$

$$\implies \mathbf{g}^{(t)} + \tfrac{1}{2}[2\mathcal{H}^{(t)}\mathbf{w} - 2\mathcal{H}^{(t)^T}\mathbf{w}^{(t)}] = 0$$

$$\implies \mathbf{g}^{(t)} + \mathcal{H}^{(t)}\mathbf{w} - \mathcal{H}^{(t)}\mathbf{w}^{(t)} = 0$$

$$\implies \mathcal{H}^{(t)}\mathbf{w} = \mathcal{H}^{(t)}\mathbf{w}^{(t)} - \mathbf{g}^{(t)}$$

$$\implies \mathbf{w} = \mathcal{H}^{(t)^{-1}}[\mathcal{H}^{(t)}\mathbf{w}^{(t)} - \mathbf{g}^{(t)}]$$

$$\implies \mathbf{w} = \mathbf{w}^{(t)} - \mathcal{H}^{(t)^{-1}}\mathbf{g}^{(t)}$$

$$\therefore \mathbf{w}^{t+1} = \mathbf{w}^{(t)} - \mathcal{H}^{(t)^{-1}}\mathbf{g}^{(t)}$$

We have, $\mathcal{R}(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w}) + \frac{1}{2}\lambda\mathbf{w}^T\mathbf{w}$

$\mathcal{R}(\mathbf{w}) = \frac{1}{2}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}) + \frac{1}{2}\lambda\mathbf{w}^T\mathbf{w}$

$\mathbf{g} = \nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}) = 0 - \frac{1}{2}[-(\mathbf{y}^T\mathbf{X})^T - \mathbf{X}^T\mathbf{y} + ((\mathbf{X}^T\mathbf{X} + (\mathbf{X}^T\mathbf{X})^T)\mathbf{w}] + \frac{1}{2}\lambda(I_D + (I_D)^T)\mathbf{w}$

$\mathbf{g} = \frac{1}{2}[-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} + 2\lambda I_D\mathbf{w}]$

$\mathbf{g} = (\mathbf{X}^T\mathbf{X} + \lambda I_D)\mathbf{w} - \mathbf{X}^T\mathbf{y}$

Also $\mathcal{H} = \nabla^2_{\mathbf{w}}\mathcal{R}(\mathbf{w}) = \nabla_{\mathbf{w}}[(\mathbf{X}^T\mathbf{X} + \lambda I_D)\mathbf{w} - \mathbf{X}^T\mathbf{y}] = \mathbf{X}^T\mathbf{X} + \lambda I_D$

Numbers of iteration required to converge

We have, $\mathbf{w}^{t+1} = \mathbf{w}^{(t)} - \mathcal{H}^{(t)^{-1}}\mathbf{g}^{(t)}$

For 1st iteration.

$\mathbf{w}^{(1)} = \mathbf{w}^{(0)} - (\mathbf{X}^T\mathbf{X} + \lambda I_D)^{-1}\{(\mathbf{X}^T\mathbf{X} + \lambda I_D)\mathbf{w}^{(0)} - \mathbf{X}^T\mathbf{y}\}$

$\implies \mathbf{w}^{(1)} = \mathbf{w}^{(0)} - \mathbf{w}^{(0)} + (\mathbf{X}^T\mathbf{X} + \lambda I_D)^{-1}\mathbf{X}^T\mathbf{y}$

$\implies \mathbf{w}^{(1)} = (\mathbf{X}^T\mathbf{X} + \lambda I_D)^{-1}\mathbf{X}^T\mathbf{y}$

For 2nd iteration.

$\mathbf{w}^{(2)} = \mathbf{w}^{(1)} - (\mathbf{X}^T\mathbf{X} + \lambda I_D)^{-1}\{(\mathbf{X}^T\mathbf{X} + \lambda I_D)\mathbf{w}^{(1)} - \mathbf{X}^T\mathbf{y}\}$

$\implies \mathbf{w}^{(2)} = \mathbf{w}^{(1)} - (\mathbf{X}^T\mathbf{X} + \lambda I_D)^{-1}\{(\mathbf{X}^T\mathbf{X} + \lambda I_D)(\mathbf{X}^T\mathbf{X} + \lambda I_D)^{-1}\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y}\}$

$\implies \mathbf{w}^{(2)} = \mathbf{w}^{(1)} - (\mathbf{X}^T\mathbf{X} + \lambda I_D)^{-1}(\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y})$

$\implies \mathbf{w}^{(2)} = \mathbf{w}^{(1)} - (\mathbf{X}^T\mathbf{X} + \lambda I_D)^{-1} \times 0$

$\implies \mathbf{w}^{(2)} = \mathbf{w}^{(1)}$

Hence it converges in **2** iterations.

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

Student Name: Abhas Kumar
Roll Number: 20111001
Date: October 30, 2020

QUESTION 5

Let $Y \in \{1,2,3,4,5,6\}$ and $P(Y=j)=\pi_j$ where j=1,2,3,4,5,6. Then we have Multinoulli

$$P(\mathbf{Y}|\pi) = Multinoulli(\mathbf{Y}|\pi) = \prod_{j=1}^{6} {\pi_j}^{I(Y=j)}$$

where $I(Y=j)=1$ if $Y=j$ and $I(Y=j)=0$ otherwise.
Assuming i.i.d, the likelihood for a sequence of N trials, $\mathbf{D}=(y_1,y_2,..,y_N)$.

$$P(\mathbf{D}|\pi) = \prod_{n=1}^{N}\prod_{j=1}^{6} {\pi_j}^{I(y_n=j)} = \prod_{j=1}^{6} {\pi_j}^{N_j}$$

where $N_j = \sum_{n=1}^{N} I(y_n=j)$ is the number of times Y=j.

MAP ESTIMATION

For Multinoulli likelihood **Dirichlet** will be a good choice of prior.

$$P(\pi) = Dirichlet(\pi:\alpha) = \frac{\Gamma(\sum_{j=1}^{6}\alpha_j)}{\prod_{j=1}^{6}\Gamma(\alpha_j)} \prod_{j=1}^{6} {\pi_j}^{\alpha_j-1}$$

where $\alpha=[\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6]$ is vector of positive real numbers.

From Bayes theorem we have,

$$P(\pi|\mathbf{D}) = \frac{P(\pi)P(\mathbf{D}|\pi)}{P(\mathbf{D})}$$

MAP estimation approach reports the maxima of the posterior i.e $P(\pi|\mathbf{D})$ needs to be maximised and since $P(\mathbf{D})$ is constant w.r.t π ,we get

$$\mathcal{L}(\pi) = \arg\max_{\pi}[P(\pi|\mathbf{D})] = \arg\max_{\pi}[P(\pi)P(\mathbf{D}|\pi)]$$

By ignoring the terms of $P(\pi)$ which doesn't depend on π we get

$$\mathcal{L}(\pi) = \arg\max_{\pi}[\prod_{j=1}^{6} {\pi_j}^{N_j} \prod_{j=1}^{6} {\pi_j}^{\alpha_j-1}] = \arg\max_{\pi}[\prod_{j=1}^{6} {\pi_j}^{\alpha_j+N_j-1}]$$

Taking log of $\mathcal{L}(\pi)$ we get,

$$L\mathcal{L}(\pi) = \arg\max_{\pi}[\log\prod_{j=1}^{6} {\pi_j}^{\alpha_j+N_j-1}] = \arg\max_{\pi}[\sum_{j=1}^{6}(\alpha_j+N_j-1)\log\pi_j]$$

For the constraint $\sum_{j=1}^{6} \pi_j = 1$ **Lagrange multiplier** can be used, then constrained cost function becomes

$$L\mathcal{L}(\pi) = \arg\max_{\pi}[\sum_{j=1}^{6}(\alpha_j + N_j - 1)\log \pi_j + \lambda[1 - \sum_{j=1}^{6} \pi_j]]$$

$$\frac{\partial L\mathcal{L}(\pi)}{\partial \pi_j} = 0$$

$$\implies \frac{\alpha_j + N_j - 1}{\pi_j} + \lambda(-1) = 0 \implies \frac{\alpha_j + N_j - 1}{\lambda} = \pi_j$$

Also, $\frac{\partial L\mathcal{L}(\pi)}{\partial \lambda} = 0$

$\implies 1 - \sum_{j=1}^{6} \pi_j = 0 \implies \sum_{j=1}^{6} \pi_j = 1$

So we get, $\alpha_j + N_j - 1 = \lambda\pi_j \implies \sum_{j=1}^{6} \alpha_j + N_j - 1 = \sum_{j=1}^{6} \lambda\pi_j$

$\implies N + \sum_{j=1}^{6} \alpha_j - 6 = \lambda$

$\therefore \pi_{MAP} = \frac{\alpha_j + N_j - 1}{N + \sum_{j=1}^{6} \alpha_j - 6}$

**MAP solution would be better than the MLE solution in case our data set is small and we don't have $Y = j$ for some j , in that case $\pi_{MLE} = N_j/N$ will be 0.

FULL POSTERIOR

$$P(\pi|\mathbf{D}) = \frac{\left(\frac{\Gamma(\sum_{j=1}^{6}\alpha_j)}{\prod_{j=1}^{6}\Gamma(\alpha_j)}\prod_{j=1}^{6}{\pi_j}^{\alpha_j-1}\right)\left(\prod_{n=1}^{N}\prod_{j=1}^{6}{\pi_j}^{I(y_n=j)}\right)}{\int_{\pi}\left(\frac{\Gamma(\sum_{j=1}^{6}\alpha_j)}{\prod_{j=1}^{6}\Gamma(\alpha_j)}\prod_{j=1}^{6}{\pi_j}^{\alpha_j-1}\right)\left(\prod_{n=1}^{N}\prod_{j=1}^{6}{\pi_j}^{I(y_n=j)}\right)d\pi}$$

Since Dirichlet and Multinoulli forms Conjugate pairs, entire integral in the denominator together with terms independent of π in numerator can be taken as proportionality constant.

$$\propto \left(\prod_{j=1}^{6}{\pi_j}^{\alpha_j-1}\right)\left(\prod_{j=1}^{6}{\pi_j}^{N_j}\right) \propto \left(\prod_{j=1}^{6}{\pi_j}^{N_j+\alpha_j-1}\right)$$

i.e **Dirichlet**($\pi : \alpha_j + N_j$), where j=1,2,3,4,5,6.

Hence π_{FP} is Expectation of π under this Dirichlet distribution $= \frac{\alpha_j + N_j}{\sum_{j=1}^{6}\alpha_j + N_j} = \frac{\alpha_j + N_j}{N + \sum_{j=1}^{6}\alpha_j}$

**Clearly, given this posterior, we can estimate both MLE and MAP without solving the MLE and MAP optimization problems explicitly. MLE estimate (Nj/N) is equivalent to full posterior with $\alpha_j = 0$ and MAP estimate (π_{MAP}) is equivalent to full posterior with $\alpha_j = \alpha_j - 1$ where $j = 1, 2, 3, 4, 5, 6$.