Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION 1

Assignment Number: 1
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: September 10, 2017

Let $\mathbf{z}^1$ = (1,0)(*green*) and $\mathbf{z}^2$ = (0,1)(*red*) and $\mathbf{z}$ =(x,y) be any arbitrary point on the decision boundary. Let M be a Mahalanobis metric. Now every point on decision boundary follows:

$$d(\mathbf{z}, \mathbf{z}^1) = d(\mathbf{z}, \mathbf{z}^2)$$
$$\left\langle \mathbf{z} - \mathbf{z}^1, M(\mathbf{z} - \mathbf{z}^1) \right\rangle = \left\langle \mathbf{z} - \mathbf{z}^2, M(\mathbf{z} - \mathbf{z}^2) \right\rangle$$
$$(\mathbf{z} - \mathbf{z}^1)^T M^T (\mathbf{z} - \mathbf{z}^1) = (\mathbf{z} - \mathbf{z}^2)^T M^T (\mathbf{z} - \mathbf{z}^2)$$

1. Here $M = U$ where $U = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$

$$\begin{bmatrix} x-1 \\ y \end{bmatrix}^T \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}^T \begin{bmatrix} x-1 \\ y \end{bmatrix} = \begin{bmatrix} x \\ y-1 \end{bmatrix}^T \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}^T \begin{bmatrix} x \\ y-1 \end{bmatrix}$$
$$\begin{bmatrix} 3(x-1) & y \end{bmatrix} \begin{bmatrix} x-1 \\ y \end{bmatrix} = \begin{bmatrix} 3x & (y-1) \end{bmatrix} \begin{bmatrix} x \\ y-1 \end{bmatrix}$$
$$3(x-1)^2 + y^2 = 3x^2 + (y-1)^2$$
$$\mathbf{y = 3x - 1}$$

In this case mathematical expression for the decision boundary is a linear equation in 2-D representing a straight line $\mathbf{y = 3x - 1}$.

2. Here $M = V$ where $V = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

$$\begin{bmatrix} x-1 \\ y \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}^T \begin{bmatrix} x-1 \\ y \end{bmatrix} = \begin{bmatrix} x \\ y-1 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}^T \begin{bmatrix} x \\ y-1 \end{bmatrix}$$
$$\begin{bmatrix} (x-1) & 0 \end{bmatrix} \begin{bmatrix} x-1 \\ y \end{bmatrix} = \begin{bmatrix} x & 0 \end{bmatrix} \begin{bmatrix} x \\ y-1 \end{bmatrix}$$
$$(x-1)^2 = x^2$$
$$\mathbf{x = 1/2}$$

In this case mathematical expression for the decision boundary is a linear equation in 2-D representing a straight line $\mathbf{x = 1/2}$.

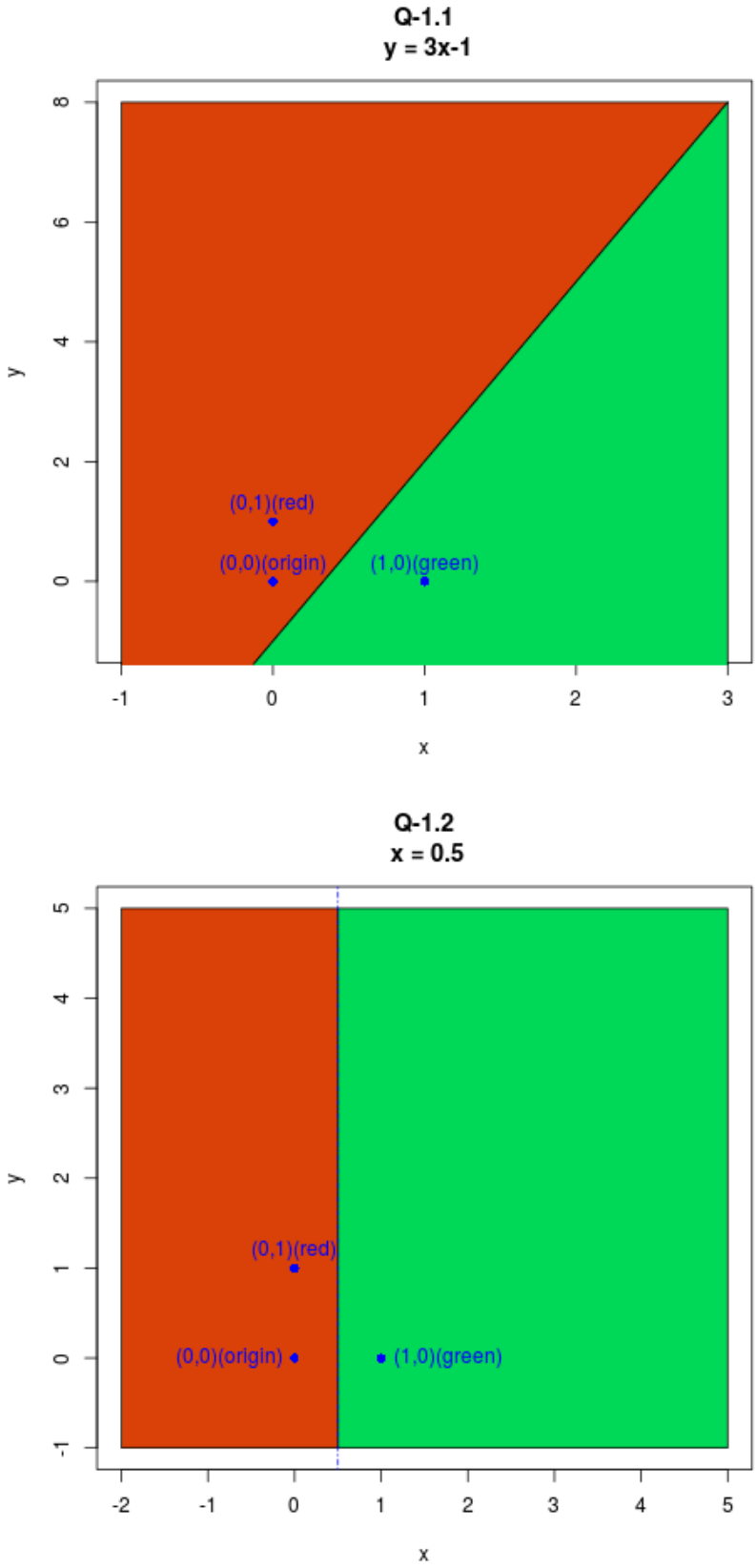


Figure 1: These figures corresponds to the decision boundary of Question-1. Top one corresponds to Q1-1 part representing straight line y = 3x-1 and below one corresponds to Q1-2 part representing x = 1/2.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION

2

Assignment Number: 1
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: September 10, 2017

Constraint given for problem is $\|\mathbf{w}\|_2 \leq r$.
We can write it as $\|\mathbf{w}\|_2 = k$ s.t k = r - δ and $0 \leq \delta \leq r$ .Now

$$\|\mathbf{w}\|_2 = k$$
$$(\|\mathbf{w}\|_2 - k)^2 = 0$$

Lagrangian solution for $\hat{\mathbf{w}}_{\text{cls}}$ will be :

$$\hat{\mathbf{w}}_{\text{cls}} = \arg\min \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \lambda(\|\mathbf{w}\|_2 - k)^2$$

$$\bigtriangledown \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \bigtriangledown \lambda(\|\mathbf{w}\|_2 - k)^2 = \bigtriangledown \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + 2\lambda(\|\mathbf{w}\|_2 - k)$$

Likelihood distribution $\mathbb{P}\left[y^i | \mathbf{x}, \mathbf{w}\right] = \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2}{2\sigma^2}\right)$ and $\mathbb{P}\left[\mathbf{y} | \mathbf{X}, \mathbf{w}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{y^i} | \mathbf{x^i}, \mathbf{w}\right]$

Prior distribution $\mathbb{P}\left[\mathbf{w}\right] = \mathcal{N}(k.\mathbf{1}_d, \rho^2.\mathbf{I}_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp\left(-\frac{(\|\mathbf{w} - k.\mathbf{1}_d\|_2)^2}{2\rho^2}\right)$ where k = r - δ and $0 \leq \delta \leq r$

Now for our functions we have

$$\hat{\mathbf{w}}_{\text{new}} = \arg\min \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \lambda(\|\mathbf{w} - k\|_2)^2$$

$$\bigtriangledown \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \bigtriangledown \lambda(\|\mathbf{w} - k\|_2)^2 = \bigtriangledown \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + 2\lambda(\|\mathbf{w}\|_2 - k)$$

We have now $\hat{\mathbf{w}}_{\text{cls}} = \hat{\mathbf{w}}_{\text{new}}$. Hence these(above) are our required functions.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 1
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: September 10, 2017

QUESTION 3

Let $\forall$ i ϵ [n] we have,

$$x_j^{i\prime} = \frac{x_j^i}{\sqrt{\alpha_j}} \forall \text{ j } \epsilon[d]$$
$$y^{i\prime} = y_i$$

Considering **Likelihood distribution** $\mathbb{P}\left[y^{i\prime}|\mathbf{x}\prime, \mathbf{w}\prime\right] = \mathcal{N}(\left\langle \mathbf{w}\prime, \mathbf{x}^{i\prime}\right\rangle, \sigma^2)$ and **Prior distribution** $\mathbb{P}\left[\mathbf{w}\prime\right] = \mathcal{N}(\mathbf{0}, \rho^2.\mathbf{I}_d)$ we have

$$(\hat{\mathbf{w}}_{\text{fr}})\prime = \arg\min \sum_{i=1}^{n}(y^{i\prime} - \left\langle \mathbf{w}\prime, \mathbf{x}^{i\prime}\right\rangle)^2 + \sum_{j=1}^{d}(\mathbf{w}\prime_j)^2$$
$$(\hat{\mathbf{w}}_{\text{fr}})\prime = (X'X^{T\prime} + \lambda I)^{-1}X'y\prime$$

Now $(\hat{\mathbf{w}}_{\text{fr}})\prime$ is a d X 1 matrix. Let's say

$$(\mathbf{w})\prime_j = \sqrt{\alpha_j}\mathbf{w}_j \ \forall \text{ j } \epsilon \ [d]$$

Now

$$\begin{aligned}\left\langle \mathbf{w}\prime, \mathbf{x}^{i\prime}\right\rangle &= \sum_{j=1}^{d} x_j^{i\prime}\mathbf{w}\prime_j \\ &= \frac{x_j^i}{\sqrt{\alpha_j}} * \sqrt{\alpha_j}\mathbf{w}_j \\ &= \sum_{j=1}^{d} x_j^i\mathbf{w}_j \\ \left\langle \mathbf{w}\prime, \mathbf{x}^{i\prime}\right\rangle &= \left\langle \mathbf{w}, \mathbf{x}^i\right\rangle\end{aligned}$$

Thus by substituting

$$(\hat{\mathbf{w}}_{\text{fr}})\prime = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \ \sum_{i=1}^{n}(y^i - \left\langle \mathbf{w}, \mathbf{x}^i\right\rangle)^2 + \sum_{j=1}^{d}\alpha_i(\mathbf{w}_i)^2$$

We now need $\hat{\mathbf{w}}_{\text{fr}}$ which is

$$\begin{aligned}(\hat{\mathbf{w}}_{\text{fr}})_j &= \frac{(\hat{\mathbf{w}}_{\text{fr}})\prime_j}{\sqrt{\alpha_j}} \forall \text{ j } \epsilon[d] \\ &= \frac{[(X'X^{T\prime} + \lambda I)^{-1}X'y\prime]_j}{\sqrt{\alpha_j}}\end{aligned}$$

Finally we have **Likelihood distribution** $\mathbb{P}\left[y^i|\mathbf{x\prime}, \mathbf{w\prime}\right] = \mathcal{N}(\left\langle \mathbf{w\prime}, \mathbf{x\prime}^i \right\rangle, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{i\prime} - \left\langle \mathbf{w\prime}, \mathbf{x}^{i\prime} \right\rangle)^2}{2\sigma^2}\right)$ and $\mathbb{P}\left[\mathbf{y}|\mathbf{X\prime}, \mathbf{w\prime}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{y^i}|\mathbf{x^{i\prime}}, \mathbf{w\prime}\right]$

Prior distribution $\mathbb{P}\left[\mathbf{w}\right] = \mathcal{N}(\mathbf{0}, \rho^2.\mathbf{I}_d) = \frac{1}{\sqrt{(2\pi)^d \rho^2}} \exp\left(-\frac{(\|\mathbf{w}\|_2)^2}{2\rho^2}\right)$

Closed-form expression for $\hat{\mathbf{w}}_{\text{fr}}$ is given by

$$(\hat{\mathbf{w}}_{\text{fr}})_j = \frac{[(X\prime X^{T\prime} + \lambda I)^{-1} X\prime y\prime]_j}{\sqrt{\alpha_j}} \forall \text{ j } \epsilon[d]$$

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION

4

Assignment Number: 1
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: September 10, 2017

For the first case ie. $\left\{\mathbf{W}^0, \{\xi_i^0\}\right\}$ are an optimum for (P_1) then shall prove that $\mathbf{W}^0$ must be an optimum for $(P2)$

$\left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i, \forall i, \forall k \neq y^i, \xi_i \geq 0$

$\xi_i \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle, \forall i, \forall k \neq y^i, \xi_i \geq 0$

$\xi_i \geq \max_{k \neq y} \left[\left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \right], \forall i, \xi_i \geq 0$

$\xi_i \geq 1 + \max_{k \neq y} \left[\left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \right], \forall i, \xi_i \geq 0$

$\xi_i \geq [1 + \max_{k \neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i]_+, \forall i$

Now since we have to make loss minimum possible hence we have

$\xi_i = [1 + \max_{k \neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i]_+, \forall i$

It is known that

$$\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) = [1 + \max_{k \neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i]_+$$

Hence

$$\xi_i = \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)$$

Hence proved that our first case.

For the second case we have to prove that if $\mathbf{W}^1$ is an optimum for $(P2)$ then there must exist $\left\{\xi_i^1\right\} \geq 0$ such that $\left\{\mathbf{W}^1, \{\xi_i^1\}\right\}$ are an optimum for (P_1)

$$\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) = [1 + \max_{k \neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i]_+, \forall i$$

$$\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) = [1 + \max_{k \neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i], \ell_{\text{cs}} \geq 0, \forall i$$

$$\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) = 1 + \max_{k \neq y} \left[\left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \right], \forall i, \ell_{\text{cs}} \geq 0$$

$$\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) = 1 + \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle, \forall i, \forall k \neq y^i, \ell_{\text{cs}} \geq 0$$

Therefore there exists

$$\xi_i = \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)$$

Hence if $\mathbf{W}^1$ is an optimum for $(P2)$ then there is $\left\{\xi_i^1\right\} \geq 0$ such that $\left\{\mathbf{W}^1, \{\xi_i^1\}\right\}$ are an optimum for (P_1)

QUESTION
5

Assignment Number: 1
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: September 10, 2017

We have

$$f(\mathbf{w}) = \sum_{i=1}^{n} [1 - y^i p \mathbf{w} \mathbf{x}^i]_+$$

Let's consider it for one point ie. for n = 1 we get

$$f(\mathbf{w}) = [1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle]_+$$

$$f(\mathbf{w}) = \begin{cases} 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle & \text{if } \; y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle < 1 \\ \mathbf{0} & \text{if } \; y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1, \end{cases}$$

We also get $\mathbf{g} = \mathrm{h}^i$

$$\mathbf{g} = \mathbf{h}^i = \begin{cases} -y^i \cdot \mathbf{x}^i & \text{if } \; y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle < 1 \\ \mathbf{0} & \text{if } \; y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1, \end{cases}$$

Consider for $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle < 1$ we have

$$\begin{aligned} f(\mathbf{w}) + \left\langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \right\rangle &= 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + \left\langle -y^i \cdot \mathbf{x}^i, \mathbf{w}' - \mathbf{w} \right\rangle \\ &= 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle \\ &= 1 - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle \\ f(\mathbf{w}) + \left\langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \right\rangle &= f(\mathbf{w}') \end{aligned}$$

Thus here it satisfies $f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$.

Consider for $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1$ we have

$$\begin{aligned} f(\mathbf{w}) + \left\langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \right\rangle &= 0 + \left\langle 0, \mathbf{w}' - \mathbf{w} \right\rangle \\ &= 0 \end{aligned}$$

Now $f(\mathbf{w}') \geq 0$. Hence it satisfies $f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$.

Hence $\mathbf{g} \in \partial f(\mathbf{w})$ i.e. $\mathbf{g}$ is a member of the subdifferential of f at $\mathbf{w}$. As we have shown that for every $\mathbf{w}' \in \mathbb{R}^d$, $f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$.

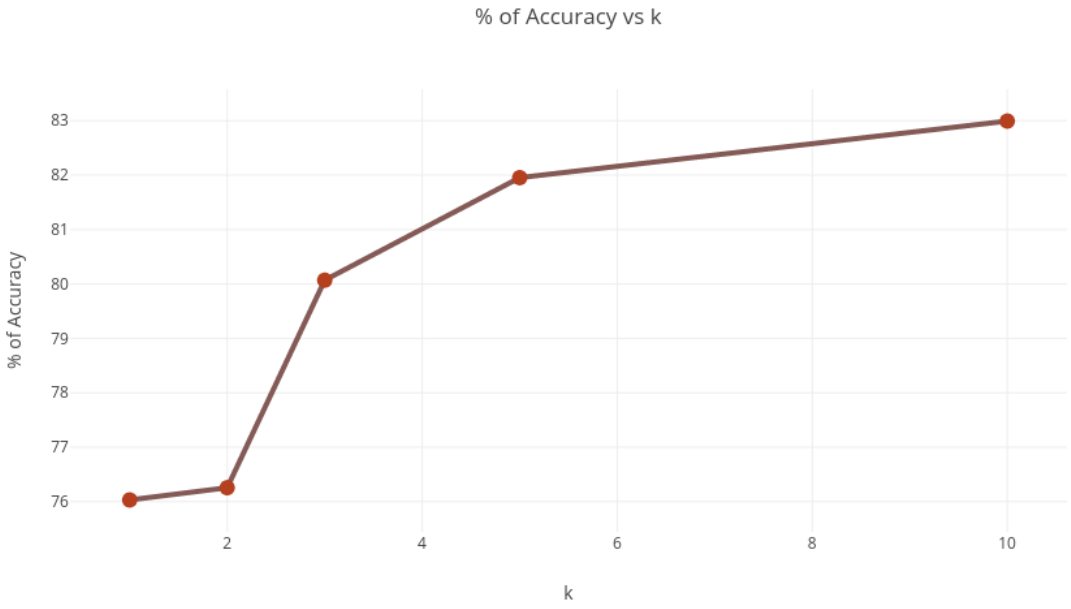
Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 1
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: September 10, 2017

QUESTION 6

1. Test error table for the 20K sized dataset

Value of k	No. of mistakes	% of error	% Accuracy
1	4794	23.97	76.03
2	4749	23.745	76.255
3	3986	19.93	80.07
5	3609	18.045	81.955
10	3401	17.005	82.995



We observe that when value of k is smaller it tends to overfits data as it takes the value of its nearest neighbour which tends to give us huge errors because of outliers and **overfitting**. As our k increases accuracy increases as it becomes more and more aware of false data points and hence increase accuracy. We also observe that if we keep on increasing value our accuracy tends to decrease because of **underfitting**. So moderate value of k gives best results.

2. Here we uses 3-fold validation technique ie. take first one-third and second one-third of the dataset points for the first round, for second round take all points except for first one-third and for the third(last) round take all points except second one-third of points. We will take value of k to be $\mathbf{k = 10}$.

3. Value of chosen k is $\mathbf{k = 10}$ and accuracy corresponding to it is **83.75%**.