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

Assignment Number: 1
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: September 15, 2017

QUESTION

1

Part 1

Let $z = (x_1, y_1)$. We are given $z_r = (0, 1)$ and $z_g = (1, 0)$.
For the decision boundary, $d(z, z_r) = d(z, z_g)$

$\implies \langle z - z_r, U(z - z_r)\rangle = \langle z - z_g, U(z - z_g)\rangle$ where $U = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$

$\implies 3x_1^2 + (y_1 - 1)^2 = 3(x_1 - 1)^2 + y_1^2$
$\implies 3x_1^2 - 3(x_1 - 1)^2 = y_1^2 - (y_1 - 1)^2$
$\implies 6x_1 - 3 = 2y_1 - 1$
$\implies 3x_1 = y_1 + 1$
$\implies 3x = y + 1$ is the decision boundary.

Part 2

For the decision boundary, $d(z, z_r) = d(z, z_g)$

$\implies \langle z - z_r, U(z - z_r)\rangle = \langle z - z_g, U(z - z_g)\rangle$ where $U = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

$\implies x_1^2 = (x_1 - 1)^2$
$\implies 2x_1 - 1 = 0$
$\implies x_1 = \frac{1}{2}$
$\implies x = \frac{1}{2}$ is the decision boundary.

Figure 1: The figure on the left shows the decision boundary for part 1 i.e $3x = y + 1$.
The figure on the right shows the decision boundary for part 2 i.e $x = \frac{1}{2}$.

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

Assignment Number: 1
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: September 15, 2017

QUESTION 2

One likelihood distribution which leads to $\hat{\mathbf{w}}_{cls}$ as the MAP estimate for the model is the gaussian distribution with mean as $\langle \mathbf{w}, \mathbf{x}^i \rangle$

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

The prior distribution encapsulates the fact that $||\mathbf{w}||_2 \leq r$. This can be achieved by a uniform distribution over the d-dimensional ball with origin as centre and radius r.

$$\mathbb{P}[\mathbf{w}] = \mathcal{U}(\mathbf{0}, r) = \begin{cases} \frac{\Gamma(\frac{d}{2}+1)}{\pi^{\frac{d}{2}} r^d} & \text{if } ||\mathbf{w}||_2 \leq r \\ 0 & \text{if } ||\mathbf{w}||_2 > r \end{cases}$$

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

Assignment Number: 1
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: September 15, 2017

QUESTION 3

The likelihood part of the estimate is same as the previous question, therefore one likelihood distribution is the gaussian distribution with mean as $\langle \mathbf{w}, \mathbf{x}^i \rangle$

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

For the prior, we can take a multivariate gaussian distribution with mean at $\mathbf{0}$

$$\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{w}; \mathbf{0}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(- \frac{1}{2} \mathbf{w}^T \Sigma^{-1} \mathbf{w} \right)$$

Here $\Sigma = \sigma^2 \begin{bmatrix} \frac{1}{\alpha_1} & 0 & \dots & \dots & 0 \\ 0 & \frac{1}{\alpha_2} & 0 & \dots & 0 \\ \vdots & 0 & \ddots & 0 \dots & 0 \\ \vdots & \vdots & \vdots & \frac{1}{\alpha_{d-1}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\alpha_d} \end{bmatrix}$ i.e a diagonal matrix with it's i'th diagonal entry as $\frac{\sigma^2}{\alpha_i}$ where σ^2 is the variance of the likelihood distribution.

$$\log \mathbb{P}[\mathbf{w}|\mathbf{X}, \mathbf{y}] = C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{d} \alpha_j \mathbf{w}_j^2$$

$$\implies \hat{\mathbf{w}}_{\text{MAP}} = \arg\min \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \sum_{j=1}^{d} \alpha_j \mathbf{w}_j^2$$

Closed form expression:

We need to minimise

$$\mathcal{L} = \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \sum_{j=1}^{d} \alpha_j \mathbf{w}_j^2$$

$$\mathcal{L} = ||\mathbf{X}\mathbf{w} - \mathbf{Y}||^2 + \sum_{j=1}^{d} \alpha_j \mathbf{w}_j^2$$

$\nabla_w \mathcal{L} = 2\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{Y}) + 2\mathbf{D}_\alpha \mathbf{w}$ where D_α is the diagonal matrix with entries $\alpha_1, \alpha_2 \dots \alpha_d$

$$\nabla_w \mathcal{L} = 2((\mathbf{X^T X} + \mathbf{D}_\alpha)\mathbf{w} - \mathbf{X^T Y})$$

$$\nabla_w \mathcal{L} = 0$$

$$\iff (\mathbf{X^T X} + \mathbf{D}_\alpha)\mathbf{w} - \mathbf{X^T Y} = 0$$

$$\iff (\mathbf{X^T X} + \mathbf{D}_\alpha)\mathbf{w} = \mathbf{X^T Y}$$

$$\iff w = (\mathbf{X^T X} + \mathbf{D}_\alpha)^{-1} \mathbf{X^T Y}$$

QUESTION 4

Assignment Number: 1
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: September 15, 2017

The constraints for $(P1)$ can be re-written as

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

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

$$\implies \xi_i \geq 1 + \max_{k \neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i \text{ and } \xi_i \geq 0 \forall i$$

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

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

Therefore, for a fixed value of $\mathbf{W}$, if we want to minimise the given function in $(P1)$, we have to set $\xi_i = \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)$ for every i to satisfy the given constraint.
For optimum $\{\mathbf{W}, \{\xi_i\}\}$, we have $\xi_i = \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i) \forall i$

Part 1

$\left\{\mathbf{W}^0, \{\xi_i^0\}\right\}$ is an optimum for (P_1)
$\implies$ For any other $\left\{\mathbf{W}^2, \{\xi_i^2\}\right\}$, we have

$$\left(\sum_{k=1}^{K} \left\|\mathbf{w}_2^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i^2\right) - \left(\sum_{k=1}^{K} \left\|\mathbf{w}_0^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i^0\right) \geq 0 \text{ where } \xi_i \text{ are subject to the given constraints}$$

$$\implies \left(\sum_{k=1}^{K} \left\|\mathbf{w}_2^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i^2\right) - \left(\sum_{k=1}^{K} \left\|\mathbf{w}_0^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)_{\mathbf{W}_0}\right) \geq 0 \text{ since } \xi_i^0 = \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)_{\mathbf{W}_0}$$

$$\implies \left(\sum_{k=1}^{K} \left\|\mathbf{w}_2^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)_{\mathbf{W}_2}\right) - \left(\sum_{k=1}^{K} \left\|\mathbf{w}_0^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)_{\mathbf{W}_0}\right) \geq 0$$

since $\left\{\xi_i^2\right\} = \left\{\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)_{\mathbf{W}_2}\right\}$ is also a valid value in the inequality before this.

$$\implies \mathbf{W}^0 \text{ is an optimum for } (P2)$$

Part 2

$\{\mathbf{W}^1\}$ is an optimum for (P_2).
$\implies$ For any other $\{\mathbf{W}^2\}$, we have

$$\implies \left(\sum_{k=1}^{K}\left\|\mathbf{w}_2^k\right\|_2^2+\sum_{i=1}^{n}\ell_{\text{cs}}(y^i,\boldsymbol{\eta}^i)_{\mathbf{W}_2}\right)-\left(\sum_{k=1}^{K}\left\|\mathbf{w}_1^k\right\|_2^2+\sum_{i=1}^{n}\ell_{\text{cs}}(y^i,\boldsymbol{\eta}^i)_{\mathbf{W}_1}\right)\geq 0$$

We have shown earlier that

$$\left(\sum_{k=1}^{K}\left\|\mathbf{w}_2^k\right\|_2^2+\sum_{i=1}^{n}\xi_i^2\right)\geq\left(\sum_{k=1}^{K}\left\|\mathbf{w}_2^k\right\|_2^2+\sum_{i=1}^{n}\ell_{\text{cs}}(y^i,\boldsymbol{\eta}^i)_{\mathbf{W}_2}\right)$$

where ξ_i are subject to the given constraints.

$$\therefore \left(\sum_{k=1}^{K}\left\|\mathbf{w}_2^k\right\|_2^2+\sum_{i=1}^{n}\xi_i^2\right)-\left(\sum_{k=1}^{K}\left\|\mathbf{w}_1^k\right\|_2^2+\sum_{i=1}^{n}\ell_{\text{cs}}(y^i,\boldsymbol{\eta}^i)_{\mathbf{W}_1}\right)\geq 0$$

$$\implies \left(\sum_{k=1}^{K}\left\|\mathbf{w}_2^k\right\|_2^2+\sum_{i=1}^{n}\xi_i^2\right)-\left(\sum_{k=1}^{K}\left\|\mathbf{w}_1^k\right\|_2^2+\sum_{i=1}^{n}\xi_i^1\right)\geq 0 \text{ where } \left\{\xi_i^1\right\}=\left\{\ell_{\text{cs}}(y^i,\boldsymbol{\eta}^i)_{\mathbf{W}_1}\right\}\geq 0$$

$$\implies \exists\left\{\xi_i^1\right\}\geq 0 \text{ such that } \left\{\mathbf{W}^1,\{\xi_i^1\}\right\} \text{ is an optimum for } (P_1)$$

QUESTION

5

Assignment Number: 1
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: September 15, 2017

To prove: $\forall \mathbf{w}' \in \mathbb{R}^d, f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$

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

where

$$[1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle]_+ = \begin{cases} 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle & \text{if } y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle < 1 \\ 0 & \text{if } y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1 \end{cases}$$

The given vector $\mathbf{g}$ is defined as $\mathbf{g} = \sum_{i=1}^{n} \mathbf{h}^i$ where

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

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

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

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

For one value of i, we get four cases:
Case 1: $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle < 1$, $y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle < 1$
In this case, the i'th term of $f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$ summation will be
$1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + \left\langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle$
$= 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + \left\langle -y^i.\mathbf{x}^i, \mathbf{w}' - \mathbf{w} \right\rangle$
$= 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + (-y^i).\left\langle \mathbf{x}^i, \mathbf{w}' - \mathbf{w} \right\rangle$
$= 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + (-y^i).\left\langle \mathbf{w}' - \mathbf{w}, \mathbf{x}^i \right\rangle$
$= 1 - y^i \left(\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + \left\langle \mathbf{w}' - \mathbf{w}, \mathbf{x}^i \right\rangle \right)$
$= 1 - y^i \left\langle \mathbf{w} + \mathbf{w}' - \mathbf{w}, \mathbf{x}^i \right\rangle$
$= 1 - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle$
This is the same as the i'th term of $f(\mathbf{w}')$ summation.

Case 2: $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle < 1$, $y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle \geq 1$
In this case, the i'th term of $f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$ summation will be
$1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + \left\langle \mathbf{h}^i, \mathbf{w}' - \mathbf{w} \right\rangle$
$= 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle + \left\langle -y^i.\mathbf{x}^i, \mathbf{w}' - \mathbf{w} \right\rangle$
$= 1 - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle$
This is the same as the i'th term of $f(\mathbf{w}')$ summation.

Case 3: $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1$, $y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle < 1$
In this case, the i'th term of $f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$ summation will be
$0 + \langle \mathbf{0}, \mathbf{w}' - \mathbf{w} \rangle$
$= 0$
This is the less than the i'th term of $f(\mathbf{w}')$ summation i.e $1 - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle$.

Case 4: $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1$, $y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle \geq 1$
In this case, the i'th term of $f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$ summation will be
$0 + \langle \mathbf{0}, \mathbf{w}' - \mathbf{w} \rangle$
$= 0$
This is the same as the i'th term of $f(\mathbf{w}')$ summation.
Thus in 3 of the 4 cases, the terms on LHS and RHS are equal.
For the case 3, LHS is strictly greater than RHS.
Therefore, $\forall \mathbf{w}' \in \mathbb{R}^d, f(\mathbf{w}') \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{w}' - \mathbf{w} \rangle$

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

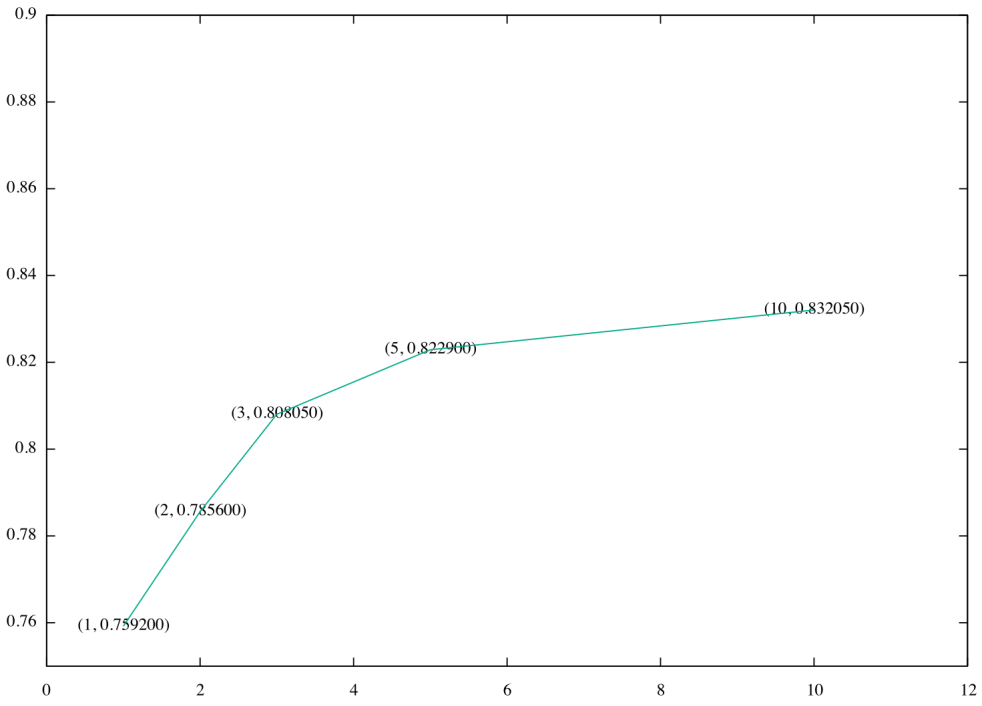
Assignment Number: 1
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: September 15, 2017

QUESTION

6

Part 1

k	Test Error(#points)
1	4816
2	4288
3	3839
5	3542
10	3359



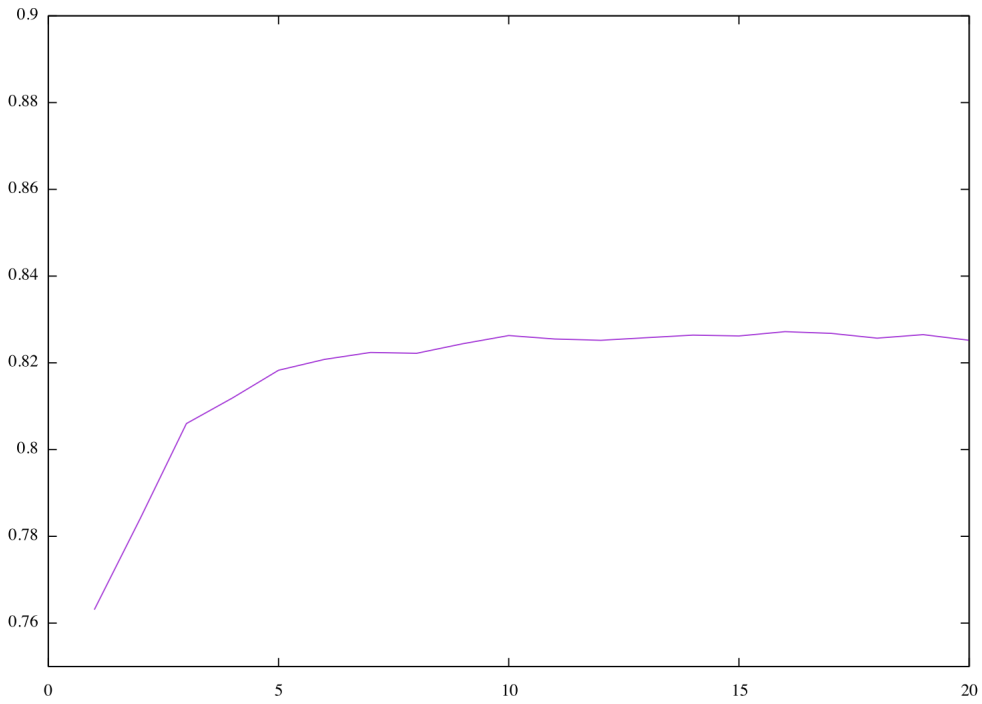
For lower values of k, the model overfits the training data. Therefore, due to the noise in the training data, the model has lesser accuracy for smaller values of k.

Part 2

I used holdout validation for tuning the value of the hyperparameter k. I took 50K random points from the given training data as the training set and the remaining 10K points as the validation set. We can see from the graph that there is not much change between the values from

k=10 to k=20. Therefore, I took the value of k=10 for faster computation although k=16(the maxima) gives slightly better performance.

k	Accuracy	k	Accuracy	k	Accuracy	k	Accuracy
1	0.7631	2	0.7842	3	0.8060	4	0.8119
5	0.8183	6	0.8208	7	0.8224	8	0.8222
9	0.8244	10	0.8263	11	0.8255	12	0.8252
13	0.8258	14	0.8264	15	0.8262	16	0.8272
17	0.8268	18	0.8257	19	0.8265	20	0.8252



Part 3

For the training of the LMNN model, I took 6K test points which were randomly selected from the given training set. I set the maxiter to 200,000. My model converged before the value reached 200,000.
I got a test accuracy of 83.585% for my metric with k=10 when I ran it on all the given 20K test points with training data as all the 60K points. Thus, we can see that we get better performance with the metric when compared to normal KNN(accuracy for normal k-NN with k=10 for the entire training set is 83.205% according to part 1).

Extra credit

I trained my ITML metric by giving it 6K random points from the train.dat set
I got an accuracy of 82.68% by using this learned metric on the 60K training points with 20K test points.