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

Assignment Number: 1
Student Name: Raktim Mitra
Roll Number: 150562
Date: September 10, 2017

QUESTION 1

Solution for d_U:

Let, $\mathbf{z_1} = (x_1, y_1)$ and $\mathbf{z_2} = (x_2, y_2)$ then $d_U(\mathbf{z_1}, \mathbf{z_2}) = \langle (z_1 - z_2), U(z_1, z_2) \rangle$. Upon calculating the inner produc we get:

$d_U(\mathbf{z_1}, \mathbf{z_2}) = 3(x_1 - x_2)^2 + (y_1 - y_2)^2$

The decision boundary is the locus of points $\mathbf{z}$ with equal distance from both $\mathbf{z_1}$ and $\mathbf{z_2}$. i.e.

$3(x_1 - x)^2 + (y_1 - y)^2 = 3(x_2 - x)^2 + (y_2 - y)^2$

Putting $\mathbf{z_1} = (1, 0)$ and $\mathbf{z_2} = (0, 1)$:

$3(x - 1)^2 + y^2 = 3x^2 + (y - 1)^2$

$\implies -6x - 1 = -2y + 1$

$\implies 2y = 6x - 2$

$\implies y = 3x - 1$ [Answer]

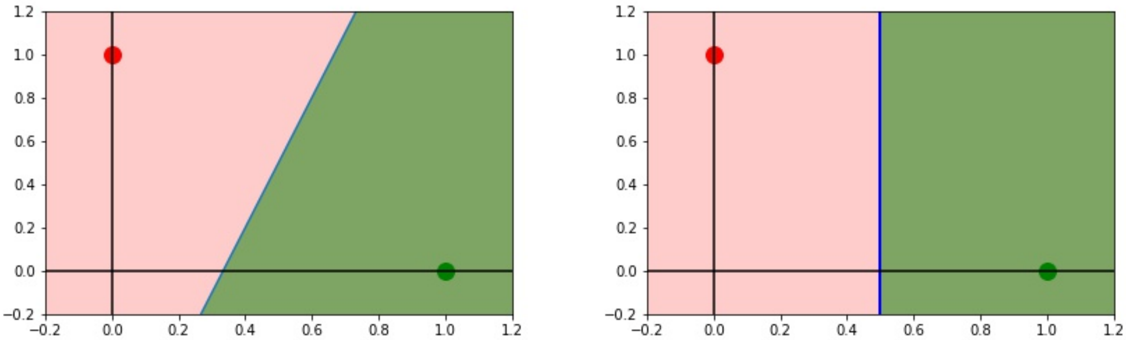

Figure 1: Learning with Prototypes: The figure on the left shows the decision boundary with d_U distance metric. The decision boundary is $y = 3x - 1$. The figure on the right shows the decision boundary with d_V distance metric. The decision boundary is $x = 0.5$.

Solution for d_V:

Let, $\mathbf{z_1} = (x_1, y_1)$ and $\mathbf{z_2} = (x_2, y_2)$ then $d_V(\mathbf{z_1}, \mathbf{z_2}) = \langle (z_1 - z_2), V(z_1, z_2) \rangle$. Upon calculating the inner product we get:

$d_V(\mathbf{z_1}, \mathbf{z_2}) = (x_1 - x_2)^2$ (Only depends on x !)

The decision boundary is the locus of points $\mathbf{z}$ with equal distance from both $\mathbf{z_1}$ and $\mathbf{z_2}$. i.e.

$(x_1 - x)^2 = (x_2 - x)^2$

Putting $\mathbf{z_1} = (1, 0)$ and $\mathbf{z_2} = (0, 1)$:

$3(x - 1)^2 = 3x^2$

$\implies 2x = 1$

$\implies x = \frac{1}{2}$ [Answer]

QUESTION 2

Assignment Number: 1
Student Name: Raktim Mitra
Roll Number: 150562
Date: September 10, 2017

We can incorporate the constraint into the prior distribution by setting prior probaility to 0 if $\|\mathbf{w}\| > r$. We also want $\mathbf{log(P(w))}$ to vanish since we want a MAP estimate that matches the given estimation function which resembles only an MLE, i.e. a MAP with constant prior. so for $\|\mathbf{w}\| \leq r$ we set the prior to be uniform distribution. The value of this constant in the probability density function is $\frac{1}{\rho}$ where ρ is the volume of hipersphere of radius r. This preserves the total probability to 1. Therefore:

$$\textbf{PDF } \mathbf{P(w)} = \frac{1}{\rho} \; if \; \|\mathbf{w}\| \leq r, \text{ 0 otherwise}$$

The likelihood $\mathbf{P}(y^i \mid \mathbf{w}, \mathbf{X_i})$ remains same as we use for the unconstrained version.

$$\textbf{PDF } \mathbf{P}(y^i \mid \mathbf{w}, \mathbf{X_i}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^i - \langle \mathbf{w}, \mathbf{X^i} \rangle)^2}{2\sigma^2}}$$

$$\text{and } \mathbf{P}(\mathbf{y} \mid \mathbf{w}, \mathbf{X}) = \prod_{i=1}^{n} \mathbf{P}(\mathbf{y^i} \mid \mathbf{w}, \mathbf{X^i})$$

This works because when $\|\mathbf{w}\| \leq r$ the MAP estimation reduces to

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)^2$$

Otherwise the cost function shoots up as $-log(\mathbf{P}(\mathbf{w})) = -log(0)$ goes to infinity.
There for the MAP estimation becomes

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)^2 \text{ where } \|\mathbf{w}\| \leq r$$

Assignment Number: 1
Student Name: Raktim Mitra
Roll Number: 150562
Date: September 10, 2017

Likelihood distribution:

$$\mathbf{P}(\mathbf{y^i} \mid \mathbf{w}, \mathbf{X^i}) = \frac{1}{\sqrt{2\pi}\rho} e^{-\frac{(y^i - \langle \mathbf{w}, \mathbf{X^i} \rangle)^2}{2\rho^2}}$$

$$\text{and } \mathbf{P}(\mathbf{y} \mid \mathbf{w}, \mathbf{X}) = \prod_{i=1}^{n} \mathbf{P}(\mathbf{y^i} \mid \mathbf{w}, \mathbf{X^i})$$

Prior Distribution:

$$\mathbf{P(w)} = \prod_{i=1}^{d} \mathbf{P(w_i)}$$

$$\text{where, } \mathbf{P(w_i)} = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{\mathbf{w_i^2}}{2{\sigma_i}^2}}$$

Explanation:

$$-log(\mathbf{P}(\mathbf{y} \mid \mathbf{w}, \mathbf{X})) = \sum_{i=1}^{n} -log(\mathbf{P}(\mathbf{y^i} \mid \mathbf{w}, \mathbf{X^i}))$$

$$= \sum_{i=1}^{n} \frac{(y^i - \langle \mathbf{w}, \mathbf{X^i} \rangle)^2}{2\rho^2} \text{ (ignoring log of constants)}$$

$$-log(\mathbf{P(w)}) = \sum_{i=1}^{d} -log(\mathbf{P(w_i)})$$

$$= \sum_{i=1}^{d} \frac{\mathbf{w_i}^2}{2{\sigma_i}^2} \text{ (ignoring log of constants)}$$

So, multiplying both with $2\rho^2$ MAP estimation becomes:

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \sum_{i=1}^{d} \frac{\rho^2}{{\sigma_i}^2} \mathbf{w_i^2}$$

Putting $\frac{\rho^2}{{\sigma_i}^2} = \alpha_i$ we get the required form of the optimisation problem :

$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \sum_{i=1}^{d} \alpha_i \mathbf{w_i^2}$$

let, $\mathbf{A}$ be a d x d diagonal matrix where diagonal entries are $\alpha_i's$ and $\mathbf{w}$ is d x 1 then $\sum_{i=1}^{d} \alpha_i \mathbf{w_i^2}$ can be written as $\mathbf{w^T A w}$. $\mathbf{X}$ is n x d, i.e. each row is a training example, then $\sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$ can be written as $\langle \mathbf{y} - \mathbf{Xw}, \mathbf{y} - \mathbf{Xw} \rangle$ which is $(\mathbf{y} - \mathbf{Xw})^{\mathbf{T}}(\mathbf{y} - \mathbf{Xw})$. Therefore the optimisation function becomes:

$$\mathbf{y^T y} + \mathbf{w^t X^T X w} - \mathbf{w^T X^T y} - \mathbf{y^T X w} + \mathbf{w^T A w}$$
$$= \mathbf{y^T y} + \mathbf{w^t X^T X w} - \mathbf{2 w^T X^T y} + \mathbf{w^T A w} \text{ (since } w^T X^T y = y^T X w = \langle \mathbf{Xw}, \mathbf{y} \rangle)$$

to find minimum set, derivative w.r.t $\mathbf{w}$ to 0:

$$\mathbf{0} + \mathbf{2X^T X w} - \mathbf{2X^T y} + \mathbf{2Aw} = 0$$
$$\implies (\mathbf{X^T X} + \mathbf{A})\mathbf{w} = \mathbf{X^T y}$$
$$\implies \widehat{\mathbf{w}} = (\mathbf{X^T X} + \mathbf{A})^{\mathbf{-1}} \mathbf{X^T y} \text{ [Answer]}$$

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

Assignment Number: 1
Student Name: Raktim Mitra
Roll Number: 150562
Date: September 10, 2017

QUESTION 4

$$\left\{\widehat{\mathbf{W}}, \{\hat{\xi}_i\}\right\} = \underset{\mathbf{W},\{\xi_i\}}{\arg\min} \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i$$
$$\text{s.t. } \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i, \forall i, \forall k \neq y^i \qquad (P1)$$
$$\xi_i \geq 0, \text{ for all } i$$

Looking at the constraints, we get $\xi_i \geq max(0, \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle + 1)$ for all i and for all k $\neq y^i$. Which is equevalent to $\xi_i \geq max(0, 1 + \max_{k\neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i)$ for all i. In notations given in problem i.e. $\xi_i \geq [1 + \max_{k\neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i]_+$ for all i.

Proof of equevalence :

P2 $\implies$ P1:

Let, $\widehat{\mathbf{W}}$ is a soution of **P2**. $\hat{\xi}_i = [1 + \max_{k\neq y} \hat{\boldsymbol{\eta}}_k^i - \hat{\boldsymbol{\eta}}_y^i]_+$ for all i,where $\hat{\boldsymbol{\eta}}^i = \left\langle \widehat{\mathbf{W}}, \mathbf{x}^i \right\rangle$ gives optimal solution of **P1**,because once **P1** reaches $\widehat{\mathbf{W}}$, there is no way $\xi_i s$ can be reduced from $\hat{\xi}_i s$. Therefore, there exists $\{\hat{\xi}_i\}$ such that $\widehat{\mathbf{W}}$,$\{\hat{\xi}_i\}$ is a solution of **P1**.

P1 $\implies$ P2:

Let, $\widehat{\mathbf{W}}$,$\{\hat{\xi}_i\}$ be an optimal solution of **P1**. That means, $\xi_i \geq max(0, \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle - \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle + 1)$ for all i and for all k, which is equevalent to $\hat{\xi}_i \geq [1 + \max_{k\neq y} \hat{\boldsymbol{\eta}}_k^i - \hat{\boldsymbol{\eta}}_y^i]_+$ for all i.

Claim: $\hat{\xi}_i = [1 + \max_{k\neq y} \hat{\boldsymbol{\eta}}_k^i - \hat{\boldsymbol{\eta}}_y^i]_+$ for all i.

Proof: Otherwise $\hat{\xi}_i > [1 + \max_{k\neq y} \hat{\boldsymbol{\eta}}_k^i - \hat{\boldsymbol{\eta}}_y^i]_+$ for all i. But in that case we can choose a smaller ξ_i which will minimise the objective function further while abiding by the connstraints. This contradicts the fact that $\widehat{\mathbf{W}}$,$\{\hat{\xi}_i\}$ is optimum.

Therefore, by replacing $\xi_i s$ with $[1 + \max_{k\neq y} \boldsymbol{\eta}_k^i - \boldsymbol{\eta}_y^i]_+$ for all i i.e. $\ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)$, we get that **P1** is restructured to **P2** while $\widehat{\mathbf{W}}$ remains optimal.

This proves $\widehat{\mathbf{W}}$ is an optimal solution of Problem **P2**.

Hence, P1 and P2 are equevalent optimisation problems. [Proved]

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

QUESTION 5

Assignment Number: 1
Student Name: Raktim Mitra
Roll Number: 150562
Date: September 10, 2017

To prove that $\mathbf{g}$ is a subdifferential of f at $\mathbf{w}$, we show that for every $\mathbf{z} \in \mathbb{R}^d$:

$$f(\mathbf{z}) \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{w} \rangle$$

We can write $\mathbf{g}$ as $\mathbf{g} = \sum\limits_{i \in S} -y^i \mathbf{X}^i$ where S is the set of $i's$ such that $y^i \langle \mathbf{w}, \mathbf{X} \rangle < 1$

$$\begin{aligned}
\mathbf{w}^T \mathbf{g} &= \sum_{i \in S} -\mathbf{w}^T y^i \mathbf{X}^i \\
\implies \mathbf{g}^T \mathbf{w} &= \sum_{i \in S} -y^i \mathbf{w}^T \mathbf{X}^i \text{ (since } \mathbf{w^T g} = \mathbf{g^T w} = \langle \mathbf{w}, \mathbf{g} \rangle) \\
\implies f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{w} \rangle &= f(\mathbf{w}) + \mathbf{g}^T \mathbf{z} - \mathbf{g}^T \mathbf{w} \\
&= \sum_{i \in S}^{n} [1 - y^i \langle \mathbf{w}, \mathbf{X}^i \rangle] + \sum_{i \in S} -y^i \mathbf{z}^T \mathbf{X}^i + \sum_{i \in S} y^i \mathbf{w}^T \mathbf{X}^i \\
&= \sum_{i \in S} 1 - y^i \langle \mathbf{z}, \mathbf{X}^i \rangle
\end{aligned}$$

To complete the proof we need to consider two cases:

1. $y^i \langle \mathbf{z}, \mathbf{X}^i \rangle < 1$ Then the term's contribution in RHS is greater than 0. By definition,it is a part of f(z).
2. $y^i \langle \mathbf{z}, \mathbf{X}^i \rangle \geq 1$ Then contribution is negative. This term reduces the RHS sum only.Hence does not contribute to f(z).

$\therefore$ f(z) is sum of all positive quantities in RHS of the expression and some other positive terms. Hence:

$$f(\mathbf{z}) \geq \sum_{i \in S} 1 - y^i \langle \mathbf{z}, \mathbf{X}^i \rangle$$

$$\implies f(\mathbf{z}) \geq f(\mathbf{w}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{w} \rangle$$

[Proved]

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

Assignment Number: 1
Student Name: Raktim Mitra
Roll Number: 150562
Date: September 10, 2017

QUESTION

6

Part 1:

Test Error for 5 different k values:

k = 1	24.075%
k = 2	21.44%
k = 3	19.19%
k = 5	17.71%
k = 10	16.795%

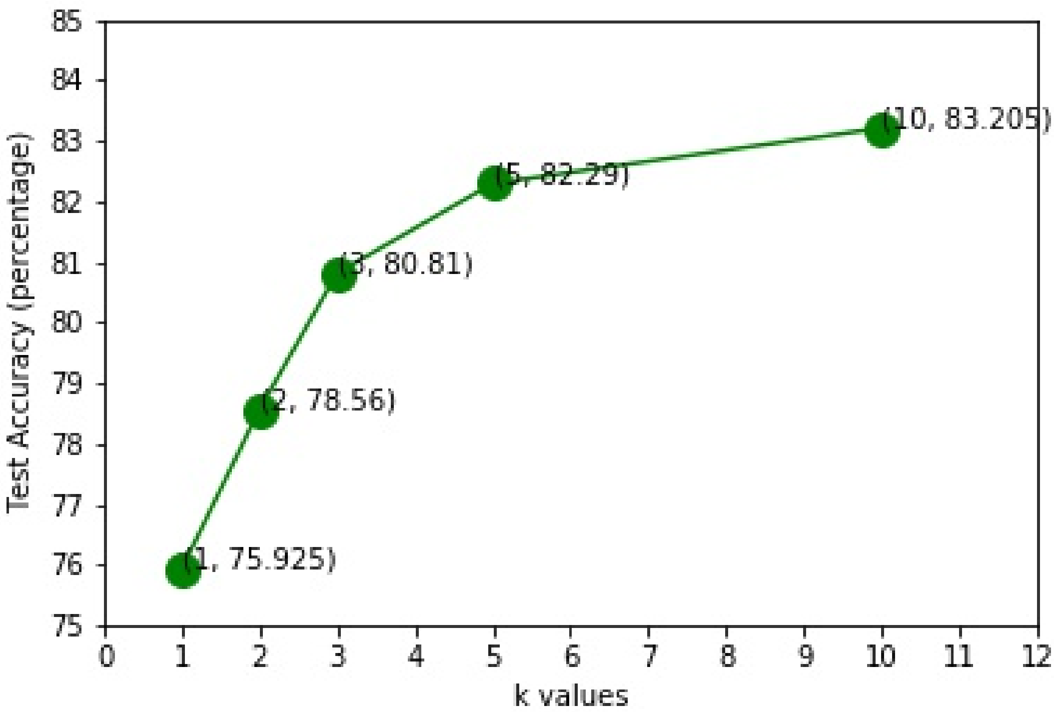

Figure 2: Plot showing test accuracy for k values {1,2,3,5,10}

Observation: As expected, number of neighbour points considered increases as we go from 1 to 10. The decision for each point takes into account more number of points, leading to minimising chanes of error.

Part 2:

I devided the train data in 8:2 ratio(train and validation) and calculated the accuracy for k = 10 to 100 with steps of 5. I found high values around 15,20. Then I reran the validation with k = 15 to 24 and found highest value at 21.

Therefore tuned value of k is 21.

Part 3:

Learned suitable metric with lmnn, used 6000 data points and 3000 iterations. Although it did not converge but as the plot shows the objective was changing very slowly after around 1000 iterations. Therefore we can say the learned metric is good enough.

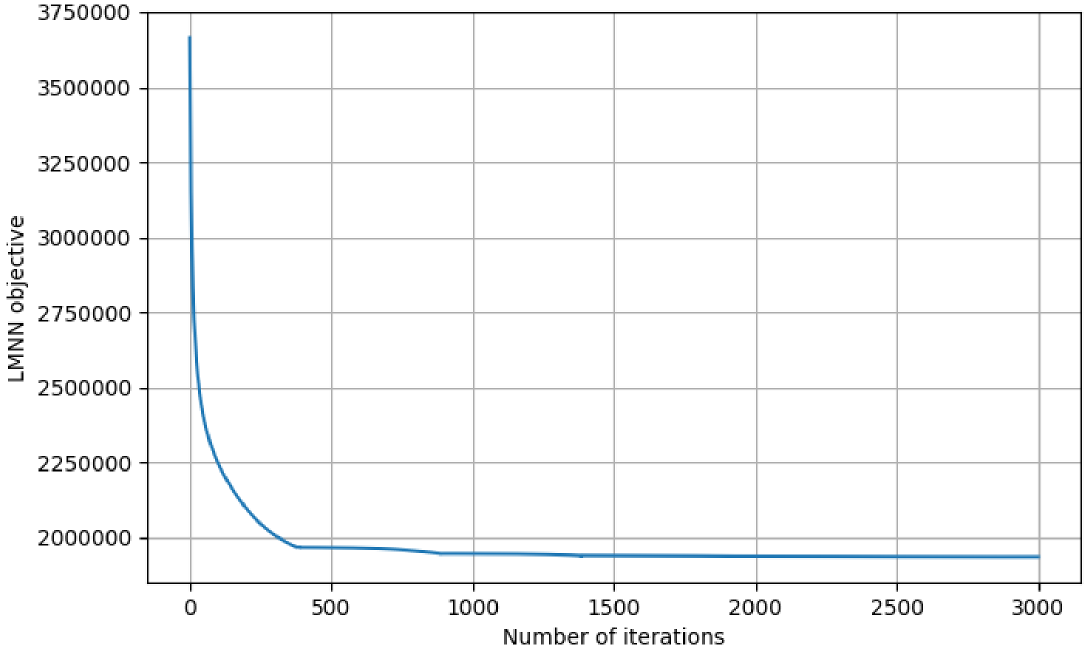

Figure 3: Plot showing lmnn objective w.r.t. number of iterations for k = 21

Using the learnt Metric and K = 21 the test accuracy is **83.975%**