COL 341 Minor 2

Chinmay Mittal

TOTAL POINTS

58 / 65

QUESTION 1

1 SVM **13 / 20**

+ 0 pts Not Attempted/ Incorrect

✓ **+ 8 pts** *Correct Primal and Lagrange Formation*

- 2 pts Condition of alpha and beta > 0 missing in conditions of Lagrange

+ 6 pts Correct differentiation equations from Lagrange

+ 3 pts Correctly substituting and deriving the final form of Lagrange

✓ **+ 3 pts** *Correctly showing constraints on dual*

+ 0 pts Minor mistakes leading to point adjustment

+ 4 pts Partially correct Primal and Lagrange

+ 4 pts Partial : 2 out of 3 derivatives are correct

+ 2 *Point adjustment*

1 This needs to be derived by substituting with derivatives

QUESTION 2

2 Euclidean Distance **10 / 10**

✓ **+ 10 pts** *Correct*

+ 8 pts Minor bugs (may be minor incompleteness)

+ 5 pts Major bugs or incompleteness with partial correctness

+ 0 pts Incorrect/ Not Attempted

QUESTION 3

Decision Tree 10 pts

3.1 (a) **5 / 5**

+ 3 pts Partial Correct

✓ **+ 5 pts** *Correct*

+ 0 pts Incorrect/NotAttempted

- 1 pts No Explanation/Justification

3.2 (b) **5 / 5**

✓ **+ 5 pts** *Correct*

+ 0 pts Incorrect/Not Attempted

QUESTION 4

AdaBoost 25 pts

4.1 (a) **2 / 2**

4a

✓ **- 0 pts** *Correct*

- 2 pts Incorrect

- 2 pts Unattempted

4.2 (b) **3 / 3**

4b

✓ **- 0 pts** *Correct*

- 3 pts Incorrect/Unattempted.

- 1 pts Wrong Exponent

4.3 (c) **10 / 10**

✓ **+ 10 pts** *Click here to replace this description.*

+ 0 pts Incorrect/ not attempted

- 2 pts steps are missing

- 1 pts did not explain the normalization term

- 1 pts did not explain the 1/m term

+ 3 pts initial steps written

4.4 (d) **10 / 10**

+ 0 pts Incorrect

+ 0 pts Not attempted

✓ **+ 10 pts** *Correct*

Student Name: CHINMAY MITTAL	Entry Number: 2020CS10336

Department of Computer Science and Engineering
Indian Institute of Technology Delhi

COL341: Fundamentals of Machine Learning

Minor 2

Time: 60 minutes | Maximum Marks: 65 | Number of Questions: 4

Instructions: Please attempt all questions. If you feel any question/statement is ambiguous, please write your assumptions clearly, and then answer as per your assumptions.

Question	1	2	3(a)	3(b)	4(a)	4(b)	4(c)	4(d)	Total
Max Marks	20	10	5	5	2	3	10	10	65
Earned Marks									

1. [20 marks] Recall the soft-margin SVM covered in the class, where the primal can be written as follows:

$$\min_{b,w,\xi} \frac{1}{2} w^\top w + C \sum_{n=1}^{m} \xi_n$$
$$\text{subject to: } y_n(w^\top x_n + b) \geq 1 - \xi_n$$
$$\xi_n \geq 0, \text{ for } n = 1, \ldots, m.$$

We derived the corresponding dual in one of the homework assignments as:

$$\min_{\alpha} \frac{1}{2} \alpha^\top G \alpha - \sum_{n=1}^{m} \alpha_n$$
$$\text{subject to: } y^\top \alpha = 0$$
$$0 \leq \alpha_n \leq C, \text{ for } n = 1, \ldots, m,$$

for an appropriate matrix G. One can look at the above as including the penalty term $C\|\xi\|_p$, with $p = 1$ in the primal, which led to the constraint $\|\alpha\|_q \leq C$, with $q = \infty$ in the dual. One can prove in general that $\frac{1}{p} + \frac{1}{q} = 1$. Your task in this question is to derive this for $p = \infty$, and $q = 1$.

The primal optimization problem becomes $\min_{b,w,\xi} \frac{1}{2} w^T w + C \|\xi\|_\infty$

s.t. $y_n(w^T x_n + b) \geq 1 - \xi_n$

$\xi_n \geq 0$

$\|\xi\|_\infty = \max_i \xi_i$

We can form an equivalent optimization problem as follows.

$$\min_{b,\omega,\xi} \frac{1}{2}\omega^T\omega + C\xi^*$$

subject to $\xi^* \geq \xi_n \ \forall n$

$y_n(\omega^T x_n + b) \geq 1 - \xi_n \ \forall n$

$\xi_n \geq 0 \ \forall n$

This encodes the constraint $\max_n \xi_n = \xi^*$

a_n, β_n, c_n for 3 types of constraints

The Lagrangian of this equivalent primal optimization is as follows

$$L(b, \omega, \xi, a_n, \beta_n, c_n) = \frac{1}{2}\omega^T\omega + C\xi^* + \sum a_n(1 - \xi_n - y_n(\omega^T x_n + b)) + \sum \beta_n(-\xi_n) + \sum c_n(\xi_n - \xi^*)$$

s.t. $a_n \geq 0, \beta_n \geq 0, c_n \geq 0$

$a_n \geq 0, \beta_n \geq 0, c_n \geq 0$

from KKT conditions

$$\frac{\partial L}{\partial \xi_n} = 0 \Rightarrow -a_n - \beta_n + c_n = 0 \ \forall n$$

$$a_n + \beta_n = c_n$$

$$\frac{\partial L}{\partial \xi^*} = 0 \Rightarrow C - \sum c_n = 0 \Rightarrow \sum c_n = C$$

Substituting this in the Lagrangian we get (β_n and c_n in terms of a_n)

$$L(b, \omega, \xi, a_n) = \underbrace{\frac{1}{2}\omega^T\omega - \sum a_n y_n(\omega^T x_n + b)}_{\text{lagrangian is same as Hard margin SVM}} + \cancel{\xi^*(C - \sum c_n)} + \cancel{\sum \xi_n(c_n - a_n - \beta_n)}$$

0, 0

lagrangian is same as Hard margin SVM subject to different constraints.

The form of the optimization problem comes out exactly the same $\min_{a} \frac{1}{2}a^T G a - \sum_{n=1}^{m} a_n$ subject to $y^T a = 0$

The new constraints are $C = \sum c_n \Rightarrow C = \sum a_n + \sum \beta_n$

$\Rightarrow \lVert a \rVert_1 \leq C$ $\Rightarrow$ When $q = \infty$ $p = 1$	since $\beta_n \geq 0$ we get $0 \leq \sum a_n \leq C$ $a_n \geq 0$

COL341: Minor 2	Student Name:	Entry Number:
	CHINMAY MITTAL	2020CS10336

2. [10 marks] Kernels are general idea which can be used to work with infinite dimensional feature space in variety of ML techniques. We saw it extensively in SVMs, and also mentioned that it could be used for k-NN classifier as well. However, to make it work in k-NN we need to compute the distances in the D-dimensional feature space, where D can be ∞ as well. Suppose we have a kernel $K(\cdot,\cdot)$, such that there is an implicit high dimensional feature map $\phi : \mathbb{R}^d \to \mathbb{R}^D$ that satisfies $\forall x, z \in \mathbb{R}^d, K(x,z) = \phi(x) \cdot \phi(z)$, where $\phi(x) \cdot \phi(z) = \sum_{i=1}^{D} \phi(x)_i \phi(z)_i$ is the dot product in the D-dimensional space. Show, how to calculate the Euclidean distance in the D-dimensional space:

$$\|\phi(x) - \phi(z)\| = \sqrt{\sum_{i=1}^{D} (\phi(x)_i - \phi(z)_i)^2},$$

without explicitly calculating the values in the D-dimensional vectors.

$$\|\phi(x) - \phi(z)\| = \sqrt{\sum_{i=1}^{D}\left(\phi(x)_i^2 + \phi(z)_i^2 - 2\phi(x)_i\phi(z)_i\right)}$$

$$= \sqrt{\sum_{i=1}^{D}\phi(x)_i \cdot \phi(x)_i + \sum_{i=1}^{D}\phi(z)_i\phi(z)_i - 2\sum_{i=1}^{D}\phi(x)_i\phi(z)_i}$$

$$= \sqrt{K(x,x) + K(z,z) - 2K(x,z)}$$

COL341: Minor 2	Student Name: CHINMAY MITTAL	Entry Number: 2020CS10336

3. Consider the following dataset with $A, B, C \in \{T, F\}$ as the input features, and $Y \in \{T, F\}$ as the output label:

A	B	C	Y
F	F	F	F
F	F	T	T
T	T	F	T
T	T	T	F

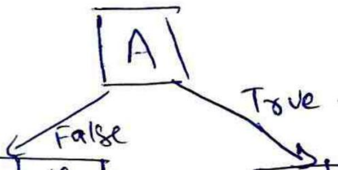
(a) [5 marks] Using the dataset above, we want to build a decision tree which classifies Y as T/F Draw the tree that would be learned by the greedy algorithm with zero training error. You do not need to show any computation but should describe the justification of each decision in English.

(b) [5 marks] Is this tree optimal (i.e. does it get zero training error with minimal depth)? Explain in less than two sentences. If it is not optimal, draw the optimal tree as well.

At the first node, the choices are. A^- [F T] B^- [F T] C^- [F T]

A^+ [T F] B^+ [T F] C^+ [T F]

All nodes have same information gain $\Rightarrow$ choose any node, I choose A

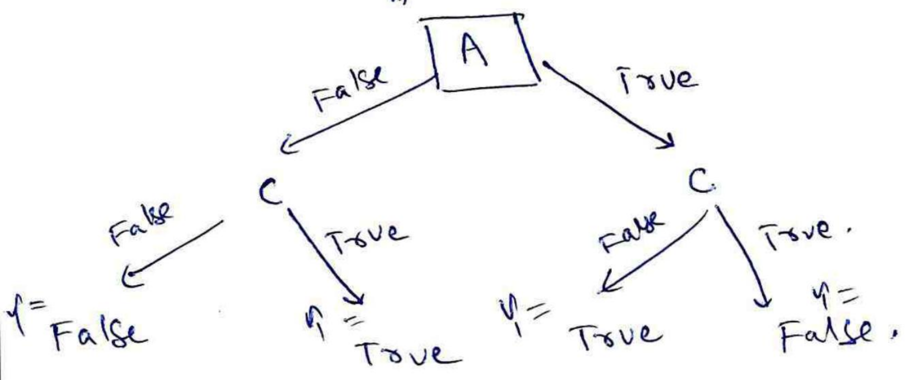


B	C	Y
F	F	F
F	T	T

B	C	Y
T	F	T
T	T	F

→ For both the nodes all data items have same value of B $\Rightarrow$ split has to be wrt C at each node

$\Downarrow$



Attributes A and B are symmetric (all data points have same value)

$\Rightarrow$ choosing B first will not give a better tree

if we choose C first, then at the second stage we can choose any amongst attribute A/B to perfectly classify all nodes $\Rightarrow$ depth of tree $= 2$

Hence the tree we created is optimal (no better tree exists)

$\hookrightarrow$ Hence amongst all possibilities no tree is better than ours $\Rightarrow$ our tree is optimal.

$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$

COL341: Minor 2	Student Name: CHINMAY MITTAL	Entry Number: 2020CS10336

4. We discussed AdaBoost algorithm in the class to learn a classifier $f(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$. In each time step t we compute the error, $\text{err}_t = \frac{\sum_{i=1}^{m} w_t^i \mathbb{I}(h_t(x^i) \neq y^i)}{\sum_{i=1}^{m} w_t^i}$, where m denotes the number of samples. We use w to estimate $\alpha_t = \frac{1}{2} \log\left(\frac{1-\text{err}_t}{\text{err}_t}\right)$. We then reweigh the samples according to the expression: $w_{t+1}^i = w_t^i \exp(-\alpha_t y^i h_t(x^i))$.
   (a) [2 marks] Show that $y^i h_t(x^i) = \mathbb{I}(h_t(x^i) \neq y^i)$
   (b) [3 marks] Show that $\exp(-\alpha_t y^i h_t(x^i)) \propto \exp(2\alpha_t \mathbb{I}(h_t(x^i) \neq y^i))$
   (c) [10 marks] Assuming $Z_t = \sum_{i=1}^{m} w_t^i$, and we initialize weight of each sample as uniform in $t = 1$, show that $w_{t+1}^i = \frac{\exp(-y^i \sum_{s=1}^{t} \alpha_s h_s(x^i))}{m \prod_{s=1}^{t} Z_s}$.
   (d) [10 marks] Assuming that the AdaBoost minimizes the loss $\mathcal{L}(f) = \frac{1}{m} \sum_{i=1}^{m} \exp(-y^i f(x^i))$, show that the loss at time step t, $\mathcal{L}(f_t) = \prod_{s=1}^{t} Z_s$.

(a) $\frac{1}{2}(1 - h_t(x^i)\, y^i)$ When the two are not equal $h_t(x^i) \neq y^i$

$\Rightarrow h_t(x^i) = 1,\ y^i = -1$ or

$h_t(x^i) = -1\ \ y^i = 1 \Rightarrow h_t(x^i) y^i = -1$

$\Rightarrow \frac{1}{2}(1 - h_t(x^i) y^i) = \frac{1}{2}(1-(-1)) = 1 = \mathbb{I}(h_t(x^i) \neq y^i)$ When the two are not equal.

When the two are equal

$h_t(x^i) = y^i = 1$ or $h_t(x^i) = y^i = -1$

$\Rightarrow \frac{1}{2}(1 - h_t(x^i) y^i) = \frac{1}{2}(1 - 1) = 0 = \mathbb{I}(h_t(x^i) \neq y^i)$

(b) $\exp\left(\frac{1}{2}(1 - h_t(x^i) y^i)\right) = \exp\left(\mathbb{I}(h_t(x^i) \neq y^i)\right)$

$\Rightarrow \exp(1 - h_t(x^i) y^i) = \exp\left(2\,\mathbb{I}(h_t(x^i) \neq y^i)\right)$

$\Rightarrow \exp(-h_t(x^i) y^i) = \exp(-1) \cdot \exp\left(2\,\mathbb{I}(h_t(x^i) \neq y^i)\right)$

$\Rightarrow \exp(-\alpha_t h_t(x^i) y^i) = \exp(-\alpha_t) \cdot \exp\left(2\alpha_t\, \mathbb{I}(h_t(x^i) \neq y_i)\right)$

$\propto \exp\left(2\alpha_t\, \mathbb{I}(h_t(x^i) \neq y_i)\right)$

initial weight : $1/m$

In my proof I explicitly normalise weights at each step.

(c) $\cancel{w_2^i} = \cancel{w_1^i} \exp(-\alpha_1 y^i h_1(x^i)) / Z_1$

normalization

$\cancel{w_3^i} = \cancel{w_2^i} \exp(-\alpha_2 y^i h_2(x^i)) / Z_2$

$\vdots$

$\cancel{w_{T+1}^i} = \cancel{w_T^i} \exp(-\alpha_T y^i h_T(x^i)) / Z_T$

Multiplying all equations

$$w_{t+1}^i = \frac{1}{m} \frac{\exp\left(-y^i \left(\sum_{t=1}^{T} \alpha_t h_t(x^i)\right)\right)}{\prod_{s=1}^{T} Z_s}$$

We had normalized in the computation

(d) $$w_{t+1}^i = \frac{1}{m} \frac{\exp(-y^i f_t(x^i))}{\prod_{s=1}^{T} Z_s}$$ Since $f_t(x^i) = \sum_{t=1}^{T} \alpha_t h_t(x^i)$

$$\sum_{i=1}^{m} w_{t+1}^i = \frac{1}{m} \frac{1}{\prod_{s=1}^{T} Z_s} \sum_{i=1}^{m} \exp(-y^i f_t(x^i))$$

$= 1$

Since weights are normalized

$$\Rightarrow 1 = \frac{1}{\prod_{s=1}^{T} Z_s} \frac{1}{m} \sum_{i=1}^{m} \exp(-y^i f_t(x^i))$$

$$\Rightarrow \prod_{s=1}^{T} Z_s = L(f_t)$$

ROUGH

COL341: Minor 2	Student Name: ~~2020CS10336~~ CHINMAY MITTAL	Entry Number: 2020CS10336.

$1 \cdot 0 = 0$

$\overline{1} - 1$

-1

$$\sqrt{\sum_{i=1}^{D} \left(\phi(x)_i - \phi(z)_i\right)^2}$$

$$= \sqrt{\sum_{i=D} \left(\phi_i^2(x) + (\phi(z)_i)^2 - 2\,\phi(x)_i\,\phi(z)_i\right)}$$

$$= \sqrt{\sum_{i=D} \phi(x)_i \cdot \phi(x)_i + \phi(z)}$$

$$= \sqrt{K(x,x) + K(z,z) - 2K(x,z)}$$

$e^{\frac{1}{2}(1 - h(x^i)(y^i))}$

$= e^{\frac{\mathrm{II}}{2\,4}}$

$e^{1}\, e^{-h_t(x^i)y^i)} = e$

$\frac{1}{m}\frac{\exp(y^i(}{Z_1}$

$w_{t+1}^i = w_t^i \exp(-\alpha_t\, y^i h_t(x^i)) \,/\, Z_t$

$w_t \qquad w_2^i = w_1^i \exp(-\alpha_1\, y^i h_1(x^i)) \,/\, Z_1$

$w_3^i = w_2^i \exp(-\alpha_2\, y^i h_2(x^i)) \,/\, Z_2$

(d) $w_{t+1}^{i} = \frac{\exp(-y^i f_t(x^i))}{m \prod_{s=1}^{t} z_s}.$

$\sum w_{t+1}^{i} = \frac{L(f_t)}{\prod^{t} z_s}$

$\min_{b,w,\xi} \frac{1}{2} w^T w + C\xi$

$\xi \geq \xi_i \geq 0$

$L \rightarrow \frac{1}{2} w^T w + C\xi + \sum \alpha_n (1 - \xi_n - y_n (w^T x_n + b)) + \sum \beta_n (-\xi_n) + \sum c_n (\xi_n - \xi)$

$\underline{C - \alpha_n - \beta_n = 0}$

$C = \sum c_n$

$-\alpha_n - \beta_n + c_n = 0$

$L \rightarrow \frac{1}{2} w^T w - \sum \alpha_n y_n (w^T x_n + b))$ $\boxed{c_n = \alpha_n + \beta_n}$

$\alpha_n \geq 0$

$\beta_n \geq 0$

$c_n \geq 0$

$\sum \alpha_n + \sum \beta_n = C$