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

Assignment Number: 2
Student Name: Gurpreet Singh
Roll Number: 150259
Date: October 9, 2017

QUESTION 1

Part 1

No, name is not a useful attribute in learning a decision tree as splitting on this attribute will have no gain. Non-statistically speaking, there is a lot of variance in the attribute itself, and it will be very inefficient to decide on the basis of name as it can often be a newly observed value.

Part 2

It is not possible to classify the data given perfectly. Looking at data #4 and #6, they have the same values for all fields except for 'name', however have different target values. Since we shall not use 'name' as a decision classifier (as discussed in Part A), therefore we can say that is is not possible to classify the data perfectly.

Part 3

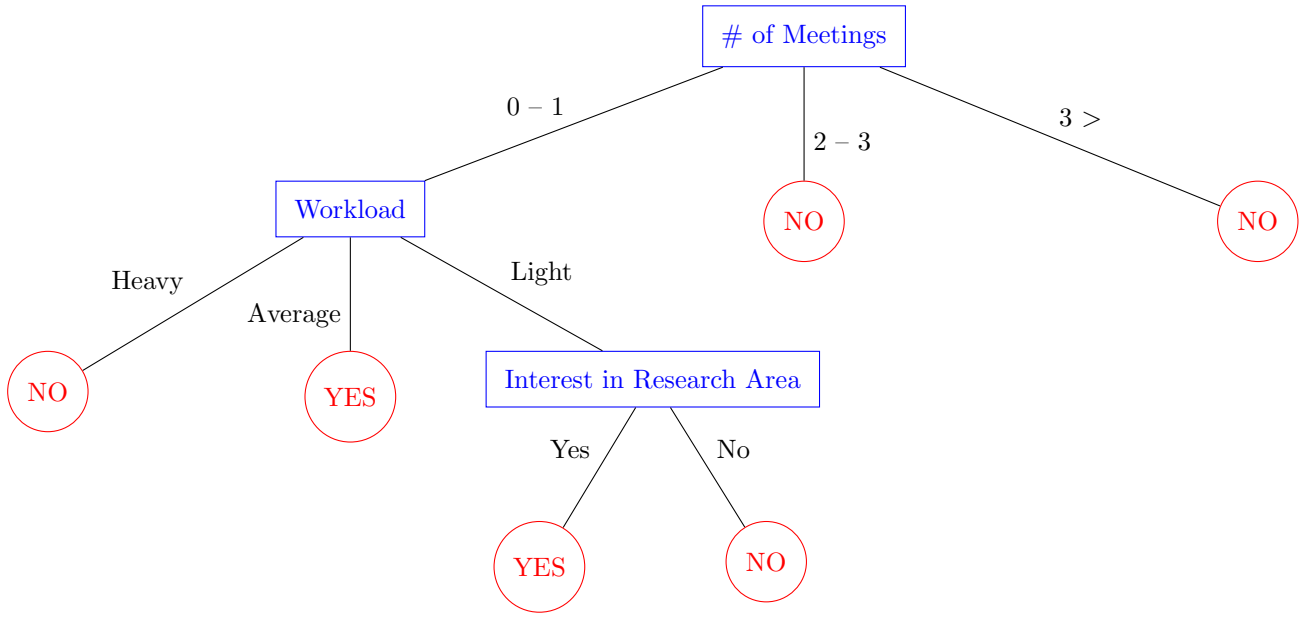

Figure 1: ID3 Tree for the given data

ID3 is a greedy strategy to generate decision trees using Information Gain. The information gain at each level/depth is mentioned below

Level 1

Entropy $S = 0.9183$

IG: Size of Research Group $= S - 0.2600 - 0.2667 - 0.3237 = 0.0679$
IG: Interest in Research Area $= S - 0.2667 - 0.6199 = 0.0317$
IG: Workload $= S - 0.4000 - 0.2163 - 0.2406 = 0.0614$
IG: Number of Meetings $= S - 0.6667 = 0.2516$

Clearly the attribute *Number of Meetings* provides the most IG, and hence we use this attribute to split at the first level.

Level 2 — Number of Meetings

Possible distictions — '$0-1$', '$2-3$', '>3'

Since for '$2-3$' and '>3', all labels are 'No' (**i.e.** entropy is 0), hence we do need to split here, and we add leaf nodes.

For the samples remaining with value '$0-1$' of the attribute *Number of Meetings*, we will further extend the decision tree.

Extropy $S = 1$

IG: Size of Research Group $= S - 0.0000 - 0.4000 - 0.4855 = 0.1145$
IG: Interest in Research Area $= S - 0.2755 - 0.6897 = 0.0348$
IG: Workload $= S - 0.0000 - 0.0000 - 0.3610 = 0.6390$

Clearly the attribute *Workload* gives the highest gain, and hence we use this attribute to further construct the decision tree.

Level 3 — Workload

Possible distinctions — 'Light', 'Average', 'Heavy'

Since there is purity in the samples with 'Average' and 'Heavy' value in the *Workload* attribute (*i.e.* entropy is 0), we do not split further in this case. Therefore we are only left with two samples, which have the value 'Light'.

Entropy $S = 1$

IG: Size of Research Group $= S - 1.0000 = 0.0000$
IG: Interest in Research Area $= S - 0.0000 = 1.0000$

Hence we further split using *Interest in Research Area* attribute. Since there is only one sample per distinction, we cannot split further. Hence we have the decision tree showed in Figure 1

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

Assignment Number: 2
Student Name: Gurpreet Singh
Roll Number: 150259
Date: October 9, 2017

QUESTION 2

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

Assignment Number: 2
Student Name: Gurpreet Singh
Roll Number: 150259
Date: October 9, 2017

QUESTION 3

Part 1

We have the optimization problem $(P1)$ as follows

$$\min_{\mathbf{w},\{\xi_i\}} \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 \qquad (P1)$$
$$s.t. \forall i \in [n] \quad 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq \xi_i$$
$$\xi_i \geq 0$$

Now consider an optimization problem $(P2)$, which is similar to the optimization problem given in equation **??**, except that we do not have the condition $\xi_i \geq 0$ for any $i \in [n]$. Clearly, the convexity still holds, and hence we will have a solution for $(P2)$.

Consider this solution to be $\mathbf{w}_0, \{\xi_i^0\}$. If we do not have any $k \in [n]$ such that $\xi_k^0 < 0$. Then we can safely claim that the solution for both the optimization problems $(P1)$ and $(P2)$ is the same. Otherwise, we will have at least one $k \in [n]$ such that the condition holds, *i.e.* $\xi_k^0 < 0$. Since this is a solution of the $(P2)$, we can say that this satisfies the condition $1 - y^k \left\langle \mathbf{w}, \mathbf{x}^k \right\rangle \leq \xi_k^0$.
Now consider the pair $\mathbf{w}, \{\xi_i^1\}$ where

$$\xi_i^1 = \begin{cases} \xi_i^0 & if i \neq k \\ 0 & if i = k \end{cases}$$

Since $0 > \xi_k^0$, we can say that it follows all constraints of $(P2)$ and $0 < (\xi_k^0)^2$. For all other ξ_i^1, they are the same and hence do not change anything (since all ξ_i are independent). Hence, we can say that $w, \{\xi_i^1\}$ is a solution of $(P2)$, and $val(P2, (w, \{\xi_i^1\})) < val(P2, (w, \{\xi_i^1\}))$.

However, since $w, \{\xi_i^0\}$ was claimed to be a valid solution, this is not possible. Hence the claim that can exist a $k \in [n]$ such that $\xi_k^0 < 0$ is false. Hence, the solution of $(P2)$ is always the same as that of $(P1)$, which suggests that the second constraint is vacuous.

Part 2

Since we have already proved that the constraint $\xi_i \geq 0$ is vacuous, we can alter the optimization problem as follows

$$\min_{\mathbf{w},\{\xi_i\}} \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 \qquad (P1)$$
$$s.t. \forall i \in [n] \quad 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq \xi_i$$

We can convert this problem to an unconstrained optimization problem using Langrange Multipliers.

$$\min_{\mathbf{w},\{\xi_i\}} \max_{\boldsymbol{\lambda}} \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \boldsymbol{\lambda}_i \big(1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle - \xi_i\big) \qquad (P1)$$

Note: *Everywhere,* $\boldsymbol{\lambda} \in \mathbb{R}^n$

Part 3

We can convert the optimization problem $(P1)$ given in equation **??** to a dual problem. Assuming strong duality, we can switch the optimization variables.

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{w},\{\xi_i\}} \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \boldsymbol{\lambda}_i \big(1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle - \xi_i\big) \qquad (P1)$$

Now using this, we can first optimize based on the inner optimization (using differentiation). That is, we can partially differentiate w.r.t. to $\mathbf{w}$ and $\{\xi_i\}$ separately, in order to obtain a dual problem.

$$\begin{aligned} \frac{\partial P1}{\partial \mathbf{w}} = &\quad 2\mathbf{w} - \sum_{i=1}^{n} \boldsymbol{\lambda}_i y^i \mathbf{x}^i = 0 \\ \implies &\quad \mathbf{w}' = \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i y^i \mathbf{x}^i \end{aligned}$$

$$\begin{aligned} \frac{\partial P1}{\partial \xi_k} = &\quad 2\xi_k - _k = 0 \\ \implies &\quad \xi_k' = \frac{\boldsymbol{\lambda}_k}{2} \end{aligned}$$

Hence, we can replace these values in the optimization problem, while solving the inner optimization problem.

$$\begin{aligned} &\max_{\boldsymbol{\lambda}} \quad \left\|\mathbf{w}'\right\|_2^2 + \sum_{i=1}^{n} \xi_i'^2 + \sum_{i=1}^{n} \boldsymbol{\lambda}_i \big(1 - y^i \left\langle \mathbf{w}', \mathbf{x}^i \right\rangle - \xi_i'\big) \\ =&\max_{\boldsymbol{\lambda}} \quad \left\| \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i y^i \mathbf{x}^i \right\|_2^2 + \frac{1}{4} \sum_{i=1}^{n} \boldsymbol{\lambda}_i^2 - \sum_{i=1}^{n} \boldsymbol{\lambda}_i - \sum_{i=1}^{n} \boldsymbol{\lambda}_i y^i \left\langle \frac{1}{2} \sum_{j=1}^{n} \boldsymbol{\lambda}_j y^j \mathbf{x}^j, \mathbf{x}^i \right\rangle - \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i^2 \\ =&\max_{\boldsymbol{\lambda}} \quad \left\| \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i y^i \mathbf{x}^i \right\|_2^2 - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\lambda}_i \boldsymbol{\lambda}_j y^i y^j \left\langle \mathbf{x}^j, \mathbf{x}^i \right\rangle - \frac{1}{4} \sum_{i=1}^{n} \boldsymbol{\lambda}_i^2 - \sum_{i=1}^{n} \boldsymbol{\lambda}_i \end{aligned}$$

We can expand the first term as

$$\begin{aligned} \left\| \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i y^i \mathbf{x}^i \right\|_2^2 &= \frac{1}{4} \left\| \sum_{i=1}^{n} \boldsymbol{\lambda}_i y^i \mathbf{x}^i \right\|_2^2 \\ \implies \left\| \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i y^i \mathbf{x}^i \right\|_2^2 &= \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\lambda}_i \boldsymbol{\lambda}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle \end{aligned}$$

Replacing back in the original equation

$$= \max_{\boldsymbol{\lambda}} \quad -\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_j y^i y^j \left\langle \mathbf{x}^j, \mathbf{x}^i \right\rangle - \frac{1}{4}\sum_{i=1}^{n}\boldsymbol{\lambda}_i^2 - \sum_{i=1}^{n}\boldsymbol{\lambda}_i$$

We can alter this optimization problem by multiplying with -4 and inverting the optimizing condition. Therefore, the final dual optimization problem can be represented as follows

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^n} \boldsymbol{\lambda}^T Q\boldsymbol{\lambda} + \sum_{i=1}^{n}\boldsymbol{\lambda}_i^2 - 4\boldsymbol{\lambda}_i \qquad (D1)$$

where

$$Q_{ij} = y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

Part 4

Let us state the original dual problem.

$$\min_{\boldsymbol{\alpha}\in[0,C]^n} \boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \sum_{i=1}^{n}\boldsymbol{\alpha}_i$$

The main difference that seems to be there between the two objectives is that the dual variable (each value) in the original problem is upper bounded by a constant C.

It seems that the dual variable in $(D1)$ has no upper bound, however there is still a tradeoff between the high and low values of our dual variable. Consider $\boldsymbol{\lambda} > 2$. In this case, the second term will become positive. Hence, we do not want very hight values of $\boldsymbol{\lambda}$. Looking at the primal problem of the original SVM, we can say that in our case (by similarity of terms), C = 2. Hence we are trying to keep the value of λ within the range $[0, C]^n$, however it is a loose bound.

Therefore, the only difference is that in the original SVM problem, we have a strict upper bound on the value of the dual variable, whereas in our case, the upper bound is loose, but still existant.

Part 5

No. From our approach, we have proved that $\xi_i < 0$ only increases the value of the objective function while $\xi_i = 0$ is a solution, and hence the lower bound is vacuous, however, this is not the case with the original SVM. In the original problem, we are hoping to get negative values of ξ_i which would mean that we are giving weightage to the points which are far from the margin, which is not a correct optimization problem, as it essentials ignores the large margin problem. Hence, in that case, we need to explicitly add the lower bound.

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

Assignment Number: 2
Student Name: Gurpreet Singh
Roll Number: 150259
Date: October 9, 2017

QUESTION 4