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

QUESTION
1

Assignment Number: 2
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: October 15, 2017

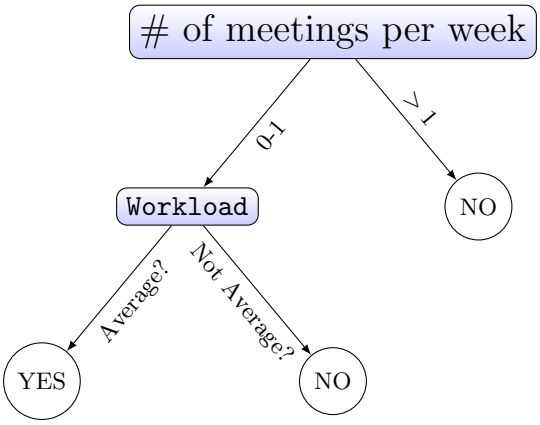
1 Name:

No.
Because whether a professor is a good advisor or not does not depend on their name.
Therefore, although there might be a correlation between some function of the name(for example whether the first letter is after 'g' or before 'g' in the alphabet) and the good/not good advising, it is only limited to the data set we have and it would not be true if we had a very large dataset.
Therefore, we should ignore this attribute while building our decision tree.

2 Perfect classification of given dataset:

No.
Look at the 4th and 6th entries of the given dataset.
Both of them have the same values for all the important attributes(i.e all except name) yet they have different values for whether the advisor is Good/Not Good.
Therefore, any classification algorithm we come up with can only be correct for either entry 4 or entry 6 but not both simultaneously(if it is a deterministic algorithm).
In this case, randomised algorithms might give correct results for both entries in some runs of the algorithm. But they are not perfect since they give wrong results also in other runs.

3 Decision Tree:



3.1 Information gain values:

At root node, information gains for
$Size = 0.067, Like = 0.031, Workload = 0.061, Meetings = 0.251 \implies Meetings$

For depth one node, where # meetings = 0-1, information gains for
$Size = 0.114, Like = 0.035, Workload = 0.439 \implies Workload$

3.2 Further splitting:

For root node's children, # meetings > 1 is not split into $2 - 3$ and > 3 because all the entries which fall into these categories are classified in the 'No' class.
For Workload's children, the 'Average' child has all values as YES(for average workload 3/3) and the 'Not average' child has majority NO values(6 out of 8). I chose to merge the light and heavy workload resulting in 6/8 cases being NO. I could have chosen to merge the light workload(1 YES, 1 NO) with either the average or the heavy workload part but both the splits give the same accuracy for the test data.

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

Assignment Number: 2
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: October 15, 2017

QUESTION 2

This is for Question 2

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

Assignment Number: 2
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: October 15, 2017

QUESTION
3

1. Vacuous constraints:

To prove: The given optimization problem has the same solutions with and without the constraint: $\xi_i \geq 0,$ for all $i \in [n]$.

Proof:
We will prove this by contradiction.
Suppose $\hat{\mathbf{w}}, \left\{\hat{\xi}_i\right\}$ are solutions to the optimization problem without the last constraint for some given inputs such that $\hat{\xi_k} < 0$ for at least one $k \in [n]$.
Consider a new proposed solution to the optimization problem $\hat{\mathbf{w}'}, \left\{\hat{\xi}'_i\right\}$ such that
$\hat{\mathbf{w}'} = \hat{\mathbf{w}},\ \hat{\xi}'_k = 0$ and $\hat{\xi}'_i = \hat{\xi}_i\ \forall\ i \in [n] \setminus \{k\}$.

It is easy to see that the new proposed solution satisfies the constraints that
$y^i \left\langle \hat{\mathbf{w}'}, \mathbf{x}^i \right\rangle \geq 1 - \hat{\xi}'_i,$ for all i.
For $i \neq k$, the constraint is satisfied since we took it $\hat{\xi}'_i = \hat{\xi}_i$ and $\hat{\xi}_i$ is a part of the solution of the initial optimisation problem.
For $i = k$, we have
$y^i \left\langle \hat{\mathbf{w}}, \mathbf{x}^i \right\rangle \geq 1 - \hat{\xi}_i\ \because$ it was part of the initial solution
$\implies y^i \left\langle \hat{\mathbf{w}}, \mathbf{x}^i \right\rangle \geq 1 - \hat{\xi}_i$
$\implies y^i \left\langle \hat{\mathbf{w}}, \mathbf{x}^i \right\rangle \geq 1\ \because\ \hat{\xi}_k < 0$
$\implies y^i \left\langle \hat{\mathbf{w}'}, \mathbf{x}^i \right\rangle \geq 1\ \because\ \hat{\mathbf{w}'} = \hat{\mathbf{w}}$
$\implies y^i \left\langle \hat{\mathbf{w}'}, \mathbf{x}^i \right\rangle \geq 1 - \hat{\xi}'_i\ \because\ \hat{\xi}'_k = 0$

Thus, we have a solution which satisfies all the given constraints AND
$\left\|\hat{\mathbf{w}'}\right\|_2^2 + \sum_{i=1}^n \hat{\xi}_i'^2 < \|\hat{\mathbf{w}}\|_2^2 + \sum_{i=1}^n \hat{\xi}_i^2$.
Thus, $\hat{\mathbf{w}}, \left\{\hat{\xi}_i\right\}$ could not have been an optima if $\hat{\xi_k} < 0$ for at least one $k \in [n]$
$\implies$ for an optimal solution, $\xi_k \geq 0\ \forall\ k \in [n]$

2. Lagrangian:

$$\mathcal{L}(\mathbf{w}, \{\xi_i\}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \|\mathbf{w}\|_2^2 + \sum_{i=1}^n \xi_i^2 + \sum_{i=1}^n \boldsymbol{\alpha}_i (1 - \xi_i - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle) + \sum_{i=1}^n \boldsymbol{\beta}_i (-\xi_i)$$

3. Dual Problem:

To get the dual problem, we need to eliminate the primal variables $\mathbf{w}$ and $\{\xi_i\}$ from the Lagrangian.

We first take the gradient of $\mathcal{L}$ wrt $\mathbf{w}$ and equate it to zero.

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, \{\xi_i\}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$$

$$\implies 2\mathbf{w} - \sum_{i=1}^{n} \boldsymbol{\alpha}_i y^i \mathbf{x}^i = 0$$

$$\implies \mathbf{w} = \sum_{i=1}^{n} \frac{\boldsymbol{\alpha}_i}{2} y^i \mathbf{x}^i$$

After substituting this in the equation for the Lagrangian, we get:

$$\mathcal{L}(\{\xi_i\}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \mathbf{x}^i.\mathbf{x}^j + \sum_{i=1}^{n}\xi_i^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \mathbf{x}^i.\mathbf{x}^j + \sum_{i=1}^{n}\boldsymbol{\alpha}_i(1-\xi_i) + \sum_{i=1}^{n}\boldsymbol{\beta}_i(-\xi_i)$$

$$\implies \mathcal{L}(\{\xi_i\}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \mathbf{x}^i.\mathbf{x}^j + \sum_{i=1}^{n}\xi_i^2 + \sum_{i=1}^{n}\boldsymbol{\alpha}_i(1-\xi_i) + \sum_{i=1}^{n}\boldsymbol{\beta}_i(-\xi_i)$$

We now take the gradient wrt $\boldsymbol{\xi}$ and equate it to zero.

$$\nabla_{\boldsymbol{\xi}}\mathcal{L}(\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$$

$$\implies 2\boldsymbol{\xi} - \boldsymbol{\alpha} - \boldsymbol{\beta} = 0$$

$$\implies \boldsymbol{\xi} = \frac{1}{2}(\boldsymbol{\alpha} + \boldsymbol{\beta})$$

Substituting this into the equation for Lagrangian, we get

$$\implies \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \mathbf{x}^i.\mathbf{x}^j + \sum_{i=1}^{n}\frac{1}{4}(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i)^2 + \sum_{i=1}^{n}\boldsymbol{\alpha}_i - \sum_{i=1}^{n}\frac{1}{2}(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i)^2$$

$$\implies \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \mathbf{x}^i.\mathbf{x}^j - \frac{1}{4}\sum_{i=1}^{n}(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i)^2 + \sum_{i=1}^{n}\boldsymbol{\alpha}_i$$

Therefore, the dual problem is:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0} \left(-\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \mathbf{x}^i.\mathbf{x}^j - \frac{1}{4}\sum_{i=1}^{n}(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i)^2 + \sum_{i=1}^{n}\boldsymbol{\alpha}_i \right)$$

4. Comparision with SVM dual:

The dual problem for original SVM is:

$$\max_{0\leq\boldsymbol{\alpha}\leq 1} \left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \mathbf{x}^i.\mathbf{x}^j + \sum_{i=1}^{n}\boldsymbol{\alpha}_i \right)$$

and the dual for (P1) is:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0} \left(-\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \mathbf{x}^i.\mathbf{x}^j - \frac{1}{4}\sum_{i=1}^{n}(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i)^2 + \sum_{i=1}^{n}\boldsymbol{\alpha}_i \right)$$

Some important differences between the two duals are:

- The dual for (P1) involves two sets of Lagrange multipliers $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ for $i \in [n]$ whereas the dual for the original SVM problem involves only $\boldsymbol{\alpha}_i$
- $\boldsymbol{\alpha}$ is bounded on both the sides for the dual of the original SVM problem but it is bounded only on one side for the dual problem of (P1).
- There is a dependence on $(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i)^2$ in the case of the dual problem for (P1) but in the case of the dual problem for the original SVM problem, there is no dependence of such square terms.
- There is a lesser dependence on the cross terms $\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j$ in the case of the dual for (P1) when compared to the dual of the original SVM problem.

5. Bonus:

No, the positivity constraints are not vacuous for the original SVM problem.
If we remove the positivity constraint from the original SVM problem, then we might end up with different solutions.
For example, consider a solution set in which for some value of i, we have $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle > 1$, with the positivity constraint, we would need $\xi_i = 0$ but if we remove the positivity constraint, then we will have at least one better solution i.e keep everything else the same and make $\xi_i' = 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle$.

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

Assignment Number: 2
Student Name: Talla Aravind Reddy
Roll Number: 14746
Date: October 15, 2017

QUESTION

4

3.

For the GD solver, the current iterate $\mathbf{w}^t$