

Assignment Number: 2

Student Name: Talla Aravind Reddy

Roll Number: 14746

Date: October 10, 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.

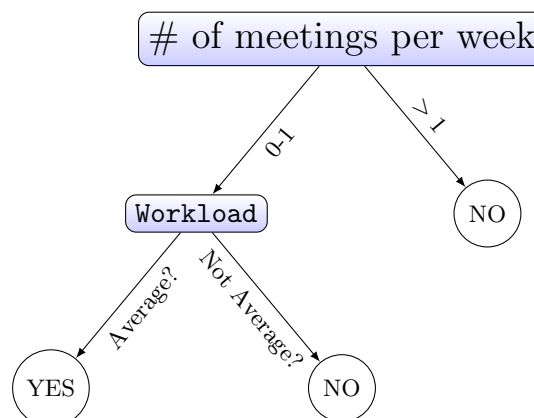
Look at the 4th and 6th entries of the given dataset.

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 $\# \text{ meetings} = 0-1$, information gains for $Size = 0.114, Like = 0.035, Workload = 0.439 \implies Workload$

3.2 Further splitting:

For root node's children, $\# \text{ meetings} > 1$ 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.

Assignment Number: 2

Student Name: Talla Aravind Reddy

Roll Number: 14746

Date: October 10, 2017

This is for Question 2

Assignment Number: 2

Student Name: Talla Aravind Reddy

Roll Number: 14746

Date: October 10, 2017

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 $\hat{\mathbf{w}}, \{\hat{\xi}_i\}$ 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 $\hat{\mathbf{w}}', \{\hat{\xi}'_i\}$ 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

$$y^i \langle \hat{\mathbf{w}}', \mathbf{x}^i \rangle \geq 1 - \hat{\xi}'_i, \text{ for all } i.$$

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

$$y^i \langle \hat{\mathbf{w}}, \mathbf{x}^i \rangle \geq 1 - \hat{\xi}_i \quad \because \text{ it was part of the initial solution}$$

$$\implies y^i \langle \hat{\mathbf{w}}, \mathbf{x}^i \rangle \geq 1 - \hat{\xi}_i$$

$$\implies y^i \langle \hat{\mathbf{w}}, \mathbf{x}^i \rangle \geq 1 \quad \because \hat{\xi}_k < 0$$

$$\implies y^i \langle \hat{\mathbf{w}}', \mathbf{x}^i \rangle \geq 1 \quad \because \hat{\mathbf{w}}' = \hat{\mathbf{w}}$$

$$\implies y^i \langle \hat{\mathbf{w}}', \mathbf{x}^i \rangle \geq 1 - \hat{\xi}'_i \quad \because \hat{\xi}'_k = 0$$

Thus, we have a solution which satisfies all the given constraints AND

$$\|\hat{\mathbf{w}}'\|_2^2 + \sum_{i=1}^n \hat{\xi}_i'^2 < \|\hat{\mathbf{w}}\|_2^2 + \sum_{i=1}^n \hat{\xi}_i^2.$$

Thus, $\hat{\mathbf{w}}, \{\hat{\xi}_i\}$ 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:

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

The primal problem is:

$$\hat{\mathbf{w}}_p, \{\hat{\xi}_i\}_p = \arg \min_{\mathbf{w}, \{\xi_i\}} \left(\arg \max_{\boldsymbol{\alpha} \geq 0} \left(\mathcal{L}(\mathbf{w}, \{\xi_i\}, \boldsymbol{\alpha}) \right) \right)$$

3. Dual Problem:

The primal problem is:

$$\hat{\mathbf{w}}_p, \{\hat{\xi}_i\}_p = \arg \min_{\mathbf{w}, \{\xi_i\}} \left(\arg \max_{\alpha \geq 0} \left(\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \xi_i^2 + \sum_{i=1}^n \alpha_i (1 - \xi_i - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle) \right) \right)$$

The corresponding dual problem is:

$$\hat{\mathbf{w}}_d, \{\hat{\xi}_i\}_d = \arg \max_{\alpha \geq 0} \left(\arg \min_{\mathbf{w}, \{\xi_i\}} \left(\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \xi_i^2 + \sum_{i=1}^n \alpha_i (1 - \xi_i - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle) \right) \right)$$

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

$$\begin{aligned} \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \{\xi_i\}, \alpha) &= 0 \\ \implies 2\mathbf{w} - \sum_{i=1}^n \alpha_i y^i \mathbf{x}^i &= 0 \\ \implies \mathbf{w} &= \sum_{i=1}^n \frac{\alpha_i}{2} y^i \mathbf{x}^i \end{aligned}$$

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

$$\begin{aligned} \mathcal{L}(\{\xi_i\}, \alpha) &= \frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^i y^j \mathbf{x}^i \cdot \mathbf{x}^j + \sum_{i=1}^n \xi_i^2 - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^i y^j \mathbf{x}^i \cdot \mathbf{x}^j + \sum_{i=1}^n \alpha_i (1 - \xi_i) \\ \implies \mathcal{L}(\{\xi_i\}, \alpha) &= -\frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^i y^j \mathbf{x}^i \cdot \mathbf{x}^j + \sum_{i=1}^n \xi_i^2 + \sum_{i=1}^n \alpha_i (1 - \xi_i) \end{aligned}$$

We now take the gradient wrt ξ and equate it to zero.

$$\begin{aligned} \nabla_{\xi} \mathcal{L}(\xi, \alpha) &= 0 \\ \implies 2\xi - \alpha &= 0 \\ \implies \xi &= \frac{1}{2} \alpha \end{aligned}$$

Substituting this into the equation for Lagrangian, we get

$$\begin{aligned} \implies \mathcal{L}(\alpha) &= -\frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^i y^j \mathbf{x}^i \cdot \mathbf{x}^j + \sum_{i=1}^n \frac{1}{4} \alpha_i^2 + \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \frac{1}{2} \alpha_i^2 \\ \mathcal{L}(\alpha) &= -\frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^i y^j \mathbf{x}^i \cdot \mathbf{x}^j - \frac{1}{4} \sum_{i=1}^n \alpha_i^2 + \sum_{i=1}^n \alpha_i \end{aligned}$$

Therefore, the dual problem gets converted to:

$$\arg \max_{\alpha \geq 0} \left(-\frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^i y^j \mathbf{x}^i \cdot \mathbf{x}^j - \frac{1}{4} \sum_{i=1}^n \alpha_i^2 + \sum_{i=1}^n \alpha_i \right)$$

4. Comparison with SVM dual:

The dual problem for original SVM is:

$$\arg \max_{0 \leq \alpha \leq 1} \left(-\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^i y^j \mathbf{x}^i \cdot \mathbf{x}^j + \sum_{i=1}^n \alpha_i \right)$$

and the dual for (P1) is:

$$\arg \max_{\alpha \geq 0} \left(-\frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y^i y^j \mathbf{x}^i \cdot \mathbf{x}^j - \frac{1}{4} \sum_{i=1}^n \alpha_i^2 + \sum_{i=1}^n \alpha_i \right)$$

Some important differences between the two duals are:

- α is bounded on both the sides for the dual of the original SVM problem but it is bounded only on one side for (P1).
- There is a dependence on the α_i^2 in the case of the dual problem for (P1) but not in the case of the dual problem for the original SVM problem.
- There is a lesser dependence on the cross terms $\alpha_i \alpha_j$ 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 $y^i \langle \mathbf{w}, \mathbf{x}^i \rangle > 1$, 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 $\xi'_i = 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle$

Assignment Number: 2

Student Name: Talla Aravind Reddy

Roll Number: 14746

Date: October 10, 2017

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis feugiat vehicula dolor, sed ultricies leo. Phasellus euismod dictum felis in euismod. Proin pretium vel neque in placerat. Proin imperdiet egestas vulputate. Etiam faucibus accumsan ante non viverra. Duis ultrices ac odio vel sodales. In maximus gravida dolor, ut commodo lacus. Pellentesque ante massa, venenatis id aliquam et, posuere sed dui. Duis dignissim justo sit amet augue posuere fringilla. Suspendisse at nisi gravida, mattis justo sit amet, elementum elit. Praesent et massa ornare, consequat dui eget, ornare risus. Duis est nibh, sollicitudin nec mattis non, mattis in leo. Donec finibus justo sed massa sagittis, non fermentum nibh dictum. Pellentesque et congue purus. Donec porta pretium porttitor.

Morbi euismod risus eu tortor ornare malesuada. Nunc sed sollicitudin neque, efficitur rhoncus tellus. Cras malesuada augue arcu. Sed sem odio, tincidunt quis laoreet ac, facilisis ut nibh. Quisque gravida dolor at egestas aliquam. Aenean mollis massa sit amet enim mattis, vel fermentum tortor facilisis. Donec pellentesque est velit, vitae posuere lorem tristique ut.