Determining the Restrictions on Concatenating Features to a Linear Regression Model: A Proof

Given N weak learners/features $h_i : x \in [-1,1]$, N weights $w_i \in \{0,1\}$ ($i \in \{1,2,\ldots,N\}$), the output of the strong classifier for S input programs $y_s \in \{-1,1\}$ ($s \in \{1,2,\ldots,S\}$), and a strong classifier $H_{1,N}(x) = \text{sign}\left(\sum_{i=1}^{N} w_i h_i(x)\right)$, the loss for a particular ordered N-tuple of binary weights $\boldsymbol{w}^{1,N} = (w_1, w_2, \ldots, w_N)$ is equal to the output of the following quadratic cost function: $L_{1,N}(\boldsymbol{w}^{1,N}) = \sum_{s=1}^{S}\left(\frac{1}{N}\sum_{i=1}^{N} w_i h_i(x_s) - y_s\right)^2$. This loss function is directly correlated to the linear regression model. We are given that $\boldsymbol{w}_{\mathbf{0}}^{1,N}$ denotes the N-tuple of binary weights under the condition that $L_{1,N}(\boldsymbol{w}_{\mathbf{0}}^{1,N})$ is the least possible value of the function $L_{1,N}(\boldsymbol{w}^{1,N})$. Similarly, let $\boldsymbol{w}_{\mathbf{1}}^{1,N}$ be an N-tuple of binary weights such that $L_{1,N}(\boldsymbol{w}_{\mathbf{1}}^{1,N})$ is extremely close to the lowest possible value of our given cost function.

Given another set of M weak learners $h_i : x \in [-1,1]$ (completely disjoint to our original set of N weak learners) and M binary variables $w_i \in \{0,1\}$ ($i \in \{N+1, N+2, \ldots,, N+M\}$), we have a corresponding strong classifier $H_{N+1,N+M}(x) = \text{sign}\left(\sum_{i=N+1}^{N+M} w_i h_i(x)\right)$, and a corresponding loss function for those weak classifiers $L_{N+1,N+M}(\boldsymbol{w}^{N+1,N+M}) = \sum_{s=1}^{S}\left(\frac{1}{M}\sum_{i=N+1}^{N+M} w_i h_i(x_s) - y_s\right)^2$. Similarly, let $\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}$ denote the M-tuple of binary weights under the condition that $L_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M})$ is relatively close to the least possible value of the function $L_{N+1,N+M}(\boldsymbol{w}^{N+1,N+M})$. Our goal is to optimize the strong classifier $H_{1,N+M}(x) = \text{sign}\left(\sum_{i=1}^{N+M} w_i h_i(x)\right)$ by minimizing the loss of the quadratic function $L_{1,N+M}(\boldsymbol{w}^{1,N+M}) = \sum_{s=1}^{S}\left(\frac{1}{N+M}\sum_{i=1}^{N+M} w_i h_i(x_s) - y_s\right)^2$, given the information that the function $L_{1,N+M}(\boldsymbol{w}^{1,N+M})$ presents a quadratic function that is too large a problem size to be solved by D-Wave's quantum annealer. We will only use the information provided by $\boldsymbol{w}_{\mathbf{0}}^{1,N}$, $\boldsymbol{w}_{\mathbf{1}}^{1,N}$, and $\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}$, because those vectors can be obtained by using D-Wave's quantum annealer to directly optimize $L_{1,N}(\boldsymbol{w}^{1,N})$ and $L_{N+1,N+M}(\boldsymbol{w}^{N+1,N+M})$, two problems that are sufficiently small for a quantum annealer.

If we let $\boldsymbol{w}_{\mathbf{00}}^{1,N+M}$ be the vector formed by concatenating $\boldsymbol{w}_{\mathbf{0}}^{1,N}$ and $\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}$, and $\boldsymbol{w}_{\mathbf{10}}^{1,N+M}$ be the vector formed by concatenating $\boldsymbol{w}_{\mathbf{1}}^{1,N}$ and $\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}$, we would like to investigate when $L_{N+M}(\boldsymbol{w}_{\mathbf{10}}^{N+M}) < L_{N+M}(\boldsymbol{w}_{\mathbf{00}}^{N+M})$. For the sake of convenience in our proof let us assume that the value of $L_{1,N}(\boldsymbol{w}_{\mathbf{1}}^{1,N})$ is extremely close to the value of $L_{1,N}(\boldsymbol{w}_{\mathbf{0}}^{1,N})$. Write the loss function $L_{1,N}(\boldsymbol{w}^{1,N})$ as $L_{1,N}(\boldsymbol{w}^{1,N}) = \sum_{s=1}^{S}\left(\frac{1}{N}f_{1,N}(\boldsymbol{w}^{1,N}, x_s) - y_s\right)^2$, utilizing the substitution $f_{k_1,k_2}(\boldsymbol{w}^{k_1,k_2}, x) = \sum_{i=k_1}^{k_2} w_i h_i(x)$. Then $L_{1,N+M}(\boldsymbol{w}^{1,N+M}) = \sum_{s=1}^{S}\left(\frac{1}{M+N}\left(f_{1,N}(\boldsymbol{w}^{1,N}, x_s) + f_{N+1,N+M}(\boldsymbol{w}^{N+1,N+M}, x_s)\right) - y_s\right)^2$. Making the algebraic substitution $C_s = f_{1,N}(\boldsymbol{w}^{1,N}, x_s) - y_s N$, we find that $L_{1,N}(\boldsymbol{w}^{1,N}) = \sum_{s=1}^{S}\left(\frac{C_s}{N}\right)^2$. Making the same substitution into our other quadratic function we obtain $L_{1,N+M}(\boldsymbol{w}^{1,N+M}) = \sum_{s=1}^{S}\left(\frac{C_s - y_s M + f_{N+1,N+M}(\boldsymbol{w}^{N+1,N+M}, x_s)}{N+M}\right)^2$. However, since $C_s = f_{1,N}(\boldsymbol{w}^{1,N}, x_s) - y_s N$ and $|y_s N| \geq |f_{1,N}(\boldsymbol{w}^{1,N}, x_s)|$, the sign of C_s is uniquely determined by the sign of y_s (in fact, the two values have opposite signs); therefore, if we impose the constraint $C_s \geq 0$ we can simplify $L_{1,N+M}(\boldsymbol{w}^{N+M}) = \sum_{s=1}^{S}\left(\frac{C_s - \text{sign}(y_s) f_{N+1,N+M}(\boldsymbol{w}^{N+1,N+M}, x_s) + M}{N+M}\right)^2$.

Since we specified earlier that the value of $L_{1,N}(\boldsymbol{w}_{\mathbf{0}}^{1,N})$ is extremely close to the value of $L_{1,N}(\boldsymbol{w}_{\mathbf{1}}^{1,N})$, we set the equation: $L_{1,N}(\boldsymbol{w}_{\mathbf{0}}^{1,N}) \approx L_{1,N}(\boldsymbol{w}_{\mathbf{1}}^{1,N})$. We also set the following two equalities: $L_{1,N}(\boldsymbol{w}_{\mathbf{0}}^{1,N}) = \sum_{s=1}^{S}\left(\frac{C_{s0}}{N}\right)^2$, $L_{1,N}(\boldsymbol{w}_{\mathbf{1}}^{1,N}) = \sum_{s=1}^{S}\left(\frac{C_{s1}}{N}\right)^2$ (where C_{s0} and C_{s1} are similar to the aforementioned C_s). This implies that $\sum_{s=1}^{S} C_{s0}^2 \approx \sum_{s=1}^{S} C_{s1}^2$, an equality very important to keep in mind moving forward. Based off of the algebraic insight developed earlier, we have $L_{1,N+M}(\boldsymbol{w}_{\mathbf{00}}^{1,N+M}) = \sum_{s=1}^{S}\left(\frac{C_{s0} - \text{sign}(y_s) f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}, x_s) + M}{N+M}\right)^2$

and $L_{1,N+M}(\boldsymbol{w}_{\mathbf{10}}^{1,N+M}) = \sum_{s=1}^{S}\left(\frac{C_{s1}-\text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M},x_s)+M}{N+M}\right)^2$. Expansion of these two expressions yields $L_{1,N+M}(\boldsymbol{w}_{\mathbf{00}}^{1,N+M}) = \sum_{s=1}^{S}\frac{C_{s0}^2+2C_{s0}(M-\text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M},x_s))+(M-\text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M},x_s))^2}{M^2+2MN+N^2}$ and $L_{1,N+M}(\boldsymbol{w}_{\mathbf{10}}^{1,N+M}) = \sum_{s=1}^{S}\frac{C_{s1}^2+2C_{s1}(M-\text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M},x_s))+(M-\text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M},x_s))^2}{M^2+2MN+N^2}$.

However, note that the values of $L_{1,N+M}(\boldsymbol{w}_{\mathbf{00}}^{1,N+M})$ and $L_{1,N+M}(\boldsymbol{w}_{\mathbf{10}}^{1,N+M})$ are only dependent on the sums $G_0 = \sum_{s=1}^{S} 2C_{s0}(M - \text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}, x_s))$ and $G_1 = \sum_{s=1}^{S} 2C_{s1}(M - \text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}, x_s))$ respectively. This is due to the fact that in our expanded representations of $L_{1,N+M}(\boldsymbol{w}_{\mathbf{00}}^{1,N+M})$ and $L_{1,N+M}(\boldsymbol{w}_{\mathbf{10}}^{1,N+M})$, we can equate $\sum_{s=1}^{S} C_{s0}^2 \approx \sum_{s=1}^{S} C_{s1}^2$, and the constant term $\sum_{s=1}^{S}(M-\text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}, x_s))^2$ appears in both expansions. Therefore, if we analyze the term $G = \sum_{s=1}^{S} 2C_s(M-\text{sign}(y_s)f_{N+1,N+M}(\boldsymbol{w}_{\mathbf{0}}^{N+1,N+M}, x_s))$ for a fixed binary vector of dimension M concatenated to a variable binary vector of dimension N, we could determine whether or not the binary vector of dimension N should be the lowest energy binary vector found by the quantum annealer. We could not accurately predict what G would be equal to, since this term would vary from problem to problem, but a computer program can easily compute G, and G only needs to be computed for solutions to $L_{1,N}(\boldsymbol{w}^{1,N})$ close to the lowest energy solution $\boldsymbol{w}_{\mathbf{0}}^{1,N}$, otherwise our equality $\sum_{s=1}^{S} C_{s0}^2 \approx \sum_{s=1}^{S} C_{s1}^2$ no longer holds. Hence, this process can easily be used to determine which binary vector of dimension N close to the lowest energy solution should be concatenated to by a binary vector of dimension M. $\square$