CS6140 - Machine Learning

HW1

Kartikeya Shukla - 001727475

Question 3:

- Assume a decision tree with a query repeating along some path. While traversing through that feature/threshold pairs, it is understood that we will not be visiting any other branches. So, the branches that were not part of the query can be pruned and along with any branches of other occurrence of the query in the tree.
  This would make the lower occurrences of invalid since, all the nodes would have a single child, and can simply be replaced by the left node of the branch.
  Applying this procedure to every possible path and with every query we can convert any given tree to a tree with distinct splits on each path.

a) Assume an arbitrary decision tree, which has a root X and has a branching ratio B. i.e. we can represent the root node X and its B branches as

$$(X, \{b_1\ b_2, b_3, b_4 \ldots b_B\})$$

We can also assume that all of the B children are themselves a root to a sub-tree of arbitrary depth.

Now if B=2, this becomes a Binary Tree.
If B=3, the tree can be represented as $(X, \{b_1, b_2, b_3\})$, which can be written as

$$(X, \{b_1, X_2\})$$

and sub-tree $(X_2, \{b_2, b_3\})$.

Therefore, for B ≥ 3, every root can be represented as $(X_{k-1},\{b_{k-1},b_k\})$.
Applying this to every node in a tree can convert the tree to a binary tree.

b) The upper bound will be log(B), since, in the first level the two nodes would have B/2 each, and further B/4, B/8 etc.
The lower bound is 2.

c) Lower bound: 3 and Upper bound: 2*B - 1 , where B is the branching factor.

Question 4:

a) Consider there is a Yes/No decision tree,
Suppose at a node there is maximum randomness, which in this case means, the node contains half the counts of Yes and No. Then, if we are to split this node, the best scenario we can achieve is separating each label into a child node (pure node -> Entropy 0).
This would give the biggest gain in entropy.

So, Entropy on parent node

$$= - 0.5 \log 0.5 - 0.5 \log 0.5$$
$$= - \log_2 0.5$$
$$= - \log_2 1 + \log 2$$
$$= 1$$

Therefore, here we can see that the maximum drop in entropy can be at most 1.

b) Similarly, for trees with Branching ratio B > 1, entropy will not be greater than $\log_2(m)$, where m is the number of classes.

Question 5:

5. Given the datapoints (x_i, y_i) where $i = 1, 2, \cdots n$ we can now represent them as matrix

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{bmatrix}$$

Then the error associated with othe regressor $h_w(x) = \sum_d w^d x^d$

$$E = \begin{bmatrix} h_w(x_1) - y_1 \\ \cdots \\ h_w(x_m) - y_m \end{bmatrix} = \begin{bmatrix} x_1 w \\ \vdots \\ x_m w \end{bmatrix} - \begin{bmatrix} y_1 \\ \cdots \\ y_m \end{bmatrix} = Xw - Y$$

can be written as

$$E^2 = \sum_1 [y_i - f(x_1, a_1, a_2 \cdots a_n)]^2$$

for a linear fit

$$f(a, b) = a + bx$$

so $E^2(a, b) = \sum_1 [y_i - (a + bx_i)]^2$

$$\frac{\partial (R^2)}{\partial a} = -2 \sum_{i=1}^{n} [y_i - (a + bx)] = 0$$

and $\dfrac{\partial R^2}{\partial b} = -2 \sum_{i=1} [y_i - (a + bx_i)] x_i = 0.$

$$\Rightarrow \quad na + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$= \quad a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

In matrix form,

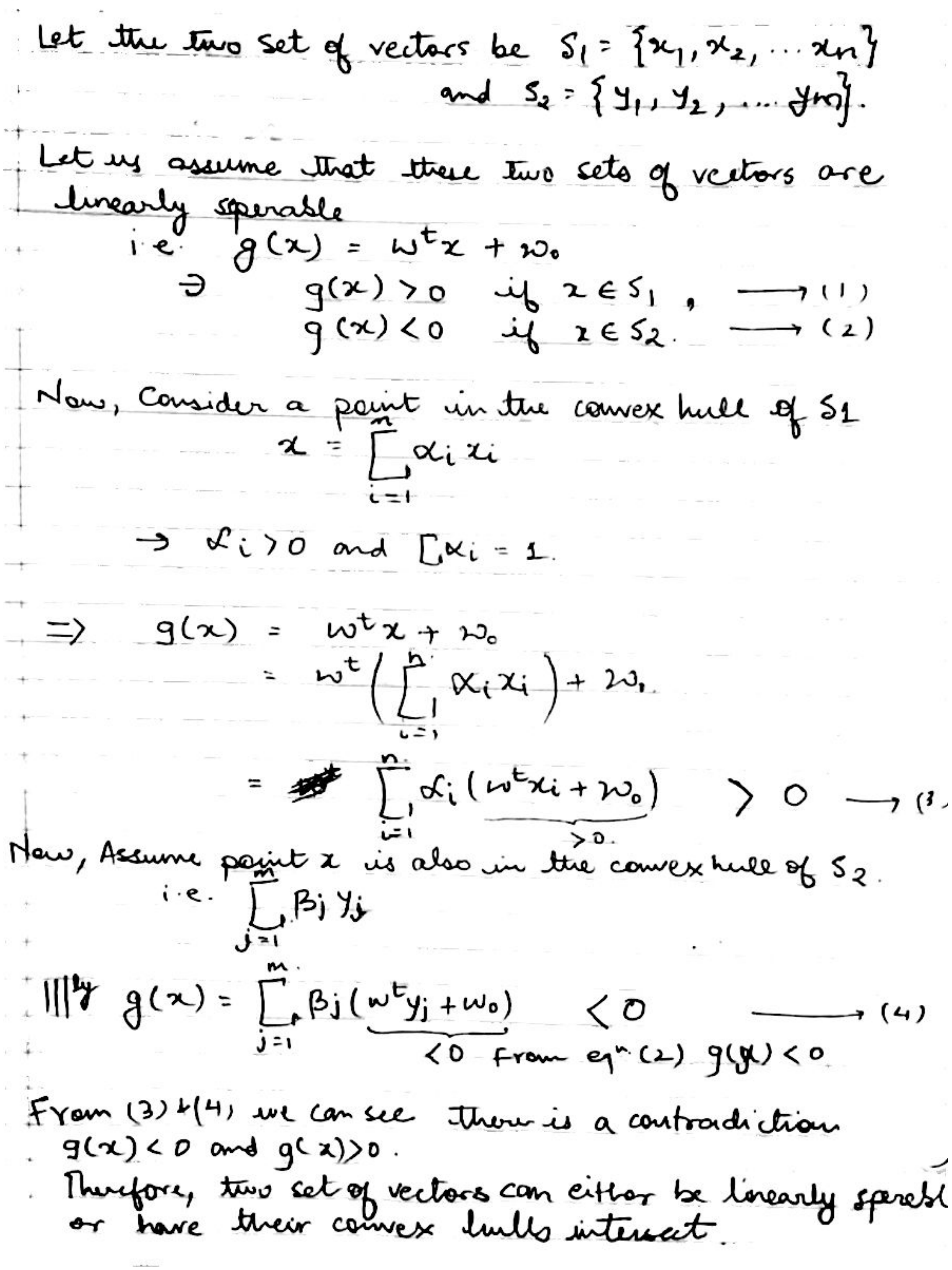
$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_1^n x_i \\ \sum_1^n x_i & \sum_1^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_1^n y_i \\ \sum_1^n x_i y_i \end{bmatrix}.$$

Question 7:

Consider two set of vectors that are linearly separable, which means that there exists a hyper-plane for which one set of points would give a result greater than zero and the other less than 0 (when substituting the vector points in the equation of the hyper-plane).

Now if these set of vectors have intersecting convex hulls (by definition means intersection of all convex sets containing that vector) - we can see that there is a contradiction.

And therefore, two sets of vectors can be one or the other.



Let the two set of vectors be $S_1 = \{x_1, x_2, \dots x_n\}$
and $S_2 = \{y_1, y_2, \dots y_m\}$.

Let us assume that these two sets of vectors are linearly sperable

i.e. $g(x) = w^t x + w_0$

$\ni \quad g(x) > 0$ if $x \in S_1$, $\longrightarrow (1)$

$g(x) < 0$ if $x \in S_2$. $\longrightarrow (2)$

Now, Consider a point in the convex hull of S_1

$$x = \sum_{i=1}^{n} \alpha_i x_i$$

$\rightarrow \alpha_i > 0$ and $\sum_1 \alpha_i = 1$.

$$\Rightarrow \quad g(x) = w^t x + w_0$$

$$= w^t \left(\sum_{i=1}^{n} \alpha_i x_i \right) + w_0.$$

$$= \sum_{i=1}^{n} \alpha_i (\underbrace{w^t x_i + w_0}_{>0.}) > 0 \longrightarrow (3$$

Now, Assume point x is also in the convex hull of S_2.

i.e. $\sum_{j=1}^{m} \beta_j y_j$

$$|||^{ly} \quad g(x) = \sum_{j=1}^{m} \beta_j (\underbrace{w^t y_j + w_0}_{<0 \text{ From eq}^n (2)\ g(y) < 0}) < 0 \longrightarrow (4)$$

From (3) & (4), we can see there is a contradiction
$g(x) < 0$ and $g(x) > 0$.
Therefore, two set of vectors can either be linearly sperabl or have their convex hulls intersect.