CSCI-599 Machine Learning Theory
Assignment - 1

Sree Priyanka Uppu
(uppu@usc.edu)

1 Relationships between concept classes

a.

Say a decision list is represented as "if l_1 then b_1 else if l_2 then b_2,... else b_m" where each l_i is a literal and each $b_i \in \{0,1\}$.

Disjunctions are special case of Decision Lists: Using the above notation, for any clause of literals $(x_1 \vee \overline{x_2} \vee x_3)$ we can define it by "if l_1 then 1 else if l_2 then 1, ... else 0", which in effect is a disjunction.

Conjunctions are special case of Decision Lists: Using the above notation, for any clause of literals $(x_1 \wedge \overline{x_2} \wedge x_3)$ we can transform it to a disjunction using De' Morgan's law and then transform it to a decision list as mentioned above. Another way to do is to define as "if $\overline{l_1}$ then 0 else if $\overline{l_2}$ then 0, ... else 1", which in effect is a conjunction.

b.

Decisions lists are a special case of Linear Threshold Functions: Assume a decision list is represented as "if x_1 then b_1 else if x_2 then b_2,... else x_n" where each $b_i \in \{+,-\}$. And assume that the linear function is of the form f(x) = + iff $w_1\ x_1 + w_2\ x_2 ....w_n\ x_n \geq w_0$, 0 otherwise; and we transform the decision list to linear threshold function for an x_i as : if b=+ then w_i=2^{-i} and if b=- then w_i=-2^{-i}. The w_0 would be 1 if the final else is + otherwise it will be -1. On the other hand, for a $\overline{x_i}$ it would be replaced with $1 - x_i$ and the transformation for weights would be same as above. The w_0 would be 1 if the final else is + otherwise it will be -1.The intuition behind this is that the weight of the first variables if true will always be greater than the sum of the absolute weights of the rest of the terms.

2 VC-dimension of simple concept classes

(a)

VC-dim(Parity function) = n
Let S=<...0001000...>,<1000...,<000...1> be a set of bits with only one bit set. For each element is S, h(s_j)=1 j ∈ S, else 0. Hence, VC-Dim ≥ n. To prove VC-Dim ≤ n: There are 2^n possible parity functions. For a finite class C: VC-Dim(C) ≤ log ‖ C ‖ Applying to our case, we have: VC-Dim ≤ log 2^n= n Thus, VC-dim=n.

(b)

VC-dim(Axis Aligned Rectangles) = 2n
For n=2, the VC-dim=4,Positive examples are points inside the rectangle, and negative examples are points outside the rectangle. VC-dim<5:Given a set of five points in the plane, there must be some point that is neither the extreme left, right, top or bottom point of the five. If we label this non-external point negative and the remaining four external points positive, no rectangle can satisfy the assignment. The same justification applies for n=3 and so on. Generalizing, for any axis aligned rectangle in any n dimension, VC-dim=2n.

(c)

VC-dim(Boolean conjunctions) = n
Let X_i be the string of all 1's except in i'th position, where it is 0. Let us define, S = X_i for all i=1 to n. For example if n = 3,we have T={011, 101, 110}. We now show that S is shattered by taking a x∈ S. Let S = {i|$x_i$is a negative example} and c = conjunction of all x_i. If i ∈ S, x_i is a negative example for c since x_i is in the conjunction; otherwise, O_i is positive. Hence, it is shattered and generalizing VC-Dim=n.

3 Problem 3

Let C be a concept class for an online non-conservative learning algorithm A, with a finite mistake bound of M. Assume that we run the algorithm A and end up getting k examples with the mistake. Let A' be the conservative algorithm for the same concept class C. If we run the algorithm A' on the k

mistaken examples from A, we would still get a mistake bound of k. Hence, the mistake bound of a $A^{'}$ would also be M.

4 Problem 4

Given: Initial hypothesis: w=w^{init}
Linear threshold function: w.x≥0
For a true label of x, if prediction is incorrect w←w-(w.x)x
$\|x\|_2$=1
Convergence Upper Bound Proof: For every mistake the update is as below:
$w^{'}$=w-(w.x)x
Squaring on both sides
$\| w^{'} \|^2 = \| w \|^2$ - 2$\| w(w.x)x \| + \| (w.x)x \|^2$
$\| w^{'} \|^2 = \| w \|^2$ - $\| (w.x) \|^2$ Since, $\|x\|_2$=1

$$\frac{\| w^{'} \|^2}{\| w \|^2} = 1 - \frac{\| (w.x) \|^2}{\| w \|^2} \leq 1 - \delta^2$$

$$\| w^{'} \|^2 \leq 1 - \delta^2, after 1 mistake$$

After m mistakes,

$$\| w^{'} \|_m^2 \leq (1 - \delta^2)^m$$

$$\| w^{'} \|_m^2 \leq \sqrt{1 - \delta^2)^m} - -(1)$$

Convergence Lower Bound Proof:
w=w_{init}=-(w.x)x
w-w_{init}= -(w.x)x
After m mistakes,

$$w_m - w_{init} = \sum_{i=0}^{m} -(w.x)x$$

Multiplying by V on both sides,

$$w_m V - w_{init} V = -\sum_{i=0}^{m} (w.x)xV$$

$$w_m V = -\sum_{i=0}^{m} (w.x)xV + w_{init} V$$

$$\| w_m V \| \geq \gamma, \| V \|_2 = 1$$

$$\| w_m \| \geq \gamma - -(2)$$

Combining (1) and (2)

$$\sqrt{1-\delta^2)^m} \geq \gamma$$

Taking Log on both sides,

$$\frac{m}{2} ln(1-\delta^2) \geq ln(\gamma)$$

$$-\frac{m}{2} ln(1-\delta^2) \geq -ln(\gamma)$$

We have, -ln(1-x) $\geq$ x and replacing it in above equation, we get

$$\delta^2 \leq -\frac{2ln(\gamma)}{m}$$

Simplifying we get,

$$m \leq \frac{2ln(1/\gamma)}{\delta^2}$$

5 Problem 5

We know that the mistake bound, M $\leq \frac{n}{\delta^2}$. Intuitively, it means that for higher dimensions of n and where δ (angle between w and v) is very small the M would be really high making it exponential. For example, data in DNA sequence which is high dimensional will be a good example to show the exponential run time of the algorithm. Below is an example, let C=<1/2,-1/4,1/8,-1/16,....> and the examples provided to that be of the form {<1,0,0,...>,<1,1,0,...>,<1,1,1,0...>,...<1,1,1...1,1,1>} and by maintaining the order of examples we can get the hypothesis to come close to C exponentially. At each step, the example changes by at most 1 bit.Since, there are n bit and 2^n examples, hence exponential number of changes in the hypothesis.