Intro To ML – HW3

students:

Dor Bank - 301380416 - [dorbank@gmail.com](mailto:dorbank@gmail.com),

Avi Caciularu - 203056585 - avi.c33@gmail.com

**Theory Questions**

**Question 1:**

We will show a lower bound: . let be a unit vector from the standard basis of size , where '' appears in the place.

We will begin by showing that the set can be shattered by . In order to show that, we need to show that every label assignment for is consistent with some hypothesis in . let be the assignment for the vectors . Then define hypothesis as:

.

Furthermore, for the computation of the result at the output node we will set:

We get

Note that here we assume that . the permutation for is trivial.

**Question 2:**

In this questions we will use the following lemmas, proved in recitation 7:

Lemma 1:

Lemma 2:

Lemma 3:

1. Denote the function family of each node *j* in layer *i* as . In our case, since the node separators are independent of each other, we get that for each layer *i*, . Because there are d nodes at each layer, we can see that all are equal to our wanted . From Radon's theorem, we get that for each , . from Lemma 2 we get that
2. is simply composed from concatenating the family times. Explicitly:

. From Lemma 3 we get that

1. For every node, we have parameters for the weight vector w, and more for the bias . In total, we have parameters, layers with nodes, and layer (the final one) with one node. Combining all together we get

.

1. Assume . First notice that . otherwise we get which is surely not true for a large enough .

Now, by taking the log of each size we get:

, where the second transition is true as stated above.

1. first notice that . Now, let there be a set of size that is shattered. Then . combining this with the inequality mentioned we get . from section d we get that . Thus, we conclude that .

**Question 3:**

1. Consider the new primal problem of SVM:

The regular SGD step was:

Denote so

We will show that:

And then the new SGD with the “Subgradient Projection Method” step would be:

The definition of the projection of on ( ) is .

If then the trivial solution is , so we will assume that :

* If so and then .
* If so and we get that the normalized vector and then , because is the optimal solution for .

1. Let be a convex set, and .

Because is a convex set, there is so that .

Let . Particularly, this equality is correct for : Hence . From the Pythagoras’s theorem, we get that meaning that . Hence, we get by the triangle inequality.

1. Property of convex set (without proof):

Let be a closed convex set, and , so the non-expensive operator maintains .

We have the same requirements of the original proof. the minimizer demands since it is in the last iteration. Hence, the expression in the equation after (12) fulfills: , from the above property. More intuitively- we minimize the norm of the vector by applying the ball projection on it, so all the expression’s norm should be smaller.

After that we continue the analyze and get:

and after that (only after equation (13)), the rest of the proof remains the same, and Theorem 1.1 still holds for stochastic Subgradient Projection Method.

**Question 4:**

1. By assigning K=2 in the given problem we get:

case

case

The left summation is just the regularization factor. notice that by setting (meaning 2 maps to 1 and 1 maps to -1) and we get that

1. The subgradient is of course

Using it for our new step in our SGD algorithm we get a step of

**Question 5:**

proof by induction on .

base: when there can be no samples so only 1 node is needed to always return the required answer (1 or 0).

step: assume the claim is true for a classifiers with domain dimensions up to a specific d. thus means that there is a decision tree of height at most d+1 as stated in the question. for a classifier with d+1 dimensions, we use the first node as .

we get 2 sub trees with d dimension (which are relevant), and by the induction assumption the those sub trees are of height at most d+1. Therefore, the new tree is of height at most (d+1)+1 as needed.

now please notice that each binary tree of height d+1 has leaf nodes (and therefore different paths). Also, notice that over the domain of there can be maximum different sample vectors.

For a lower bound, we take the sample group S of size which contains different vectors. Every node in the d+1 height binary tree will map to a specific sample. Each node will return the label of the corresponding label, and the correctness is obvious.

For an upper bound we tale a sample group S of size . Based on what said before, S contains 2 samples such that and . Both of these samples will be mapped by the tree model to the same leaf node. So, by labeling those samples differently, we get that S is not shattered.

Thus we conclude that the VC-dimension of the class of decision trees over the domain of is .