Computational Learning Theory
Michaelmas Term 2023

Problem Sheet 3

Instructions: The problem sheets are designed to increase your understanding of the material taught in the lectures, as well as to prepare you for the final exam. You should attempt to solve the problems on your own after reading the lecture notes and other posted material, where applicable. Problems marked with an asterisk are optional. Once you have given sufficient thought to a problem, if you are stuck, you are encouraged to discuss with others in the course and with the lecturer during office hours. You are *not permitted* to search for solutions online.

1 VC Dimension of Linear Halfspaces in $\mathbb{R}^n$

We will show that the concept class of linear halfspaces in $\mathbb{R}^n$ has VC-dimension $n+1$.

1. Give a set of $n+1$ points in $\mathbb{R}^n$ that is shattered by the class of linear halfspaces.
2. We want to show that no set of $m = n+2$ points in $\mathbb{R}^n$ can be shattered by the class of linear halfspaces. For this you can use what is called as Radon's theorem, described below.
3.* Prove Radon's theorem.

Given a set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathbb{R}^n$, the convex hull of S is the set,

$$\{\mathbf{z} \in \mathbb{R}^n \mid \exists \lambda_1, \ldots, \lambda_m \in [0,1], \sum_{i=1}^{m} \lambda_i = 1, z = \sum_{i=1}^{m} \lambda_i \mathbf{x}_i\}$$

Radon's Theorem: Let $m \geq n+2$, then S must have two disjoint subsets S_1 and S_2 whose convex hulls intersect.

Hints:

1. Consider the set consisting of the origin and the basis vectors.
2. Show that if a set of points is contained in a linear halfspace of $\mathbb{R}^n$, then the entire convex hull of those points is also contained in that linear halfspace.
3. For some $\mu_i \in \mathbb{R}$, argue that when $m \geq n+2$, there is a non-trivial solution to the following set of equations:

$$\sum_{i=1}^{m} \mu_i \mathbf{x}_i = 0,$$
$$\sum_{i=1}^{m} \mu_i = 0.$$

2 Properties of AdaBoost

Consider the AdaBoost algorithm as described in the lecture notes and assume that the weak learning algorithm succeeds with probability 1 at each iteration.

1. Show that the error of h_t with respect to the distribution D_{t+1} is exactly $1/2$.
2. What is the maximum possible value of $D_t(\mathbf{x}_i)$ for some $1 \leq t \leq T$ and $1 \leq i \leq m$?
3. Fix some example, say i, let t_i be the first iteration such that $h_{t_i}(\mathbf{x}_i) = y_i$. How large can t_i be?

Hints

1. Use the definition of α_t and simply calculate.
2. Consider what is the largest value of $D_t(\mathbf{x}_i)$ for which it may still be the case that h_t is incorrect on $\mathbf{x}_i$ and follow the calculations from the previous part.
3. By how much must $D_t(\mathbf{x}_i)$ increase if h_t is incorrect on $\mathbf{x}_i$? Use this and the bound from the previous part.

3 Weak Learning CONJUNCTIONS and PARITIES

Consider the instance space $X_n = \{0,1\}^n$. Consider the following hypothesis class:

$$H_n = \{0, 1, z_1, \overline{z}_1, z_2, \overline{z}_2, \ldots, z_n, \overline{z}_n\}.$$

The hypothesis class contains $2n + 2$ functions. The functions "0" and "1" are constant and predict 0 and 1 on all instances in X_n. The function "z_i" evaluates to 1 on any $\mathbf{x} \in \{0,1\}^n$ satisfying $x_i = 1$ and 0 otherwise. Likewise, the function "$\overline{z}_i$" evaluates to 1 on any $\mathbf{x} \in \{0,1\}^n$ satisfying $x_i = 0$ and 0 otherwise. Thus a single bit of the input determines the value of these functions; for this reason these functions are sometimes referred to as *dictator* functions.

1. Show that the class CONJUNCTIONS is $\frac{1}{10n}$-weak learnable using H.
2. Let CONJUNCTIONS_k denote the class of conjunctions on at most k literals. Give an algorithm that PAC-learns CONJUNCTIONS_k and has sample complexity polynomial in k, $\log n$, $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. What would be the sample complexity if you had used the algorithm for learning CONJUNCTIONS discussed in the lectures?
3. Show that there does not exist a weak learning algorithm for PARITIES using H.

Hints:

1. The factor 10 is not particularly important, just a sufficiently large constant.
2. First show that the weak learning algorithm in the previous part can be modified to be a $\frac{1}{10k}$-weak learner in this case.

3. Consider the uniform distribution over $\{0,1\}^n$ and see if any of the hypotheses in H_n would work.