Homework 1

Technical details

TA in Charge: Guy Uziel
Email: guziel@campus.technion.ac.il

Each student must submit individually.

Submit only electronically in the webcourse site

Submit a pdf file, edited by a document editor (not scanned).

1 VC-Dimension of Finite Cardinality Sets

Let $X = [n]$ (the set of numbers $1 \ldots n$) and let $\mathcal{H} = \{ h_S : S \subseteq [n], |S| = k \}$ denote the set of hypotheses defined by

$$h_S(x) = \begin{cases} +1 & x \in S \\ -1 & \text{otherwise} \end{cases}.$$  

In words, this is the set of hypotheses for which exactly $k$ data points are positive (and the rest are negative). What is the cardinality of $\mathcal{H}$? What is the VC-dimension of $\mathcal{H}$?

2 Upper bounds using homogeneous half-spaces

In class we have seen that the VC-dimension of the set of homogeneous half-spaces in $\mathbb{R}^d$ is $d$. This result can be very helpful in proving an upper bound on some other hypothesis set. Consider the following set of non-homogeneous half-spaces:

$$\mathcal{X} = \mathbb{R}^d, \mathcal{H}_d = \{ h_{w,b}(x) = [[\langle w, x \rangle > b]] | w \in \mathbb{R}^d \}$$

Each classifier is a half space, but the margin doesn’t necessarily go through the origin.

(a) Show that we can identify each classifier in $\mathcal{H}_d$ with an homogeneous half spaces in $\mathbb{R}^{d+1}$
(b) Using the result from (a) give an upper bound on the VC-dimension of \( H_d \)

c) Prove that the upper bound you found in (b) is indeed the VC-dimension of \( H_d \)

d) Using the method used in (a), give an upper bound on the following hypothesis set

\[ X = \mathbb{R}^2, H = \{ h_{a,r} = \left[ (x_1 - a_1)^2 + (x_2 - a_2)^2 < r^2 \right] = a \in \mathbb{R}^2, r \in \mathbb{R} \} \]

3 Estimation Error and Approximation Error

ABG Marketing is attempting to train a predictor that would predict, based on the text of a product review, whether it is positive or negative. For this purpose they collected 200,000 reviews (which we will assume were independently and randomly drawn from all reviews of interest) for which they also have labels indicating whether the review is positive or not. Avi, an employee, learns a linear predictor over word counts. That is, he uses the feature representation \( \phi(x) \in \mathbb{R}^{1036} \), where \( \phi_A(x)[i] \) is the number of appearance of word \( i \), for the 1036 distinct words that appeared in the reviews. He splits the data randomly into \( S_1 \) of size 150,000 and \( S_2 \) of size 50,000, and learns a predictor of the form \( \text{sign}(\langle w, \phi_A(x) \rangle) \) by minimizing the 0/1 loss over \( S_1 \): \( \hat{w}_A = \arg\min_w L_{S_1}(\text{sign}(\langle w, \phi_A(x) \rangle)) \). (Ignore computational complexity problems at this point). Avi found that \( L_{S_2}(\hat{w}_A) = 0.29 \). This was not good enough for the company.

Benny, another employee, instead suggests using as features counts of three-word-phrases. That is, using \( \phi_B(x) \in \mathbb{R}^{99,029} \), where \( \phi_B(x)[i] \) is the number of appearances in review \( x \) of one of 99,029 distinct phrases of up to three consecutive words (e.g. “is not good” or “best HW ever”). He repeated the same protocol as Avi, learning \( \hat{w}_B = \arg\min_w L_{S_1}(\text{sign}(\langle w, \phi_B(x) \rangle)) \) and found that \( L_{S_2}(\hat{w}_B) = 0.55 \)

(a) Was the increase in error mostly due to an increase in the estimation error or in the approximation error? Explain why.

(b) For each of the following suggestions, might it reasonably help Benny improve over Avi’s performance? Why?

(i) Use only phrases of up to two words.

(ii) Use four-word phrases as well.

(iii) Use a significantly larger training set.

4 Learning an Equivalence Relation

Let \( V \) be some set of size \( n \). Denote an instance space

\[ \mathcal{X} := \{(u, v) : u, v \in V\} = V \times V. \]

(In word, \( \mathcal{X} \) is the set of pairs of \( V \).) Define \( \mathcal{H}^0 \) to be the concept class consisting of all binary functions \( h : \mathcal{X} \mapsto \{0, 1\} \) such that

\[ \forall u \neq v : h((u, v)) = h((v, u)) \]

\[ \forall u : h((u, u)) = 1 \]

\[ \forall \text{distinct } u, v, y \in V : \text{if } h((u, v)) = 1 \text{ and } h((v, y)) = 1 \text{ then } h((u, y)) = 1. \]
A function satisfying (4.1) is called symmetric. If it satisfies (4.2) it is called reflexive. If it satisfies (4.3) then it is called transitive. If a binary relation is symmetric, reflexive and transitive then it is an equivalence relation. Given an equivalence relation \( h \in H^n \), then we can uniquely (up to reordering) partition of \( V \) into disjoint sets \( V_1, \ldots, V_{k=\#(h)} \) such that \( h((u, v)) = 1 \) for all \( u, v \) in the same \( V_i \), and \( h((u, v)) = 0 \) whenever \( u \in V_i, v \in V_j \) for \( i \neq j \). The sets \( V_i \) are called equivalence classes (induced by \( h \)).

(a) Give the best upper bound you can on the cardinality of \( H^n \).

(b) Establish the VC-dimension of \( H^n \) exactly, by showing how to shatter some \( D \) points and that no \( D + 1 \) points can be shattered. (Hint: for the upper bound, think of cycles and for the lower bound, use a tree.)

(c) Consider the class \( H^n_k \), defined as
\[
H^n_k := \{ h \in H^n : h \text{ induces at most } k \text{ equivalence classes.} \}
\]
Is it true that \( VCDim(H^n_{\log(n)}) = o(VCDim(H^n)) \)?

5 VC-dimension of Sparse Linear Classifiers Using Concept Class Union

Arguments

Fix an instance space \( X \). Let \( \{ H_i \} \) be a family of concept classes over \( X \), where \( i \) ranges from 1 to some integer \( r \). Define the concept \( H \) to be the union \( \cup_{i=1}^{r} H_i \).

(a) Give an upper bound for the growth function \( \tau_m(H) \) in terms of the growth functions \( \tau_m(H_i) \) for \( i = 1 \ldots r \).

(b) If the VC-dimension of all \( H_i \) are bounded by \( D \), i.e. \( VCdim(H_i) \leq D \), give a bound on the growth function \( \tau_m(H) \) in terms of \( m, r \) and \( D \) (hint: use Sauer’s lemma, which we saw in lecture 2 and appears in section 6.5.1 of the course book).

(c) From (b) conclude that \( VCdim(H) = O(\max(D, \log(r) + D \log(\log(r)/D))) \). Try to give an exact bound, without the \( O(\cdot) \) notation.

For the remainder or the question, consider a feature mapping \( \phi : X \to \mathbb{R}^d \). For a vector \( w \in \mathbb{R}^d \), we say that \( w \) is \( k \)-sparse if the number of coordinates \( j \) for which \( w(j) \neq 0 \) is at most \( k \). We define the following concept class:
\[
H^{(k)} := \{ h_w(x) = \text{sign}(\langle w, \phi(x) \rangle) \mid \text{s.t. } w \text{ is } k\text{-sparse} \}
\]

(d) Using (c) above, show that the VC-dimension of \( H^{(k)} \) is at most \( O(k \log(d/k)) \).

(e) Show how to shatter a subset of size \( \Omega(\log d) \) with respect to \( H^{(1)} \), establishing tight upper and lower bounds on the VC-dimension of \( H^{(1)} \).

(f*) Show how to shatter a subset of size \( \Omega(k \log(d/k)) \) with respect to \( H^{(k)} \), establishing tight upper and lower bounds on the VC-dimension of \( H^{(k)} \).
6 ERM for Intervals (Taken from exercise 1 from section 8.6 in the course book)

Let $\mathcal{X} = \mathbb{R}$, and define the concept class of intervals to be

$$\mathcal{H} := \{h \in \{0, 1\}^\mathcal{X} | \exists a \leq b \text{ s.t. } h(x) = 1 \iff a \leq x \leq b\}.$$ 

Propose an implementation of the ERM$_\mathcal{H}$ rule in the agnostic case which, given a sample of size $m$, runs in time $O(m^2)$. \textit{Hint: Use dynamic programming.}

\textbf{Good Luck!}