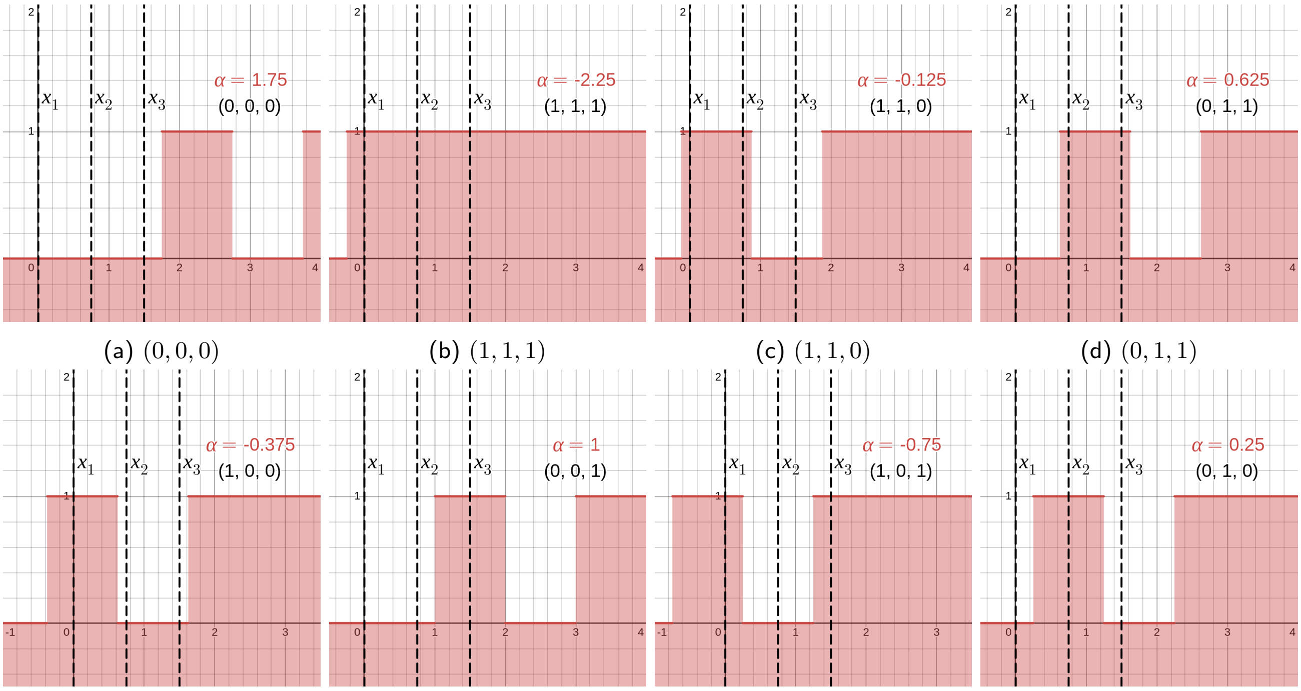
Theory of Machine Learning: Homework 5

1. Let $I_\alpha = [\alpha, \alpha+1] \cup [\alpha+2, \infty)$, and $h_\alpha(x) = 1_{I_\alpha}(x)$ be the indicator function associated with this set. Let $\mathcal{H}$ be the set of all h_α. We want to find the VC dimension of $\mathcal{H}$.

   Proof. We claim that $\mathrm{VCdim}(\mathcal{H}) = 3$.

   Step 1: $\mathrm{VCdim}(\mathcal{H}) \geq 3$

   Let $C = \{0,\ \frac{3}{4},\ \frac{3}{2}\}$, which is shattered. For a set of size 3, there are 8 dichotomies which can all be realized:



   (a) $(0, 0, 0)$ (b) $(1, 1, 1)$ (c) $(1, 1, 0)$ (d) $(0, 1, 1)$

   (e) $(1, 0, 0)$ (f) $(0, 0, 1)$ (g) $(1, 0, 1)$ (h) $(0, 1, 0)$

   Step 2: $\mathrm{VCdim}(\mathcal{H}) < 4$

   Let $A = [\alpha, \alpha+1]$ and $B = [\alpha+2, \infty)$. For a set of 4 samples, consider the dichotomy $(1, 0, 1, 0)$. WLOG, denote the 4 samples as $x_1 \leq x_2 \leq x_3 \leq x_4$. Both positive samples must lie in $A \cup B$. Suppose both samples belonged to the same disjoint subset, either A or B. Then the ordering $x_1 \leq x_2 \leq x_3$ requires that x_2 belongs to either A or B as well, in which case x_2 is labeled as positive. So it must be that x_1 and x_3 do not belong to the same disjoint subset. The assumption of ordering and the fact that every element in A is strictly less than every element in B requires that $x_1 \in A$ and $x_3 \in B$.

   Then $x_3 \leq x_4 \implies x_4 \in B = [\alpha+2, \infty)$, meaning the 4th sample could never be labeled as negative. So this dichotomy is not realizable for any choice of 4 reals. QED

2. Let $\mathcal{X}$ have cardinality n. Fix $k \in [0, n]$. Let $\mathcal{H}$ consist of all binary functions on $\mathcal{X}$ that assign exactly k samples positive labels. We want to find the VC dimension of $\mathcal{H}$.

   Proof. We claim that $\mathrm{VCdim}(\mathcal{H}) = \min\{k,\ n-k\}$.

   Step 1: $\mathrm{VCdim}(\mathcal{H}) \geq \min\{k,\ n-k\}$

   Choose any $C \subset \mathcal{X}$ such that $|C| = m = \min\{k,\ n-k\}$. Denote the count of positively-labeled samples in C by p. By construction, $p \leq m \leq k$. So we can allocate p positive labels for all appropriate samples in C. We must choose $k-p$ additional samples from $\mathcal{X} \setminus C$ to label as positive. We know $|\mathcal{X} \setminus C| = n - m$.

$$m \leq n-k \implies n-m \geq k \implies n-m \geq k-p$$

   So, we can always select $k-p$ additional points outside C to make the total number of positive samples k. So every labeling of C corresponds with some $h \in \mathcal{H}$, shattering C.

   Step 2: $\mathrm{VCdim}(\mathcal{H}) < \min\{k,\ n-k\} + 1$

Now, suppose $|C| = m + 1$. Recall that $m = \min\{k,\, n - k\}$.

- **Case 1.** If $m = k$, then $|C| = k + 1$. In this case, an entirely positive labeling is not realizable because $p = k + 1$ and any $h \in \mathcal{H}$ can only assign positive labels to k samples exactly.
- **Case 2.** If $m = n - k$, then $|C| = n - k + 1$. If all samples are labeled as negative, then we require that k samples in $\mathcal{X} \setminus C$ are labeled as positive. We know that

  $$|\mathcal{X} \setminus C| = n - |C| = n - (n - k + 1) = k - 1$$

  So, only $k - 1$ samples exist outside C, requiring that at least one sample in C is labeled as positive. An entirely negative labeling is not realizable for any $h \in \mathcal{H}$.

Regardless of m, there is a labeling for any C that is not realizable under any $h \in \mathcal{H}$. QED

3. $\boxed{\text{Reading}}$ Part of the paper discusses encoding an entire dataset into a single real number α by discretizing each sample and appending that binary representation to a growing decimal in $[0, 1]$. I think it is not too remarkable that this is possible because real numbers can trail infinitely, so conceptually you could encode any information within one with arbitrary precision (though not efficiently). The fact that this real parameter could be decoded with a closed-form function was much more surpising and pretty counterintuitive. The function is evidently not useful as a predictive model because from a complexity standpoint, it can shatter any dataset, so we cannot really expect meaningful generalization from it. As the author discusses at the end of the paper, this undermines parameter counting as a way of “eyeballing” complexity, as the provided solution seems to be infinitely complex with a single parameter. It also calls learning into question for over-parameterized deep networks because you could probably draw functional similarities between the roles of α and the millions of floating point weights in a deep network. The first reading showed failure of UC in deep networks and this paper seems to cast generalization in deep learning as something truly unexplainable.