2022-05-25 Wednesday

NAME: Ravi Moelchand - 7463332

Topics covered in this tutorial:

- Shattering and VC Dimension (Lecture 6)
- The VC Generalization Bound (Lecture 7)

If you write your answers directly into the notebook, it is preferred that you generate a .pdf file for submission

This tutorial contains 9 problems. Please submit one solution per person.

Shattering and VC Dimension

1. Given is a hypothesis set $\mathcal{H}$ and a particular data set $\mathcal{D}$ with N points. Given is that $\mathcal{H}$ cannot shatter $\mathcal{D}$. Is the following statement true or false? Explain.

It is certain that for N, the growth function is less than 2^N. In other words: $m_{\mathcal{H}}(N) < 2^N$

answer

True, the number of dichotomies is at most 2^N. If $\mathcal{H}$ cannot shatter $\mathcal{D}$, then the maximum number of dichotomies that you can get will be smaller than 2^N. So, the growth function will be smaller than 2^N.

---

2. Suppose that hypothesis set $\mathcal{H}$ can shatter a dataset $\mathcal{D}$ with N points. Is the following statement true or false? Explain.

It is possible that for some values $M < N$, the growth-function has a value that is less than 2^M. In other words: it is possible that there is an $M < N$ such that $m_{\mathcal{H}}(M) < 2^M$.

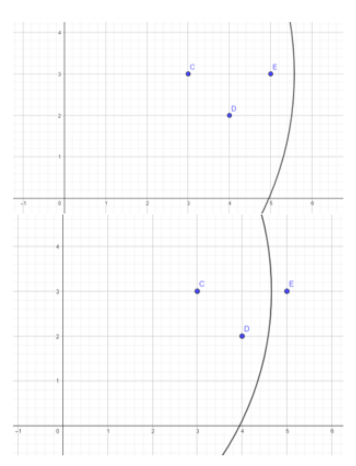
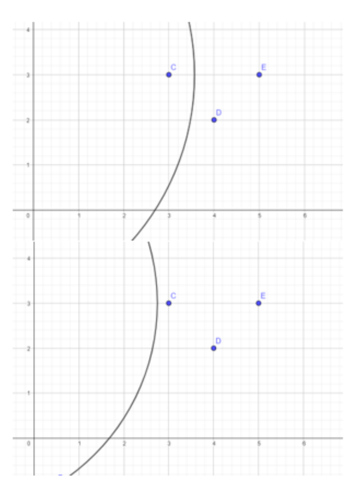
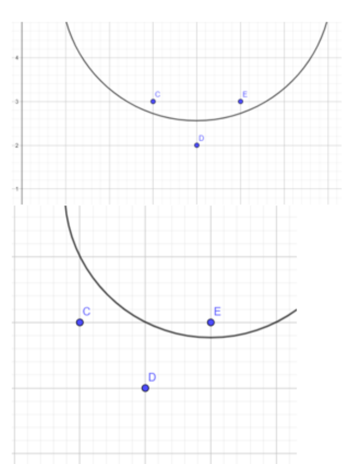
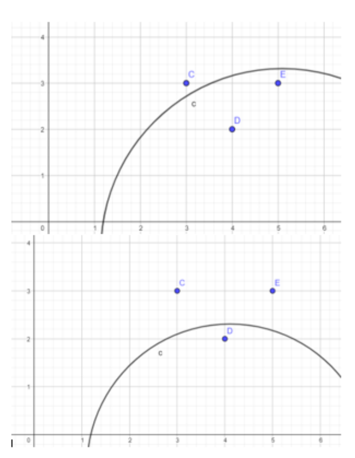
answer

True, for a certain set of M datapoints $\mathcal{H}$ shatters $\mathcal{D}$, but this does not mean that there cannot be another set of M points for which $\mathcal{H}$ does not shatter $\mathcal{D}$.

Given is the following hypothesis set $\mathcal{H}_{disk}$, with a two-dimensional input space, so $\mathcal{X} = \mathbb{R}^2$. Each $h \in \mathcal{H}_{disk}$ is a region in the form of a closed disk (closed means that the region includes the circle boundary of the disk). Also see https://en.wikipedia.org/wiki/Disk_(mathematics). A disk-region may have any diameter. Everything in the disk-region is classified as $+1$, the rest as -1.

3. Show that there is a data set with $N = 3$ that can be shattered by $\mathcal{H}_{disk}$. Show this visually in a graph, by strategically choosing a data set and drawing in that same graph $2^3 = 8$ strategically chosen circles that represent the disk-regions, which each produce another dichotomy. If the graph becomes too crowded, instead, create multiple graphs with the same data set, and in each draw a few of the circles. Also, for large circles, you can suffice with drawing a part of the circle, if it is clear to the viewer how to extend it to a full circle (also see https://en.wikipedia.org/wiki/Circular_arc).

answer: *If the point is inside the circle it is considered '+', otherwise '-'.*

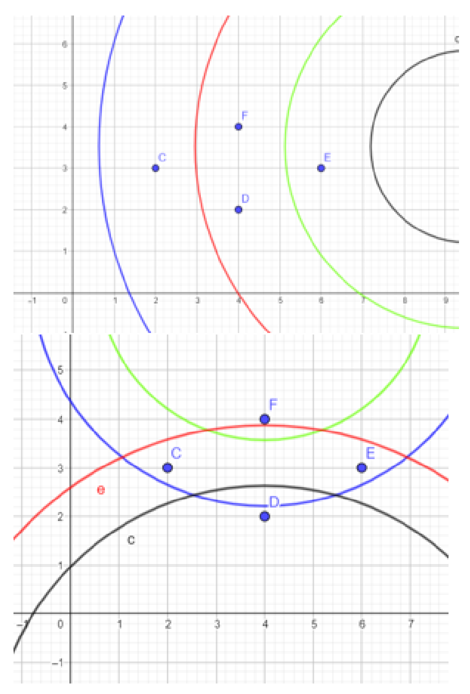
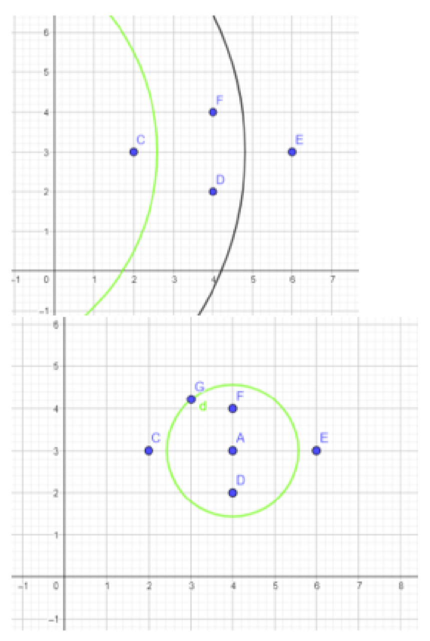
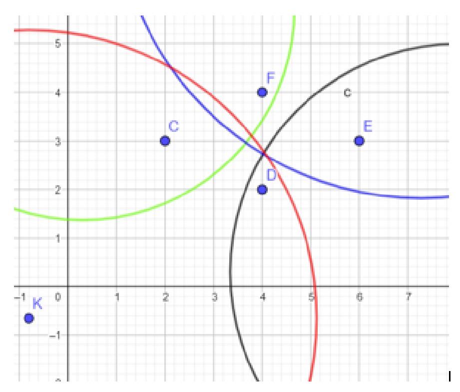
4. What is the value of the growth function for $N = 3$? So, what is $m_{\mathcal{H}_{disk}}(3)$?

answer

Since it can be shattered by $\mathcal{H}_{disk}$, $m_{\mathcal{H}_{disk}}(3) = 2^3 = 8$.

5. Determine the highest lower boundary of the value of the growth-function in $N = 4$ you can come up with. In other words, what is the maximal amount of dichotomies you are able to create for a strategically chosen data set with $N = 4$? Do this in the same way as one of the above questions, and show your graphs.

answer: *If the point is inside the circle it is considered '+', otherwise '-'. Using this dataset I could create 15 dichotomies, which is the most I could find. I could not think of a dataset that could generate all 16 possibilities.*

---

6. What is the VC-dimension of a hypothesis set $\mathcal{H}$ with growth function $m_{\mathcal{H}}(1) = 2, m_{\mathcal{H}}(2) = 4, m_{\mathcal{H}}(3) = 8, m_{\mathcal{H}}(4) = 16$, and $m_{\mathcal{H}}(N) = N^2 + 7$ for $N > 4$?

answer

The VC-dimension is the largest amount of points N such that the growth function is 2^N, which in this case is $N = 4$ since $m_{\mathcal{H}}(4) = 16$ and $2^4 = 16$.

Which of the following functions cannot be a growth-function of some hypothesis set? Which of them could be a growth-function? Explain. Tip: use the properties from Learning From Data (Abu-Mostafa et al, 2012), Section 2.1.3.

7. $f(N) = N + 1$.

8. $f(N) = N + \lfloor (1\frac{1}{2})^N \rfloor$. (The symbols indicate the floor.)

answer

7 cannot be a growth function, because $N + 1$ would only be possible if $m_{\mathcal{H}}(N) \leq N^{d_{VC}} + 1$. That would only be true if d_{VC} would be 1 for which N should be 0. But for $N = 0$ the VC dimension is infinite, which means that the rule, $m_{\mathcal{H}}(N) \leq N^{d_{VC}} + 1$, cannot be applied. Therefore, 7 is not a growth function.

8 can be a growth function with a value $\leq 2^N + N$.

The VC Generalization Bound

9. For a hypothesis set with $d_{\mathrm{VC}} = 12$, what sample size do you need (as prescribed by the generalisation bound) to have a 90% confidence that your generalization error is at most 0.03? (A variant on Problem 2.12 from Learning From Data (Abu-Mostafa et al, 2012)).

answer

$\sqrt{\frac{8}{N} \cdot ln(\frac{4(2N)^{12}+1}{0.9})} \leq 0.03$