Big Data
Exercises Infinite Hypothesis Classes

Introduction to infinite hypothesis classes

In the results we have derived so far, the size of the hypothesis class always played an important role. For example, to determine what would be a sufficient number of training examples to meet some (ϵ, δ) requirements, $|\mathcal{H}|$ always appeared prominently in the formula. Such results become useless when $\mathcal{H}$ has infinite size. In most realistic machine learning problems $\mathcal{H}$ has (at least conceptually) infinite size, so we would really like to obtain similar results for infinite hypothesis classes. We have seen that there are infinite hypothesis classes that can be learned (e.g. the threshold function) and there are infinite hypothesis classes thet cannot be learned (e.g. the set of all functions $f : X \to \{0, 1\}$ where the domain of X is infinite). So how can we distinguish the infinite classes that *are* learnable from the infinite classes that are not learnable?

As it turns out, we can characterize the complexity of infinite hypothesis classes by a number called the VC-dimension. This number will play a similar role for infinite hypothesis classes as the size of $\mathcal{H}$ did in the results for finite hypothesis classes.

The VC-dimension of a hypothesis class $\mathcal{H}$ is the size of the largest data set for which $\mathcal{H}$ can guarantee zero training error *for any assignment of class labels to the data points in that set.* When $\mathcal{H}$ can produce all possible 2^m ways to assign class labels to a set of m data points, we say that $\mathcal{H}$ *shatters* this set of data points. It is important to note that if there is *any* set of m data points that $\mathcal{H}$ can shatter, then the VC-dimension of $\mathcal{H}$ is at least m. It is not required that $\mathcal{H}$ shatters *all* sets of m data points! In other words, you may choose the set of points in the most favourable way so that $\mathcal{H}$ can shatter them.

We'll call an assignment of class labels to a set of data points a *dichotomy.* For a given set of data points, many different members of $\mathcal{H}$ will produce the same dichotomy because we only look at how h labels the set of data points concerned (as opposed to how h labels the entire space of all possible data points).

Before we start with the exercises, we make one final definition. The growth function $\tau_{\mathcal{H}}(m)$ of $\mathcal{H}$ is the maximum number of dichotomies that $\mathcal{H}$ can produce on a set of m points. Again, you get to choose the m data points so as to maximize the number of dichotomies. So $\tau_{\mathcal{H}}(m)$ is the maximum over all data sets of size m of the number of dichotomies that $\mathcal{H}$ can produce on this set. If for some number a, $\tau_{\mathcal{H}}(a) = 2^a$, then there is a set of a points that is shattered by $\mathcal{H}$, so $\mathcal{H}$ has VC-dimension at least a (at least, because there might be a bigger set that is also shattered by $\mathcal{H}$).

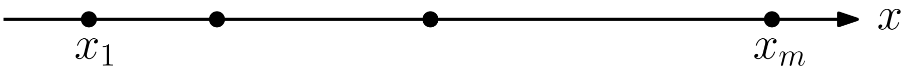
Exercise 1: Threshold functions

Consider the class $\mathcal{H}_t$ of threshold functions:

$$h(x) = \begin{cases} +1 & \text{if } x \geq t \\ -1 & \text{if } x < t, \end{cases}$$

where $x, t \in \mathbb{R}$. Notice that this is an infinite hypothesis class, since t can be any real number.

(a) Consider a set of m distinct data points $x_1, x_2, \ldots, x_m$. How many dichotomies, as a function of m, can $\mathcal{H}_t$ produce on such a set? (the answer to this question is what we call the growth function for this hypothesis class).
The following picture may be helpful:



(b) Compute the value of the growth function of $\mathcal{H}_t$ for $m = 1, 2, 3$.

(c) What is the VC-dimension of $\mathcal{H}_t$?

Exercise 2: Intervals

Consider the class $\mathcal{H}_{(t_1,t_2)}$ of intervals:

$$h(x) = \begin{cases} +1 & \text{if } t_1 \leq x \leq t_2 \\ -1 & \text{otherwise }, \end{cases}$$

where $x, t_1, t_2 \in \mathbb{R}$.

(a) Consider a set of m distinct data points $x_1, x_2, \ldots, x_m$. How many dichotomies can $\mathcal{H}_{(t_1,t_2)}$ produce on such a set?

(b) Compute the value of the growth function of $\mathcal{H}_{(t_1,t_2)}$ for $m = 1, 2, 3$.

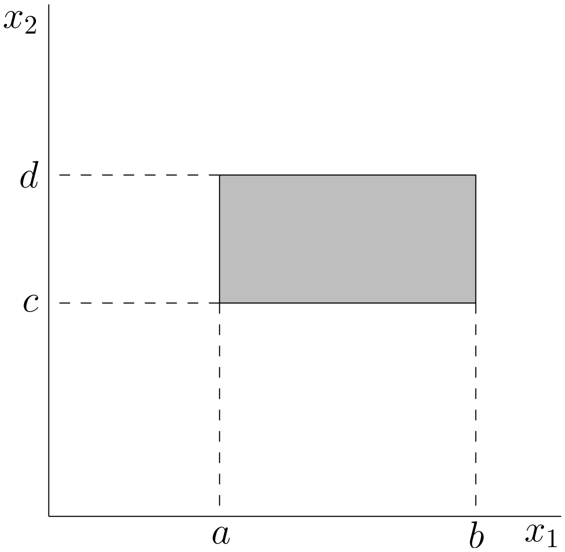
(c) What is the VC-dimension of $\mathcal{H}_{(t_1,t_2)}$?

Exercise 3: Axis-aligned rectangles

This question is about the VC-dimension of axis-aligned rectangles in $\mathbb{R}^2$.
For any $a, b, c, d \in \mathbb{R}$, let

$$h(x_1, x_2) = \begin{cases} +1 & \text{if } a \leq x_1 \leq b \text{ and } c \leq x_2 \leq d \\ -1 & \text{otherwise} \end{cases}$$

In a picture:



Let $\mathcal{H}_{rect}$ denote the class of all such axis-aligned rectangles.
Show that the VC-dimension of $\mathcal{H}_{rect}$ is 4, by showing that

(a) There is a set of 4 points that is shattered by $\mathcal{H}_{rect}$.

(b) No set of 5 points is shattered by $\mathcal{H}_{rect}$.

Exercise 4: Bounding the growth function

By Sauer's Lemma, we know that

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}, \tag{1}$$

where d is the VC-dimension of $\mathcal{H}$. Prove by induction that

$$\sum_{i=0}^{k} \binom{m}{i} \leq m^k + 1,$$

and hence

$$\tau_{\mathcal{H}}(m) \leq m^d + 1. \tag{2}$$

Exercise 5: Application of the bound on the growth function

(a) We have derived the growth function for threshold functions from “first principles” in exercise 1. Compare this bound to the bound provided by Sauer’s Lemma (equation 1). Is the bound provided by Sauer’s Lemma tight in this case?

(b) Repeat (a) for the hypothesis class of intervals (exercise 2).

(c) We have determined that the VC-dimension of axis-aligned rectangles in $\mathbb{R}^2$ is 4. Use this in combination with Sauer’s Lemma to bound the number of dichotomies that this hypothesis class can realize on $m = 10$ respectively $m = 20$ data points. Compare these numbers to the total number of dichotomies possible for $m = 10$, respectively $m = 20$.

Introduction to Linear Classifiers

A linear classifier in $\mathbb{R}^d$ is a classifier of the form:

$$h(x_1, x_2, \ldots, x_d) = \begin{cases} +1 & \text{if } w_0 + \sum_{j=1}^{d} w_j x_j \geq 0 \\ -1 & \text{if } w_0 + \sum_{j=1}^{d} w_j x_j < 0 \end{cases}$$

The $d+1$ coefficients (or weights) w_j ($j = 0, 1, \ldots, d$) can have any real number as a value. In $\mathbb{R}^2$ the decision boundary $w_0 + w_1x_1 + w_2x_2 = 0$ is a line. In $\mathbb{R}^3$ the decision boundary $w_0 + w_1x_1 + w_2x_2 + w_3x_3 = 0$ is a plane. The weight w_0 is often called the bias (not to be confused with inductive bias); if $w_0 = 0$ the decision boundary passes through the origin.

To give a concrete example, consider a linear classifier for credit approval. Let x_1 represent the income of the applicant, and let x_2 represent the age of the applicant. Furthermore, let $y = +1$ if the applicant is accepted and $y = -1$ if the applicant is rejected. For this concrete example, the linear classifier is:

$$h(\text{income}, \text{age}) = \begin{cases} \text{accept} & \text{if } w_0 + w_1 \text{ income} + w_2 \text{ age} \geq 0 \\ \text{reject} & \text{if } w_0 + w_1 \text{ income} + w_2 \text{ age} < 0 \end{cases}$$

For brevity we switch to vector notation. Let $\mathbf{x} = (x_0, x_1, \ldots, x_d)$, where $x_0 \equiv 1$, and $\mathbf{w} = (w_0, w_1, \ldots, w_d)$. Now we can conveniently write:

$$\mathbf{w} \cdot \mathbf{x} = w_0 + \sum_{j=1}^{d} w_j x_j$$

A more compact way to define the class of linear classifiers now is: $h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x})$.

Exercise 6: Linear classifiers with a single predictor variable

A linear classifier in $\mathbb{R}^1$ is a classifier of the form:

$$h(x) = \begin{cases} +1 & \text{if } w_0 + w_1 x \geq 0 \\ -1 & \text{if } w_0 + w_1 x < 0 \end{cases}$$

Consider an example where we want to predict whether someone is able to complete a programming assignment in time, based on the number of months of programming experience of this person. We have collected the following 25 examples:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
x	4	4	5	6	6	8	9	11	12	13	13	14	18	18	19	21	22	22	24	25	28	29	30	30	32
y	-1	-1	-1	-1	-1	$+1$	-1	-1	-1	-1	$+1$	-1	$+1$	-1	-1	$+1$	$+1$	$+1$	-1	$+1$	$+1$	-1	$+1$	$+1$	$+1$

Here, the x variable is *experience*: the number of months of programming experience of a person. The variable *success* is the y variable (the class label), and has the value $+1$ if the programming assignment was completed in time, and the value -1 if the assignment was not completed in time. Note that i is just an index for the training examples, it is not part of the data. For example, training example 1 ($i = 1$) represents a person who has 4 months of programming experience who did not finish the assignment in time. There are 25 training examples in total, and 11 programming assignments were completed in time.

(a) Determine a threshold value for *experience* that minimizes the empirical risk of the threshold function on this data set. What is the value of the training error? Is the ERM solution unique?

(b) Give values for w_0 and w_1 that minimize the empirical risk of the linear classifier on this data set. What is the value of the training error? Is the ERM solution unique?

(c) The threshold classifier requires one parameter less than the linear classifier (one threshold versus two weights), but it seems to have the same expressive power! Is this indeed the case? (Hint: consider the above data set, but with the labels inverted: $+1$ becomes -1 and vice versa).

(d) Determine the growth function and VC-dimension of the linear classifier with one predictor variable.

Exercise 7: The VC-dimension of linear classifiers

Let's denote the class of linear classifiers in $\mathbb{R}^d$ as $\mathcal{H}^d_{\text{lin}}$. Let's start with $\mathbb{R}^2$. We'll show that the VC-dimension is at least 3, by giving the following set of 3 points that is shattered:

	x_0	x_1	x_2
$\mathbf{x}_0$	1	0	0
$\mathbf{x}_1$	1	1	0
$\mathbf{x}_2$	1	0	1

The difference between boldface (vectors) and normal font now becomes important. For example $\mathbf{x}_0$ is the first row of the data table, that is, the first data point (excluding the class label). We number the data points beginning at zero just for convenience, as will become clear later. Formally, $\mathbf{x}_0 = (x_{0,0}, x_{0,1}, x_{0,2}) = (1, 0, 0)$ is a vector with three components. The normal face x_0 is the variable x_0 that always has the value 1 (remember it is just there to accommodate the bias weight w_0). Likewise, the normal face x_1 could stand for income, and x_2 could stand for age in the credit approval example.

(a) To show that the given set of three points (rows) is shattered, we must show that we can obtain any label assignment to them by choosing appropriate values for the weights. We'll even be a bit more demanding: show that we can obtain

$$\begin{aligned} y_0 &= \mathbf{w} \cdot \mathbf{x}_0 \\ y_1 &= \mathbf{w} \cdot \mathbf{x}_1 \\ y_2 &= \mathbf{w} \cdot \mathbf{x}_2 \end{aligned}$$

for any label assignment y_0, y_1, y_2 ($y_i \in \{-1, +1\}$) by choosing appropriate weight values w_0, w_1, w_2.

Plot the three data points in the (x_1, x_2)-plane, and draw in the decision boundary (line) for a few weight vectors corresponding to different label assignments.

(b) Generalize the result found under (a), by giving a rule for how to choose the weight values $w_0, w_1, \ldots, w_d$ depending on the labeling that has to be produced on $d + 1$ data points. Choose the data points in a similar way as was done under (a).

(c) (Hard) Show that no set of $d + 2$ data points in $\mathbb{R}^d$ is shattered by $\mathcal{H}^d_{\text{lin}}$.

Hint: you need to use the result from linear algebra that any set of $n + 1$ vectors in $\mathbb{R}^n$ is linearly dependent. Therefore, we can write one of the rows (say we pick row j) as a linear combination of the other rows:

$$\mathbf{x}_j = \sum_{i \neq j} c_i \mathbf{x}_i,$$

where not all c_i are zero because the first component of every vector has the value 1 ($x_0 \equiv 1$).

Note: we actually have $d + 2$ vectors in $\mathbb{R}^{d+1}$ (not $\mathbb{R}^d$) since we added x_0.
We follow the standard terminology however, which can be a bit confusing.