Big Data 2019
The VC dimension and frequent item sets

Introduction

We have seen that for finite hypothesis classes,

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\exists h\in\mathcal{H}:\left|L_{\mathcal{D}}(h)-L_D(h)\right|>\epsilon\right]\leq 2|\mathcal{H}|e^{-2\epsilon^2 m} \tag{1}$$

by the union bound and Hoeffding's inequality.

To adapt this result to the case of infinite hypothesis classes, one could loosely argue as follows. We can replace $|\mathcal{H}|$ by the number of dichotomies that $\mathcal{H}$ can realize on a data set of size m. All hypotheses that display the same behaviour on the training sample are counted as one hypothesis only. Hence, we can just replace $|\mathcal{H}|$ by $\tau_{\mathcal{H}}(m)$ in equation (1):

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\exists h\in\mathcal{H}:\left|L_{\mathcal{D}}(h)-L_D(h)\right|>\epsilon\right]\leq 2\tau_{\mathcal{H}}(m)e^{-2\epsilon^2 m}. \tag{2}$$

This argument is not quite correct, but in a qualitative sense it leads to the correct conclusion. The correct inequality is:

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\exists h\in\mathcal{H}:\left|L_{\mathcal{D}}(h)-L_D(h)\right|>\epsilon\right]\leq 4\tau_{\mathcal{H}}(2m)e^{-\frac{1}{8}\epsilon^2 m}. \tag{3}$$

If $\mathcal{H}$ has finite VC-dimension d, then the growth function $\tau_{\mathcal{H}}(m)$ is bounded above by m^d (see the previous exercise set). Filling in this upper bound for $\tau_{\mathcal{H}}(2m)$ in equation 3, we obtain:

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\exists h\in\mathcal{H}:\left|L_{\mathcal{D}}(h)-L_D(h)\right|>\epsilon\right]\leq \frac{4(2m)^d}{e^{\frac{1}{8}\epsilon^2 m}}.$$

As $m\to\infty$, the polynomial m^d gets annihilated by the exponential e^m, and so the probability of error goes to zero. If $\mathcal{H}$ does not have finite VC-dimension, then $\tau_{\mathcal{H}}(m)=2^m$. Filling this in in equation 3 gives:

$$\mathbb{P}\leq\frac{2^{2m+2}}{e^{\frac{1}{8}\epsilon^2 m}}.$$

This ratio does not go to zero as m goes to infinity.

Exercise 1: The VC generalization bound

Starting from the VC-inequality:

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\exists h\in\mathcal{H}:\left|L_{\mathcal{D}}(h)-L_D(h)\right|>\epsilon\right]\leq 4\tau_{\mathcal{H}}(2m)e^{-\frac{1}{8}\epsilon^2 m} \tag{4}$$

we can reverse the statement

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\forall h\in\mathcal{H}:\left|L_{\mathcal{D}}(h)-L_D(h)\right|\leq\epsilon\right]\geq 1-4\tau_{\mathcal{H}}(2m)e^{-\frac{1}{8}\epsilon^2 m} \tag{5}$$

Hence, in particular we have

$$\mathbb{P}_{D\sim\mathbb{P}(X,Y)^m}\left[\left|L_{\mathcal{D}}(h_D)-L_D(h_D)\right|\leq\epsilon\right]\geq 1-4\tau_{\mathcal{H}}(2m)e^{-\frac{1}{8}\epsilon^2 m} \tag{6}$$

where h_D denotes a hypothesis returned by an ERM algorithm.

Set $\delta = 4\tau_{\mathcal{H}}(2m)e^{-\frac{1}{8}\epsilon^2 m}$, and solve for ϵ to show that with probability $\geq (1-\delta)$:

$$L_{\mathcal{D}}(h_D)\leq L_D(h_D)+\sqrt{\frac{8}{m}\ln\left(\frac{4\tau_{\mathcal{H}}(2m)}{\delta}\right)}$$

Analyzing this expression, what can you say about how close $L_D(h_D)$ gets to $L_{\mathcal{D}}(h_D)$, as $m\to\infty$, for hypothesis classes with finite VC-dimension?

Exercise 2: d-index and d-bound

Consider the following database with 9 transactions:

tid	Items
1	ABE
2	BD
3	BC
4	ABD
5	AC
6	BC
7	AC
8	ABCE
9	ABC

Here A,B, etc. denote single items. For example, in the third transaction the items B and C were bought. tid denotes the transaction id.

(a) Compute the d-index and the d-bound for the given transaction database.

(b) Repeat, but now without the last transaction.

(c) Suppose that, just to increase its size, we replicate a transaction database a number of times. For example, we copy each transaction 10 times. Does the d-index increase, decrease, or stay the same?

(d) Consider a transaction database D with $1,000,000$ transactions on 500 items. The d-bound of D is 80. Let $\epsilon = 0.05$ and $\delta = 0.05$. How many transactions should we sample to obtain an ϵ-close approximation with probability at least $1 - \delta$?

(e) Repeat with $\epsilon = 0.01$.

Exercise 3: An even looser upper bound

(a) Suppose we also think it is too much trouble to check if two transactions are different. Define an upper bound for the VC-dimension that gets rid of this condition as well, and argue that it is indeed a valid upper bound.

(b) Suppose you have the following statistics about a supermarket transaction database (see the table on the next page). The column "size" indicates the number of items in a transaction, and the column "# transactions" contains the number of transactions in the database with the given size. For example, there are 6919 transactions that contain exactly 3 items. Give an upper bound (as small as possible) for the VC-dimension of this transaction database.

Exercise 4: Some intuition about the VC-dimension

By using the VC-dimension we try to improve over the union bound in computing the probability of a "bad event". In a sense, the VC-dimension measures the amount of overlap between bad events (recall that the union bound is valid even if the overlap is zero, that is, if the bad events are mutually exclusive). The bigger the overlap, the smaller the VC-dimension and vice versa.

Consider two transaction databases with the same number of transactions, the same number of items $|\mathcal{I}|$, and the same distribution over supports of single items. In one data base the items are independent (like shampoo and bread), and in the other they are correlated (like bread and ham). Which database do you expect to have a bigger VC-dimension?

size	# transactions	size	# transactions
1	3016	37	153
2	5516	38	123
3	6919	39	115
4	7210	40	112
5	6814	41	76
6	6163	42	66
7	5746	43	71
8	5143	44	60
9	4660	45	50
10	4086	46	44
11	3751	47	37
12	3285	48	37
13	2866	49	33
14	2620	50	22
15	2310	51	24
16	2115	52	21
17	1874	53	21
18	1645	54	10
19	1469	55	11
20	1290	56	10
21	1205	57	9
22	981	58	11
23	887	59	4
24	819	60	9
25	684	61	7
26	586	62	4
27	582	63	5
28	472	64	2
29	480	65	2
30	355	66	5
31	310	67	3
32	303	68	3
33	272	71	1
34	234	73	1
35	194	74	1
36	136	76	1