Machine Learning Problemset 3

Aimee Barciauskas

26 February 2016

Problem 12

Consider the class $\mathcal{A}$ of all sets of the form:

$$A_\alpha = \{x \in \mathbb{R} : sin(\alpha x) > 0\}$$

where $\alpha > 0$. What is the VC dimension of $\mathcal{A}$? (Note that $\mathcal{A}$ has one free parameter.)

The VC dimension of $\mathcal{A}$ is infinite.

For example, if $x_i = 2^{-i}, i = 1, ..., m$ are assigned arbitrary labels $(y_1, ..., y_m) \in \{-1, 1\}^m$, α may be chosen such that any label set is correctly classified:

$$\alpha = \pi(1 + \sum_{i=1}^{m} 2^i \frac{1 - y_i}{2})$$

Problem 13

Let $\mathcal{A}_1, ..., \mathcal{A}_k$ be classes of sets, all of the with VC dimension at most V. Show that the VC-dimension of $\cup_{i=1}^k \mathcal{A}_\rangle$ is at most $4V log_2(2V) + 4k$. You may use the fact that for $a \geq 1$ and $b > 0$, if $x \geq 4a log(2a) + 2b$ then $x \geq a log x + b$.

Can you bound the VC dimension of the class of all sets of the form:

$$A_1 \cup ... \cup A_k \text{ with } A_1 \in \mathcal{A}_1, ..., A_k \in \mathcal{A}_k$$

Part 1 Solution:

$$s_\mathcal{A} = 2^u \text{ such that } \alpha \text{ is the VC dimension of } \mathcal{A}$$

Sauer's Lemma:

$$s_\mathcal{A} \leq s_{\mathcal{A}_1} + s_{\mathcal{A}_2} + ... + s_{\mathcal{A}_k} \leq k(u + 1)^V$$

If there exists a u which satisfies:

$$s_\mathcal{A} = 2^u$$

$$2^u \leq k(u + 1)^V$$

$$u \leq log(k) + V log(u + 1)$$

$$= log(k) + Vlog(u+1)$$

$$log(k) \leq 2k$$

$$= 2k + Vlog(u+1)$$

Using $log(u+1) \approx log(u)$

$$u \leq 2k + Vlog(u)$$

Using the hints from the problem where $a = V$, $x = u$ and $b = 2k$ this can be re-written as:

$$u \leq 4Vlog(2V) + 4k$$

Part 2 Solution:

For the union of sets of $A_1 \cup ... \cup A_k$:

$$s_{\mathcal{A}}(n) \leq s_{\mathcal{A}_1}(n) \times s_{\mathcal{A}_1}(n) \times ... \times s_{\mathcal{A}_k} \leq (n+1)^{Vk}$$

$$s_{\mathcal{A}}(u) = 2^u \leq (u+1)^{Vk}$$

$$u \leq Vklog(u+1)$$

Let $a = Vk$, $b = \epsilon$, $x = u+1$, where ϵ is some small number:

$$u \leq 4Vklog(2Vk)$$

Problem 14

$$\|w_t - w_*\|^2 \leq \|w_{t-1} - w_*\|^2 - 1$$

We have $w_t = w_{t-1} + \frac{Y_t X_t}{\|X_t\|}$, so the first term can be re-written as:

$$\|w_{t-1} - w_* + \frac{Y_t X_t}{\|X_t\|}\|^2 = \|w_{t-1} - w_*\|^2 + \left(\frac{Y_t X_t}{\|X_t\|}\right)^2 + 2(w_{t-1} - w_*)\frac{Y_t X_t}{\|X_t\|}$$

$$= \|w_{t-1} - w_*\|^2 + \left(\frac{Y_t X_t}{\|X_t\|}\right)^2 + 2w_{t-1}\frac{Y_t X_t}{\|X_t\|} - 2w_*\frac{Y_t X_t}{\|X_t\|}$$

The first term above equals the first term in the RHS of our initial inequality to be proved.

The second term is 1, so we subtract it from the RHS and get an equality of the first two terms of the expanded LHS and RHS.

The last two terms formulate the inquality: We know the second to last term $2w_{t-1}\frac{Y_t X_t}{\|X_t\|} < 0$ when the percepton makes no more updates. and the last term $2w_*\frac{Y_t X_t}{\|X_t\|} \geq 1$. Something negative minus something

positive is negative, so the whole term is negative. Adding this negative term to the other side we get the inequality:

$$\|w_t - w_*\|^2 \le \|w_{t-1} - w_*\|^2 - 1$$

Using this inquality iteratively:

$$\|w_{t-1} - w_*\|^2 \le \|w_{t-2} - w_*\|^2 - 2$$
$$\|w_{t-2} - w_*\|^2 \le \|w_{t-3} - w_*\|^2 - 3$$
$$...$$
$$\|w_* - w_*\|^2 \le \|w_0 - w_*\|^2 - k$$

where k is the number of steps, and the LHS is now 0:

$$k \le \|w_0 - w_*\|^2$$

The number of steps is less than or equal to the $\|w_0 - w_*\|^2$.

Problem 15

Part 1

The expected risk of the data-dependent leave-one-out classifier:

$$\mathbb{E}\left\{R_n^D(g_n)\right\} = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{g_{n-1}(X_i, D_{n,i}) \ne Y_i}\right\}$$

The RHS is a random variable: it is the sum of the variations on the random data set being used to train the leave-one-out classifier. The law of iterated expectations shows this can be written as (where g_{n-1} is the leave-one-out classifier):

$$\mathbb{E}\left\{R_n^D(g_n)\right\} = \mathbb{E}\left\{\mathbb{E}\{R(g_{n-1})|D\}\right\} = \mathbb{E}\{R(g_{n-1})\}$$

Part 2

$$\mathbb{E}\left\{R_n^D(g_n)\right\} = \mathbb{E}\{R(g_{n-1})\}$$

If M is the number of mistakes made by the perceptron classifier, it is equivalent to the number of times the leave one out classifier also makes a mistake. Since we can estimate the expected risk using the leave-one-out classifier and the number of iterations made by the perceptron classifier is upper-bounded by $(\frac{R}{\gamma})^2$ (Novikoff, 1962), we can bound the expected risk of the perceptron classifier by $\frac{M}{n}$

Problem 16

Consider the majority classifier:

$$g_n(x, D_n) = \begin{cases} 1, & \text{if } \sum_{i=1}^n Y_i \geq \frac{n}{2} \\ 0, & \text{otherwise} \end{cases}$$

(Thus, g_n ignores x and the X_i's.) Assume that n is odd. What is the expected risk $\mathbb{E}R(g_n) = \mathbb{P}\{g_n(X) \neq Y\}$ of this classifier? Study the performance of the leave-one-out error estimate. Show that for some distributions $Var(R_n^D(g_n)) \geq c/\sqrt{n}$ for some constant c. _Hint:_ Strang things happen when the number of 0's and 1's is about the same in the data.

Solution Part 1:

Let N_n be the number of $Y_i = 0$ in the sample. So N_n is binomial $(n, 1 - p)$ with p the $P\{Y = 1\}$.

$$\mathbb{E}(R(g_n)) = p\mathbb{P}\left\{N_n \geq \frac{n}{2}\right\} + (1 - p)\mathbb{P}\left\{N_n < \frac{n}{2}\right\}$$

(asymptotically this is $min(p, 1 - p)$)

Solution Part 2:

Risk is binomial and so it's variance can be lower bounded by the expected value of the squared risk.

$$Var(R_n^D(g_n)) = \mathbb{E}\{(R_n^D)^2\} - \mathbb{E}\{R_n^D\}^2$$

$$\geq \mathbb{E}\{(R_n^D)^2\}$$

The value of the risk for which this expectation is maximized is the expected value of the risk $(R_n^D(g_n) \sim Bin(n, p))$ when $p = \frac{1}{2}$ and the variance can be lower bounded when the expectated value of the risk 1.

This is always greater than or equal to the probablity that just over half the data are 1's when we favor 0's.

Risk being equal to 1 when $\mathbb{P}\{Y = 1\} = \mathbb{P}\{Y = 0\} = \frac{1}{2}$ which happens when the data n is near as possible to (favoring the half which has realized number $\frac{n-1}{2}$) $\frac{1}{2}$ with probability:

$$\mathbb{E}\{(R_n^D)^2\} \geq \mathbb{P}\left\{Bin(n, \frac{1}{2}) = \frac{n+1}{2}\right\}$$

$$= \frac{1}{2}^n \binom{n}{\frac{n+1}{2}}$$

Using Stirling's approximation and that $n \approx n + 1$:

$$Var(R_n^D(g_n)) \geq \frac{\sqrt{\frac{2}{\pi}}}{\sqrt{n}}$$