Machine Learning Problemset 3

Aimee Barciauskas

29 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 $4Vlog_2(2V) + 4k$. You may use the fact that for $a \geq 1$ and $b > 0$, if $x \geq 4alog(2a) + 2b$ then $x \geq alogx + b$.

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:

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

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

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

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

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

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

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:

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

Using this inquality iteratively:

$$\begin{aligned}
\|w_{t-1} - w_*\|^2 &\leq \|w_{t-2} - w_*\|^2 - 2 \\
\|w_{t-2} - w_*\|^2 &\leq \|w_{t-3} - w_*\|^2 - 3 \\
&\dots \\
\|w_* - w_*\|^2 &\leq \|w_0 - w_*\|^2 - k
\end{aligned}$$

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

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

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

Problem 15

Part 1 Solution:

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

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

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):

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

Part 2 Solution:

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

The number of iterations of the perceptron classifier is upper bounded by the number of mistakes made in leave one out:

$$\mathbb{E}\Big\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{g_{n-1}(X_i,D_{n,i})\neq Y_i}\Big\}$$

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 makes a mistake - the summation in the expression above. This M is upper-bounded by $(\frac{R}{\gamma})^2$ (Novikoff, 1962), the number of iterations, so we can bound the expected risk of the perceptron classifier by:

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

$$\leq \frac{1}{n}\Big(\frac{R}{\gamma}\Big)^2$$

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.

Part 1 Solution:

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

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

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

Part 2 Solution:

The variance of the risk:

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

For distributions where $\mathbb{P}\{Y = 1\} = \mathbb{P}\{Y = 0\} = \frac{1}{2}$, the expected risk is $\frac{1}{2}$ and the variance is minimized when the empirical risk is as near the expected value as possible, given it can't be $\frac{1}{2}$ when n is odd.

In general, the probability the empirical risk is near $\frac{1}{2}$ can be expressed:

$$= \mathbb{P}\Big\{Bin(n, \frac{1}{2}) = \frac{n+1}{2}\Big\}$$

So:

$$Var(R_n^D(g_n)) = \sum \left\{ \left(\frac{n+1}{2} - \frac{1}{2} \right)^2 \mathbb{P}\{ Bin\big(n, \frac{1}{2}\big) = \frac{n+1}{2} \} \right\}$$

The first term $(\frac{n+1}{2} - \frac{1}{2})^2$ will always be greater than 1, so can be lower bounded by:

$$\geq \mathbb{P}\{ Bin(n, \frac{1}{2}) = \frac{n+1}{2} \}$$

$$= \frac{1}{2}^n \begin{pmatrix} n \\ \frac{n+1}{2} \end{pmatrix}$$

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

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