Stats 231C - Homework 1

Peter Racioppo (103953689)

April 2020

Problem 1. *Compute the VC-dimension of the following sets of hypotheses (i), (ii), (iii), (iv).*

(*i*) *Positive rays: $\mathcal{H}$ consists of all hypotheses $h : R \mapsto \{0, 1\}$ of the form $h(x) = sgn(x - a)$, i.e., the hypotheses in one dimension that are 0 left of some value a and 1 elsewhere.*

For $\forall x, y \in R$, with $x < y$, $h(x) = \text{sgn}(x - a)$ with $a = (x + y)/2$ sets $x = 0$ and $y = 1$. However, the reverse assignment cannot be made. In other words, no $h \in \mathcal{H}$ can compute more than 3 of the 4 possible dichotomies. In the case of a single point $x \in R$, $h_1(x') = \text{sgn}(x' - \delta)$ and $h_2(x') = \text{sgn}(x' + \delta)$, with $\delta > 0$, are sufficient to compute the 2 possible dichotomies. The VC-dimension of $\mathcal{H}$ is thus 1.

(*ii*) *Positive intervals: $\mathcal{H}$ consists of all hypotheses in one dimension that are 1 within some interval and 0 elsewhere.*

Clearly, $\mathcal{H}$ can shatter any set of two distinct points. For $\forall x, y \in R$, with $x < y$, let $\delta = y - x$. Let $h_1(x')$ return 1 in the interval $[x' - \delta/2, x' + \delta/2]$ and let $h_2(x')$ return 1 in the interval $[y' - \delta/2, y' + \delta/2]$. Let h_3 return 1 in the interval containing x and y and let h_4 return 1 in an interval not containing x or y. Then $\{h_1, h_2, h_3, h_4\}$ are sufficient to compute the 4 possible dichotomies. On the other hand, $\mathcal{H}$ cannot shatter a set of three distinct points. For $\forall x, y, z \in R$, with $x < y < z$, no function in $\mathcal{H}$ can return 1 for x and z and 0 for y. The VC-dimension of $\mathcal{H}$ is thus 2.

(*iii*) *Convex sets: $\mathcal{H}$ consists of all hypotheses in two dimensions that are 1 inside some convex set and 0 elsewhere.*

Consider three points in $\in R^2$. Any two of these points can be enclosed in a convex set excluding the third point, unless the three points are colinear. It follows that $\mathcal{H}$ can shatter a set of three points. Likewise, consider four points in R^2. We can draw a triangle between any three of these points. Provided none of the four points lies inside the triangle formed by the other three points, this shows that any subset of the four points can be enclosed in a convex set.

Thus, $\mathcal{H}$ can shatter a set of four points.

We can generalize this reasoning by considering a regular polygon in R^2 with n vertices. Given any set of $k < n$ of these points, it is always possible to draw a polygon between these k points, which lies inside of the n-vertex polygon. More generally, consider the case of n points arranged in a circle. In this case, given any subset of k points, one can form a convex polygon with these k points as vertices by preceding clockwise along the circle and drawing lines between neighboring points in the set. The resulting cyclic polygon clearly has interior angles strictly less than 180 degrees, and is therefore convex. It follows that any subset of k points can be assigned a value of 1 by some $h \in \mathcal{H}$ and thus that $\mathcal{H}$ can shatter a set of n points. In other words, $\mathcal{H}$ has a VC-dimension of ∞.

(iv) *Quadratic classifier: $\mathcal{H}$ consists of all functions $R^n \mapsto 0, 1$ of the form*

$$h(x) = sgn(\sum_{1 \le i \le j \le n} w_{ij} x_i x_j + \sum_{1 \le i \le n} w_i x_i + w_0), x = (x_1, ..., x_n) \in R^n$$

where w_{ij}, w_i, w_0 are some real coefficients.

In vector form, $h(x) = x^T W x + w^T x + w_0$, where w_{ij} is the ijth component of W and w_i is the ith component of w. This first term has n^2 terms. However, $x_i x_j$ and $x_j x_i$ are linearly dependent. The number of linearly independent terms in $x^T W x$ are the number of terms above or on the main diagonal of an $n \times n$ matrix: $\frac{n^2-n}{2} + n = \frac{n^2+n}{2}$. Thus, the set S contains $\frac{n^2+n}{2} + n = \frac{1}{2}n^2 + \frac{3}{2}n$ affinely-independent vectors. Let $x_2 \in R^{1 \times \frac{n^2+n}{2}}$ be a vector of the affinely-independent $x_i x_j$ and let $w_2 \in R^{1 \times \frac{n^2+n}{2}}$ be a vector of the corresponding w_{ij}. Then we can write the parameters as $(w_2, w, \theta)^T$. Following the proof of Theorem 13 in the notes, it follows that the VC-dimension of $\mathcal{H}$ is $\frac{1}{2}n^2 + \frac{3}{2}n + 1$.

Problem 2. *Let $n_0, n_1 \in N$. Compute the VC-dimension (or upper and lower bounds for the VC-dimension) of a model $\mathcal{H}$ consisting of functions $h : R^{n0} \mapsto \{0, 1\}$ of the form $h = sgn(f)$, where $f : R^{n_0} \mapsto R; x = (x_1, ..., x_{n0}) \mapsto \sum_{j=1}^{n_1} w_{1j}^{(2)} (\sum_{i=1}^{n_0} w_{ji}^{(1)} x_i)$, for some $w^{(1)} \in R^{n_1 \times n_0}$ and $w^2 \in R^{1 \times n_1}$.*

These variable names are confusing, so let $w = w^{(1)}$ and $y = w_1^{(2)}$. Then, $f = \sum_{j=1}^{n_1} w_{1j}^{(2)} (\sum_{i=1}^{n_0} w_{ji}^{(1)} x_i) = \sum_{j=1}^{n_1} \sum_{i=1}^{n_0} y_j w_{ji} x_i = \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} y_j w_{ji} x_i$. Letting $z_i = \sum_{j=1}^{n_1} y_j w_{ji}$, this is simply $\sum_{i=1}^{n_0} z_i x_i$. Thus $\mathcal{H}$ is just the set of perceptrons $h : R^{n_0} \mapsto \{0, 1\}$ of the form $h = \text{sgn}(\sum_{i=1}^{n_0} z_i x_i)$, so the VC-dimension is $n_0 + 1$.

Problem 3. *Suppose we have a simple learning model $\mathcal{H}$ whose growth function is $\prod_{\mathcal{H}}(m) = m + 1$ and hence VCdim $(\mathcal{H}) = 1$. Estimate the probability that er_P will be within 0.1 of $e\hat{r}_z$ given 10, 100, and 10,000 training examples. Explain the steps of your calculation and the results that you are using. What if*

instead of 1 the model under consideration has VC-dimension $n(n-1)/2+n+1$ *and* $n = 4$*?*

Let $P = P^m\{|er_P(h) - e\hat{r}_z(h)| \leq \epsilon \text{ for } \forall \text{ h} \in \mathcal{H}\}$ and $Q = 1 - P = P^m\{|er_P(h) - e\hat{r}_z(h)| \geq \epsilon \text{ for some } h \in \mathcal{H}\}$. By Theorem 4.3 (Theorem 18 in the notes), $Q \leq 4\prod_{\mathcal{H}}(2m)\exp(-m\epsilon^2/8)$. Thus, $Q(m) \leq 4(2m+1)\exp(-m\epsilon^2/8) = 4(2m+1)\exp(-m/800)$, since $\epsilon = 0.1$. Thus, $Q(10) \leq 84\exp(-1/80) \approx 83.0$, $Q(100) \leq 804\exp(-1/8) \approx 709.5$., and $Q(10,000) \leq 80,004\exp(-25/2) \approx 0.30$. These first two inequalities are useless since the right hand sides are greater than 1. All we can say is that $P(10) \geq 0$, $P(100) \geq 0$, and $P(10,000) \geq 0.7$. (If the VC-dimension is 1, $\mathcal{H}$ can only shatter a set with one point, which is equivalent to the case of a single decision boundary in one dimension (as in Problem 1(a)). However, we know nothing about the underlying distribution of points, so I don't think this information can be leveraged.) A plot of the log-probability bound as a function of m is shown in Fig. 1. The probability bound passes 0 at $m = 8,328$, 0.9 at $m = 10,343$, and 0.99 at $m = 12,326$.

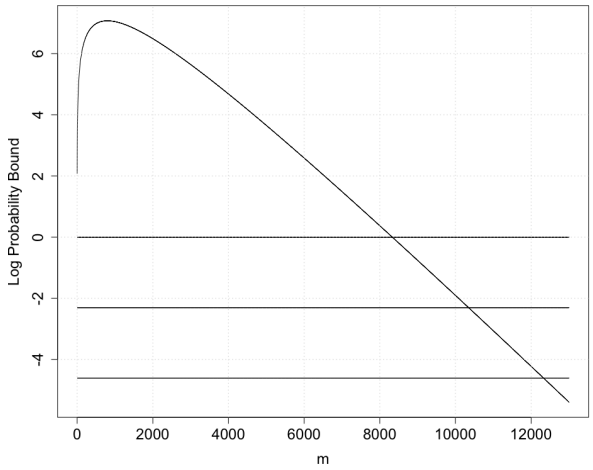

Figure 1: Log-probability bounds vs m. The horizontal lines correspond to probability bounds of 0, 0.9, and 0.99.

If instead VC-dim$(\mathcal{H}) = n(n-1)/2 + n + 1$ and $n = 4$, VC-dim$(\mathcal{H}) = 11$. By Theorem 3.7, $\prod_{\mathcal{H}}(m) = 2^m$ if $m \leq d$ and $\prod_{\mathcal{H}}(m) < (em/d)^d$ if $m > d$. Thus, $\prod_{\mathcal{H}}(10) < 2.1 \times 10^4$, $\prod_{\mathcal{H}}(10) < 2.1 \times 10^{15}$, and $\prod_{\mathcal{H}}(10,000) < 2.1 \times 10^{37}$. Again using Theorem 4.3, $Q \leq 4\prod_{\mathcal{H}}(2m)\exp(-m\epsilon^2/8)$, but for $m = 10,000$, $\exp(-m\epsilon^2/8) \approx 3.7 \times 10^{-6}$, so the bound on the growth function is much larger than the exponential term even for $m = 10,000$. In other words, these bounds tell us nothing. A plot of the log-probability bound as a function of m is shown in Fig. 1. The probability bound passes 0 at $m = 89,102$, 0.9 at $m = 91,143$,

and 0.99 at $m = 93,180$.

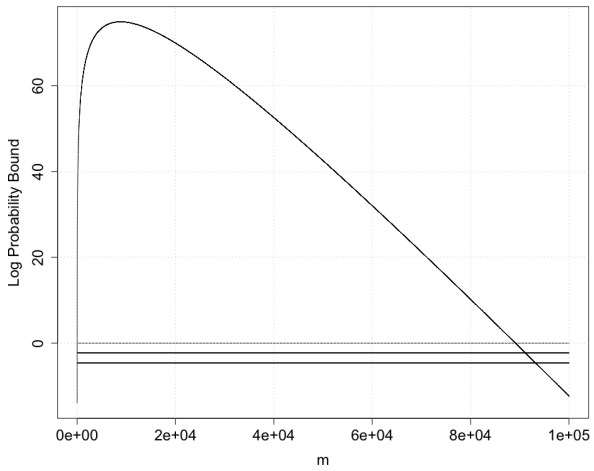

Figure 2: Log-probability bounds vs m. The horizontal lines correspond to probability bounds of 0, 0.9, and 0.99.

Problem 4. *Consider a binary classification problem. Assume that you have a data set z consisting of 600 examples, $z(i) = (x(i), y(i)), i = 1, ..., 600$. To properly test the performance of the final hypothesis, you set aside a randomly selected subset of 200 examples z_{test} which are never used in the training phase; these form a test set. You use a finite learning model which has 1000 hypotheses and select the final hypothesis g based on 400 training examples, z_{train}. You wish to estimate $er_P(g)$. You have access to two estimates: $e\hat{r}_{ztrain}(g)$, the training error on the 400 training examples z_{train}; and, $e\hat{r}_{ztest}(g)$, the test error on the 200 test examples z_{test} that were set aside.*
** Using a 5% error tolerance ($\delta = 0.05$), which estimate has the higher 'error bar' ϵ?*
** Is there any reason why you should / should not reserve even more examples for testing?*

Using Hoeffding's Inequality, as in the proof of Theorem 2.4, if $\epsilon_L(m, \delta) \geq (\frac{1}{2m}\ln(\frac{2|\mathcal{H}|}{\delta}))^{\frac{1}{2}}$, then with probability at least $1-\delta$, for every $h \in \mathcal{H}$, $|er_P(L(z)) - e\hat{r}_z(L(z))| \leq \epsilon$. With $m = 400$, $\delta = 0.05$, and $|\mathcal{H}| = 1000$, the 'error bar' for the training case is $\epsilon_{train} = (\frac{1}{2\times 400}\ln(\frac{2\times 1000}{0.05}))^{\frac{1}{2}} \approx 0.115$. In the testing case, $\mathcal{H}$ consists only of the optimal training hypothesis g. Thus, $\epsilon_{test} = (\frac{1}{2\times 200}\ln(\frac{2\times 1}{0.05}))^{\frac{1}{2}} \approx 0.070$.

Reserving more data for testing increases testing accuracy, but decreases train-

ing accuracy. If the number of hypotheses in $\mathcal{H}$ is large, ϵ_{train} will be large, and more data will be needed for training (to distinguish between the hypotheses). A plot of training error and testing error is displayed in Fig. 3. Evidently, anywhere between about 200 and 500 training samples keeps the sum of training and testing errors near its minimum.

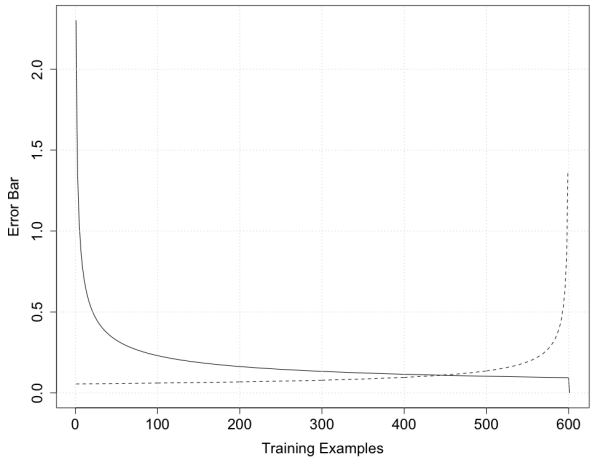

Figure 3: Error vs. Training Examples. The solid line represents the training error and the dotted lines the testing error.