CS 5350/6350: Machine Learining Spring 2015

Homework 3 Solution

Handed out: Feb 4, 2015
Due date: ~~Mar 4~~ Mar 8, 2015

1 Warm up: Feature expansion

There are several possible answers to this question. One is $\phi(x_1, x_2) = f(x_1^2,\ x_2^2)$.

To show that the function is linearly separable after this transformation, we need to produce a $\mathbf{w}$ and b such that the original function is equivalent to $sgn(\mathbf{w}^T\phi(x_1, x_2) - b)$. One such $\mathbf{w}$ is the vector $[-1, -1]^T$ and $b = -r^2$.

2 PAC Learning

1. (a) $2^N - 1$. For each part, it has 2 possible states: used or not used. So there are 2^N possible different makeup of a product. However, we exclude the case that a product is made of nothing.

   (b) $3^N - 1$ ($4^N - 1$ or $5^N - 1$ are also acceptable). Since cutting will add an additional state to the parts, if two sections are treated as identical state, then the asnwer will be $3^N - 1$. If two sections treated as distinct states, then it would be $4^N - 1$. Based on that, if someone argues that combining two section is not equivalent to use a part as a whole, then the answer will be $5^N - 1$ in this case.

   (c)
   $$m \geq \frac{1}{\epsilon}\left(\ln(|H|) + \ln(1/\delta)\right) = \frac{1}{0.01}(\ln|H| + \ln(\frac{1}{0.01}))$$

   If $|H| = 3^6 - 1$, then $m \geq 1119.55$ so the robot have to see at least 1120 examples;
   If $|H| = 4^6 - 1$, then $m \geq 1292.27$ so the robot have to see at least 1293 examples;
   If $|H| = 5^6 - 1$, then $m \geq 1426.17$ so the robot have to see at least 1427 examples.

2. Using Chernoff bound (here we change the name of variable γ to η for avoiding confusion)
   $$Pr[S/m < (1-\eta)p] \leq \exp(-mp\eta^2/2)$$
   we have
   $$Pr[\ error_S(h) < (1-\eta)error_D(h)\ ] \leq \exp(-m \cdot error_D(h)\eta^2/2)$$

where $error_S(h)$ is the training error; $error_D(h)$ is true error. Rearranging, we have

$$Pr\left[\ error_D(h) > \frac{error_S(h)}{(1-\eta)}\ \right] \leq \exp(-m \cdot error_D(h)\eta^2/2)$$

Since our goal is to find $Pr[error_D(h) > error_S(h)(1+\gamma)]$, we need to let

$$\frac{1}{1-\eta} = 1+\gamma$$

which yields

$$\eta = \frac{\gamma}{(1+\gamma)}\ , 0 \leq \eta \leq 1$$

Substitute η by $\gamma/(1+\gamma)$ we have

$$Pr[\ error_D(h) > error_S(h)(1+\gamma)\ ] \leq \exp(m \cdot error_D(h)\gamma^2/(1+\gamma)^2/2)$$

Next, we upper bound the probability that a hypothesis h has a large error by δ:

$$Pr[\exists h \in H; error_D(h) > (1+\gamma) error_S(h)] \leq |H| \exp(-mp\gamma^2/(1+\gamma)^2/2) \leq \delta$$

where $p = \min_{h \in H} error_D(h)$(minimum error of h among $|H|$ over D).

Taking log (base e) on both sides, we get

$$\ln(\delta) \leq \ln(|H|) - mp\gamma^2/(1+\gamma)^2/2$$

Therefore,

$$m \geq \frac{2(1+\gamma)^2}{p\gamma^2}(\ln(|H|) + \ln(1/\delta))$$

3 VC Dimension

1. [Shattering, 10 points] Suppose a finite set S has the following form:

$$S = \{(0, \underbrace{1, 1, \cdots, 1}_{n-1}), (1, 0, \underbrace{1, 1, \cdots, 1}_{n-2}), \cdots (\underbrace{1, 1, \cdots 1}_{n-1}, 0)\}$$

In the above definition, there are exactly n examples in S. And for each vector v of size n in S, $n-1$ bits are 1 and one bit is 0. Let v_i denotes the vector in S such that only the i^{th} bit of v_i is 0. Without loss of generality, assume there are $k(k \in [0, n])$ examples with "-" label and $n-k$ examples with "+" label. After labelling, $\{v_{n_1}, v_{n_2}, \cdots, v_{n_k}\}$ will be assigned negative label "-".

Now we claim S can be shattered by the set of all conjunctions of n boolean variables. The conjunction has the following form:

$$f(v) = b(n_1) \wedge b(n_2) \wedge \cdots \wedge b(n_k)$$

where $b(i)$ is a function that takes the i^{th} value from input vector v.

Clearly, any vector v that has 0 at its n_i-th $(1 \leq i \leq k)$ bit will result 0 and any vector doesn't have 0 at all its n_i-th $(i = 1, \cdots, k)$ bit will result 1. Thus, the above boolean conjunction function can shatter S mentioned at the beginning.

2. [10 points] Proof by contradiction: Suppose $VC(C) = d > \log|C|$, then there must exist at least one set of d points such that C can shatter each dichotomy of these points. Since the number of possible dichotomies is 2^d, we have $2^d > 2^{\log_2|C|} = |C|$. However, the maximum number of dichotomies of points that C can shatter is at most $|C|$ so $|C| = 2^d$, which contradicts $2^d > |C|$. So $VC(C) \leq \log|C|$.

3. [15 points] $VC(H) = 2$. First, we prove that $VC(H) \geq 2$. Pick 2 points $p_1(x_11, x_21)$ and $p_2(x_12, x_22)$ such that $x_11 < x_12$ and $x21 < x_22$. For all 4 possible labelings, we can always adjust a and b so that H can shatter them.
   Second, we prove $VC(H) < 3$. For the possible positions of 3 points,There are 2 cases to consider. 1) 3 points are collinear and 3) 3 points are not collinear. In each case, we cannot not find such 3 points so that H can shatter them.

   to do. Combining $VC(H) \geq 2$ and $VC(H) < 3$ we get $VC(H) = 2$.

4. [15 points] 4. If there are 4 points, then there are 16 different ways of labeling. And for each labeling, all 4 points can be shattered by the 2 intervals. However, when there are 5 points with label "+1,-1,+1,-1,+1". We need at least 3 intervals to shatter them. So 5 points cannot be shattered by union of 2 intervals.

5. [**For 6350 Students**, 10 points] $2k$. Clearly, with k intervals, $2k$ distinct points on the real line can be shattered. Now, consider there are $2k + 1$ points on the read line. If successive points are labeled with +1 and -1 alternatively (starting label is +1), there will be $k + 1$ points with +1 label and k points with -1 label. And at least $k + 1$ intervals are required to shatter these $2k + 1$ points.

6. [**For 6350 Students**, 15 points] Assume $VC(H_1) = d$. Then we know there exists a set S of size d that is shattered by H_1. And since H_2 contains all hypotheses in H_1, then we must have H_2 shatters S. Therefore, $VC(H_1) = d \leq VC(H_2)$.