
CSCI 6100 RPI
Machine Learning Spring 2001

ASSIGNMENT 3, due Thursday, Marrch 1

Discussions are encouraged but no written records can be taken away from a discussion. Books and notes can be consulted but not copied from. Homeworks are due in class or in my mail box by 4pm on the due date. For experimental problems, debugging help may be obtained but all code should be your own and all results reported should be from running your own code. In order to verify "curious" results, we may request you submit your code so DO NOT delete your code after you hand in a problem set.

1 (200pts) Coin Model / Bin Model

A data set is drawn by tossing coins (independently) as described in class – each coin representing a function. The coins are independent (a coin popping heads (for a given data point) is independent of the other coins). i.e., assume independence wherever needed. For a given coin, let the probability of heads (probability of error) be π. The probability of obtaining k heads in N tosses is given by the binomial distribution:

$$P[k|N,\pi] = \binom{N}{k}\pi^k(1-\pi)^{N-k}$$

Remember that the training error ν is $\frac{k}{N}$.

(a) Assume the sample size (N) is 10. If all the coins have $\pi = 0.05$ compute the probability that at least one coin will have $\nu = 0$ for the case of 1 coin, 1,000 coins, 1,000,000 coins. Repeat for $\pi = 0.8$.

(b) For the case $N = 6$ and 2 coins with $\pi = 0.5$ for both coins, plot the probability

$$P[\max_i |\nu_i - \pi_i| \geq \epsilon]$$

for ϵ in the range $[0, 1]$ (the max is over coins). On the same plot show the bound that would be obtained using the Chernoff bound as a function of ϵ. Remember that for a single coin, the Chernoff bound is

$$P[|\nu - \pi| \geq \epsilon] \leq 2e^{-2N\epsilon^2}$$

(Hint: use $P[A \textit{ or } B] = P[A] + P[B] - P[A \textit{ and } B] = P[A] + P[B] - P[A]P[B]$, where the last equality follows by independence to evaluate $P[\max\ldots]$)

2 (200pts) Computation of $M(N)$

In this problem, compute the maximum number of dichotomies, $M(N)$, for the various simple learning models, and hence compute the d_{VC}, the VC dimension.

(a) Positive or negative interval:

$$\mathcal{L} = \left\{ g \;\middle|\; g(x) = \begin{cases} 1 \text{ for } a < x < b, \text{ some } a, b \in \mathbf{R} \\ -1, \text{ otherwise} \end{cases} \right\} \bigcup \left\{ g \;\middle|\; g(x) = \begin{cases} -1 \text{ for } a < x < b, \text{ some } a, b \in \mathbf{R} \\ 1, \text{ otherwise} \end{cases} \right\}$$

(b) Two concentric spheres in $\mathbf{R}^d$:

$$\mathcal{L} = \left\{ g \;\middle|\; g(\mathbf{x}) = \begin{cases} 1 \text{ for } a < \|\mathbf{x}\| < b, \text{ some } a \leq b \in \mathbf{R}, \text{ and } \|\mathbf{x}\| = \sqrt{x_1^2 + \ldots + x_d^2} \\ -1, \text{ otherwise} \end{cases} \right\}$$

3 (400pts) VC Dimension

Compute the VC dimension for the following learning models.

(a) Continuous functions in $\mathbf{R}^2$:

$$\mathcal{L} = \{g \mid g(\mathbf{x}) = sign(c(\mathbf{x})), \text{ for some } c(\mathbf{x}), \text{ a continuous function}\}$$

(b) Monotonically increasing functions in $\mathbf{R}^2$:

$$\mathcal{L} = \{g \mid \mathbf{u} \geq \mathbf{v} \Rightarrow g(\mathbf{u}) \geq g(\mathbf{v})\}$$

where $\mathbf{u} \geq \mathbf{v}$ if and only if the inequality is satisfied for every component.

(Hint: Consider a set of N points generated by choosing one point and generating the next point by increasing the first component and decreasing the second component until N points are obtained.)

(c) Show that the VC dimension of the simple perceptron in d dimensions is $D_{VC} = d + 1$ as follows (or using any other method of your choice)

  (i) There is a set of $d + 1$ points that can be shattered (fully dichotomized). (Hint: Construct an invertible matrix using $d + 1$ points and conclude that a perceptron can implement any dichotomy on these points.)

  (ii) No set of $d + 2$ points can be fully dichotomized. (Hint: Use the fact that any $d + 1$ vectors in d dimensions must be linearly dependent to construct a dichotomy on those $d + 2$ points that cannot be implemented by a perceptron.

(d) Prove that the VC dimension of the positive rectangle learning model discussed in class is given by $d_{VC} = 4$ as follows. Thus, give a bound for $m(N)$.

  (i) Show that there is a set of 4 points on which all 2^4 dichotomies can be done.

  (ii) Show that on *any* set of 5 points, all dichotomies cannot be done.

4 (200pts) Bounds on the Risk

(a) For a learning model with VC dimension, $d_{VC} = 5$, approximately determine the smallest number of examples required for $|\pi - \nu| < 0.05$ to hold with probability greater than 0.95. What about if $d_{VC} = 6$?

(b) Assume the following theorem to hold

Theorem

$$P\left[\sup_{g \in \mathcal{L}} \frac{\pi_g - \nu_g}{\sqrt{\pi_g}} \geq \epsilon\right] \leq C M_{\mathcal{L}}(2N) exp\left(-\frac{\epsilon^2 N}{4}\right)$$

where C is a constant a little larger than 6.

It is useful because sometimes what you care about is not the absolute generalization error but the percentage generalization error (one can imagine that a generalization error of 0.01 is more significant when $\pi = 0.01$ than when $\pi = 0.5$). Prove that with probability greater than $1 - \eta$, the following inequality holds

$$\pi < \nu + \frac{\xi}{2}\left[1 + \sqrt{1 + \frac{4\nu}{\xi}}\right]$$

where

$$\xi = \frac{4}{N} \log \frac{C M_{\mathcal{L}}(2N)}{\eta}$$

Thus we have a bound (valid with probability close to 1) for the risk in terms of the empirical risk similar to the one derived in class.