LFD Problem Set 2

John Cohen

September 23, 2024

Exercise 1.13

Consider the bin model for a hypothesis h that makes an error with probability μ in approximating a deterministic target function f (both h and f are binary functions). If we use the same h to approximate a noisy version of f given by

$$P(y|x) = \{\lambda \qquad y = f(x), \qquad 1 - \lambda \qquad y \neq f(x)$$

(a) What is the probability of error that h makes in approximating y?
We know the probabilities by h(x) $\neq$ f(x) and when f(x) $\neq$ y. Error occurs when h(x) $\neq$ y and thus it occurs in 2 cases: h(x) = f(x) $\neq$ y (event A) and h(x) $\neq$ f(x) = y (event B). P(error) = P(A) + P(B).
P(A) = P(h(x) = f(x)) * P(f(x) $\neq$ y) = (1-μ) * (1-λ)
P(B) = P(h(x) $\neq$ f(x)) + P(f(x) = y) = (μ)*(λ)
P(error) = (1-μ) * (1-λ) + (μ)*(λ) = $1 - \lambda + \mu(2\lambda - 1)$

(b) At what value of λ will the performance of h be independent of μ? [*Hint: The noisy target will look completely random.*]
This happens when μ is out of the equation. We can do this by getting μ to multiply against 0. Because $\mu(2\lambda - 1)$, $\mu * (0)$ when $(2\lambda - 1) = 0$. That is when $\lambda = \frac{1}{2}$ or = 0.5. This also makes P(error) = 0.5.

Exercise 2.1

By inspection, find a break point k for each hypothesis set in Example 2.2 (if there is one). Verify that $m_H(N) < 2^k$ using the formulas derived in that Example.

The formulas for Positive Ray, Positive Intervals, and Convex Sets are as follows:

(i) $m_H(N) = N + 1$. The break point here is k = 2 where $(2) + 1 = 3 < 4 = 2^{(2)}$
(i) $m_H(N) = 0.5N^2+0.5N+1$. The break point here is k = 3 where $0.5(3)^2+0.5(3)+1 = 7 < 8 = 2^{(3)}$
(i) $m_H(N) = 2^N$. There is no break point here as for any number N $2^N = 2^{(N)}$ thus never validating less than. k = ∞

Exercise 2.2

(a) Verify the bound of Theorem 2.4 in the three cases of Example 2.2:
(i) Positive rays: H consists of all hypotheses in one dimension of the form $h(x) =$ sign(x-a).
(ii) Positive intervals: H consists of all hypotheses in one dimension that are positive within some interval and negative elsewhere.
(iii) Convex sets: H consists of all hypotheses in two dimensions that are positive inside some convex set and negative elsewhere.
(Note: you can use the break points you found in Exercise 2.1.)

(i) Given: Break point k=2, growth function $m_H(N) = N + 1$.
Apply Theorem 2.4:

$$\sum_{i=0}^{k-1}\binom{N}{i} = \sum_{i=0}^{1}\binom{N}{i} = \binom{N}{0} + \binom{N}{1} = 1 + N = N + 1$$

Compare $m_H(N)$ and RHS:

$$m_H(N) = N + 1 \leq N + 1 = \sum_{i=0}^{1}\binom{N}{i}$$

The bound of Theorem 2.4 holds exactly for Positive Rays.

(ii) Given: Break point k=3, growth function $m_H(N) = 0.5N^2 + 0.5N + 1$.
Apply Theorem 2.4:

$$\sum_{i=0}^{k-1}\binom{N}{i} = \sum_{i=0}^{2}\binom{N}{i} = \binom{N}{0} + \binom{N}{1} + \binom{N}{2} = 1 + N + \frac{N(N-1)}{2} = 0.5N^2 + 0.5N + 1$$

So,

$$\sum_{i=0}^{2}\binom{N}{i} = 0.5N^2 + 0.5N + 1$$

Compare $m_H(N)$ and RHS:

$$m_H(N) = 0.5N^2 + 0.5N + 1 \leq 0.5N^2 + 0.5N + 1 = \sum_{i=0}^{2} \binom{N}{i}$$

The bound of Theorem 2.4 holds exactly for Positive Intervals.

(iii) Given: break point k = ∞ (no finite break points), growth function $m_H(N) = 2^N$. There is no finite k such that $2^N < 2^N$ for all N. Convex sets do not satisfy the starting condition of Theorem 2.4 ($m_H(N) < 2^k$) so we conclude here.

(b) Does there exist a hypothesis set for which $m_H(N) = N + 2^{N/2}$ (where [N/2] is the largest integer $\leq N$)?
Yes, such a hypothesis set exists

Exercise 2.3

Compute the VC dimension of H for the hypothesis sets in parts (i), (ii), (iii) of Exercise 2.2(a).

VC dimension of H: k 1 (the most points H can shatter).
Positive rays: $m_H(N) = N + 1$, Break point k = 2, $d_{VC} = 1$
Positive rays: $m_H(N) = 0.5N^2 + 0.5N + 1$, Break point k = 3, $d_{VC} = 2$
Positive intervals: $m_H(N) = 2^N$, Break point $k = \infty$, $d_{VC} = \infty$

Exercise 2.6

A data set has 600 examples. To properly test the performance of the final hypothesis, you set aside a randomly selected subset of 200 examples which are never used in the training phase; these form a test set. You use a learning model with 1,000 hypotheses and select the final hypothesis g based on the 400 training examples. We wish to estimate $E_{out}(g)$. We have access to two estimates: $E_{in}(g)$, the in-sample error on the 400 training examples; and, $E_{test}(g)$, the test error on the 200 test examples that were set aside.

(a) Using a 5% error tolerance ($\delta = 0.05$), which estimate has the higher 'error bar'?
Given: $\delta = 0.05$, H = 1000, $N_{train} = 400$, $N_{test} = 400$, and the hypothesis set of testing is

just 1, that being g(x).

$$\Omega_{in}(N_{train}, H, \delta) = \sqrt{\frac{8}{N_{train}} ln(\frac{4|H|(2N_{train})}{\delta})}$$

$$\Omega_{in}(400, 1000, 0.05) = \sqrt{\frac{8}{400} ln(\frac{4 * 1000 * 2 * 400}{0.05})}$$

$$\Omega_{in}(400, 1000, 0.05) = 0.611$$

$$\Omega_{test}(N_{test}, 1, \delta) = \sqrt{\frac{1}{2N_{test}} ln(\frac{2 * 1}{\delta})}$$

$$\Omega_{test}(200, 1, 0.05) = \sqrt{\frac{1}{400} ln(\frac{2}{0.05})}$$

$$\Omega_{test}(200, 1, 0.05) = 0.096$$

Error Bar for $E_{in}(g)$: Larger due to the dependency on the hypothesis set's complexity (1,000 hypotheses).
Error Bar for $E_{test}(g)$: Smaller since it relies only on the size of the test set and not on the hypothesis set's complexity.
The error bar associated with $E_{in}(g)$ is higher than that of $E_{test}(g)$ because it must account for the risk of over-fitting due to the selection of g from a large hypothesis set based on the training data. Therefore, $E_{test}(g)$ provides a more reliable estimate with a smaller error bar under a 5% error tolerance (δ=0.05).

(b) Is there any reason why you shouldn't reserve even more examples for testing?
You shouldn't reserve more examples for testing because it would reduce the size of your training set, which can harm the learning algorithm's ability to produce an accurate and generalizable hypothesis. The marginal benefit of a slightly more precise test error estimate does not outweigh the significant cost of a less effective learning process due to insufficient training data. There needs to be a careful balance.

Problem 1.11

The matrix which tabulates the cost of various errors for the CIA and Supermarket applications in Example 1.1 is called a *risk* or *lost matrix.*

For the two risk matrices in Example 1.1, explicitly write down the in-sample error E_{in} that one should minimize to obtain g. This in-sample error should weight the different types of errors based on the risk matrix. [*Hint: Consider* $y_n = +1$ *and* $y_n = -1$ *separately.*]

Assume $E_{in} = \frac{1}{N}\sum_{i=1}^{N}$ Error(h(x),y(x)) where

$$\text{Error}_{\text{supermarket}}(h(x), y(x)) = \begin{cases} 1 & \text{for } h(x) = +1 \text{ and } y(x) = -1 \\ 10 & \text{for } h(x) = -1 \text{ and } y(x) = +1 \\ 0 & \text{for } h(x) = y(x) \end{cases}$$

$$\text{Error}_{\text{CIA}}(h(x), y(x)) = \begin{cases} 1000 & \text{for } h(x) = +1 \text{ and } y(x) = -1 \\ 1 & \text{for } h(x) = -1 \text{ and } y(x) = +1 \\ 0 & \text{for } h(x) = y(x) \end{cases}$$

Problem 1.12

(This problem investigates how changing the error measure can change the result of the learning process. You have N data points $y_1 \leq ... \leq y_n$ and wish to estimate a 'representative' value.

(a) If your algorithm is to find the hypothesis h that minimizes the in-sample sum of squared deviations,

$$E_{in}(h) = \sum_{n=1}^{N}(h - y_n)^2,$$

then show that your estimate will be the in-sample mean,

$$h_{mean} = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

$\frac{d}{dh}\sum_{n=1}^{N}(h - y_n)^2 = \sum_{n=1}^{N} 2(h - y_n) = 2\sum_{n=1}^{N}(h - y_n)$
$0 = 2\sum_{n=1}^{N}(h - y_n) = \sum_{n=1}^{N}(h) - \sum_{n=1}^{N}(y_n)$
$h * N = \sum_{n=1}^{N}(y_n)$
$h = \frac{1}{N}\sum_{n=1}^{N}(y_n)$
The hypothesis that minimizes the sum of squared deviations is the in-sample mean, $h_{mean} = \frac{1}{N}\sum_{n=1}^{N}(y_n)$

(b) If your algorithm is to find the hypothesis h that minimizes the in-sample sum of the absolute deviations,

$$E_{in}(h) = \sum_{n=1}^{N}|h - y_n|,$$

then show that your estimate will be the in-sample median h_{med}, which is any value for which half the data points are at most h_{med} and half the data points are at least h_{med}.

The median or h_{med} is any value for which half the data points are greater than or equal to it, and half are less than or equal to it. When h is less than the median, moving closer to the median decreases the sum of deviations from data points greater than h more than it increases the deviations from points less than h. Similarly, when h is greater than the median, moving h closer to the median decreases the deviations from data points less than h more than it increases the deviations from point greater than h. With this, we know that the median must be the value of h that minimizes the sum of absolute deviations.

(c) Suppose y_N is perturbed to $y_N + \epsilon$, where $\epsilon \implies \infty$. So, the single data point y_N becomes an outlier. What happens to your two estimators h_{mean} and h_{med}?

h_{mean} is sensitive to all data points. Therefore perturbing y_N will have an affect on the mean. As $\epsilon \implies \infty$, the mean will move toward the outlier.
h_{med}, however, is not sensitive to outliers. The median by definition is the middle of the sorted data. If y_N is a point greater than the median, it growing will have no effect on the median.
h_{mean} is directly affected by the outlier while h_{med} is stable and unchanged.