Machine Learning From Data HW3

Shane O'Brien

September 2017

Exercise 1.13

a

The probability of error that h makes in approximating y is:

$\mu(1-\lambda)+\lambda(1-\mu)$

This is because there are two possible cases of error. The first part is where h matches f, but f is wrong due to noise. The second part is when the function f is correct, but h doesn't match f.

b

The value is $\lambda = 0.5$

This is because at $\lambda = 0.5$, the noisy target f looks completely random. So now, μ has no importance, because the target is essentially random.

Exercise 2.1

The hypotheses have been charted below:

Analyzing $m_{\mathcal{H}}(N)$				
N	1	2	3	4
Positive Ray	2	**3**	**4**	**5**
Positive Intervals	2	4	**7**	**11**
Convex Set	2	4	8	...

The bold letters are below 2^N, this means that these are break points. 2^N is noted in the higher row. The convex set has no break point.

Exercise 2.2

a

To verify these bounds, we will look at the first break point of each hypotheses.

Positive Ray
Break point is $k = 2$

Based on definition, $m_{\mathcal{H}}(2) = 2 + 1 = 3$

Then the bound is:

$\Sigma_{i=0}^{1} * \binom{2}{i} = 3$

$3 \leq 3$

Positive Interval
Break point is $k = 3$

Based on definition, $m_{\mathcal{H}}(3) = \frac{N^2}{2} + \frac{N}{2} + 1 = 7$

Then the bound is:

$\Sigma_{i=0}^{2} * \binom{3}{i} = 7$

$7 \leq 7$

Convex Set
The convex set has no break point, so Theorem 2.4 cannot be used.

b

No, there is no hypotheses set for which $m_{\mathcal{H}}(N) = N + 2^{N/2}$. This is because there are only two types of hypotheses sets. They can either be polynomial, or of the form 2^N.

Exercise 2.3

Positive Ray
$d_{vc}(H) = 1$

Positive Interval
$d_{vc}(H) = 2$

Convex Set

$d_{vc}(H) = \infty$

Exercise 2.6

a

$E_{in} = \sqrt{\frac{1}{800} * ln\frac{2*1000}{0.05}} = 0.115$

$E_{test} = \sqrt{\frac{1}{400} * ln\frac{2}{0.05}} = 0.096$

E_{test} has the smaller error bar because we are only looking at one hypotheses

b

A potential reason why you shouldn't reserve even more examples for testing is because of E_{in}. If your N for E_{in} gets smaller, your error bar for E_{in} increases.

Problem 1.11

CIA

$E_{in} =$
$\frac{1}{N}\Sigma_{n=1}^{N} * (10 * (f(x_n) = +1 \ \wedge \ h(x_2) \ \neq f(x_n)) + 1 * (f(x_n) = -1 \ \wedge \ h(x_2) \ \neq f(x_n)))$

Supermarket

$E_{in} =$
$\frac{1}{N}\Sigma_{n=1}^{N} * (1 * (f(x_n) = +1 \ \wedge \ h(x_2) \ \neq f(x_n)) + 1000 * (f(x_n) = -1 \ \wedge \ h(x_2) \ \neq f(x_n)))$

Problem 1.12

a

To show this, we just need to mathematically transform

$\Sigma_{n=1}^{N}(h^2 - 2hy_n + y_n^2)$

into

$h = \frac{1}{N}\Sigma_{n=1}^{N} \ y_n$

So, follow these steps:

$\Sigma_{n=1}^{N}(h^2 - 2hy_n + y_n^2)$

$\Sigma_{n=1}^{N}h^2 - \Sigma_{n=1}^{N}2hy_n + \Sigma_{n=1}^{N}y_n^2$

Derive, set equal to 0...

$2Nh - \Sigma_{n=1}^{N}2y_n = 0$

$h = \frac{1}{N}\Sigma_{n=1}^{N}\ y_n$

b

To prove this correctly, we would need to show that

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

just picks the median value. We can do this by using an algorithm that picks a value that is off by just a tiny bit in both directions. If this slight deviation is bigger in both directions, then minimizing would have to be the mean.

Case 1: Above Median
We know that a number $k \geq \frac{N}{2}$, so that $N - k$ is the amount of data points $> median$. If our algorithm picks $median + \epsilon$, where ϵ is some very small number between 0 and 1.

We can split our $E_{in}(h) = \Sigma_{n=1}^{N}|h - y_n|$ into

$\Sigma_{n=1}^{k}|h - y_n| + \Sigma_{k}^{N}|h - y_n|$

then the differences between picking a regular h_{med} and picking $h_{med} + \epsilon$ would amount to:

$k\epsilon - (N - k)\epsilon$

Since $(2k - N)$ is a positive number, and ϵ is between [0,1], $k\epsilon - (N - k)\epsilon$ adds value onto our h. This means that if there is any value added to our h to make it above the $median$, we do not minimize our function.

Case 2: Below Median
We know that a number $k \geq \frac{N}{2}$, so that $N - k$ is the amount of data points $< median$. If our algorithm picks $median + \epsilon$, where ϵ is some very small number between 0 and 1.

We can split our $E_{in}(h) = \Sigma_{n=1}^{N}|h - y_n|$ into

$\Sigma_{n=1}^{N-k}|h-y_n| + \Sigma_{N-k}^{N}|h-y_n|$

then the differences between picking a regular h_{med} and picking $h_{med}+\epsilon$ would amount to:

$k\epsilon - (N-k)\epsilon$

Since $(2k-N)$ is a positive number, and ϵ is between [0,1], $k\epsilon-(N-k)\epsilon$ adds value onto our h. This means that if there is any value subtracted from our h to make it above the *median*, we do not minimize our function.

c
h_{mean} is just an average, so that goes off to infinity

h_{med} is just the middle value, so at most, it will shift by one. There is a 50 percent chance the original y_n is below h_{med}. In this case, h_{med} goes up by one index. There is a 50 percent chance the original y_n is above h_{med}. In this case, h_{med} stays the same.