Machine Learning from Data Assignment 3

Greg Stewart

September 24, 2018

Exercise 1.13

(a) *What is the probability of error that h makes in approximating y?* We want

$$\mathbb{P}[h(x) = y].$$

Given the functions we have, this means we need to calculate

$$\mathbb{P}[h(x) = f \text{ and } f(x) \neq y] + \mathbb{P}[h(x) \neq f \text{ and } f = y]$$

Noting that, given the independence of the two variables, $\lambda = 0.5$, this is given by

$$\begin{aligned}(1-\mu)(1-\lambda) + \mu\lambda &= 1 - \mu - \lambda + \mu\lambda + \mu\lambda \\ &= 1 - \mu - \lambda + 2\mu\lambda\end{aligned}$$

So the probability given $\lambda = 0.5$ is $\frac{1}{2}$.

(b) *At what value of λ will the performance of h be independent of μ?* Given $\lambda = 0.5$, the above evaluates to

$$\begin{aligned}1 - \mu - \lambda + 2\mu\lambda &= 1 - \mu - 0.5 + \mu \\ &= 0.5\end{aligned}$$

The value we need for λ is 0.5.

Exercise 2.1

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

1. *Positive rays.*

   The break point is $k = 2$. When there are two points, there are only three possible dichotomies. $3 < 2^k = 2^2 = 4$, which also follows from the formula derived for m_H.

2. *Positive intervals.*

   The break point is $k = 3$. Given three points, all dichotomies except one are possible—only the dichotomy where the middle point is -1 and the separated points are +1 is not possible. $m_H = 7 < 2^3 = 8$, which also follows from the given formula.

3. *Convex sets.*

   There is no break point in this case. It's clear that there is always a case where all dichotomies are possible given k points, so $m_H(k) = 2^k$ will always be true.

Exercise 2.2

(a) *Verify the bound of Theorem 2.4 in three cases of Ex. 2.2.*

(i) *Positive rays.*
$k = 2$. So

$$\begin{aligned} m_H(N) &\leq \sum_{i=0}^{k-1} \binom{N}{i} \\ &\leq \binom{N}{0} + \binom{N}{1} \\ &\leq \frac{N!}{N!} + \frac{N!}{(N-1)!} \\ &\leq 1 + N \end{aligned}$$

The result agrees with the formula given in Ex 2.2.

(ii) *Positive intervals.*
$k = 3$. So

$$\begin{aligned} m_H(N) &\leq \sum_{i=0}^{k-1} \binom{N}{i} \\ &\leq \binom{N}{0} + \binom{N}{1} + \binom{N}{2} \\ &\leq \frac{N!}{N!} + \frac{N!}{(N-1)!} + \frac{N!}{2(N-2)!} \\ &\leq 1 + N + \frac{N(N-1)}{2} \\ &\leq 1 + N + \frac{1}{2}(N^2 - N) \\ &\leq 1 + \frac{N}{2} + \frac{N^2}{2} \end{aligned}$$

This agrees with the formula from before.

(iii) *Convex sets.*
There is no break point, so we Theorem 2.4 does not apply to convex sets.

(b) *Does there exist a hypothesis set for which $m_H(N) = N + 2^{N/2}$?*
No. Either m_H is bounded by a polynomial (if there is a breakpoint), or we must have $m_H(N) = 2^N$.

Exercise 2.3

Compute the VC dimension of H for the hypothesis sets in parts (i) - (iii) of 2.2(a)

(i) *Positive rays.*
Since $k = 2$, we use $N = 1$ for the VC dimension:

$$d_{VC}(H) = 2^1 = 2$$

(ii) $k = 3$, so we use $N = 2$:

$$d_{VC}(H) = 2^2 = 4$$

(iii) There is no breakpoint, so

$$d_{VC}(h) = \infty$$

Exercise 2.6

(a) $\delta = 0.05$. $M = 1000$.

The "error bar" is given by

$$\sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}.$$

For the 400 training examples, we get an error bar of

$$\sqrt{\frac{1}{2(400)} \ln \frac{2(1000)}{0.05}} = 0.11509.$$

And for the 200 test examples, the error bar is

$$\sqrt{\frac{1}{2(200)} \ln \frac{2(1)}{0.05}} = 0.09603.$$

So the error bar for the training set E_{out} is larger.

(b) *Is there any reason why you shouldn't reserve even more examples for testing?*

The more examples reserved for testing, the less that can be used for training. This increases the error in E_{out} for training, and could lead to choosing a less than optimal final hypothesis to approximate f.

Problem 1.11

The matrix which tabulates the cost of various errors for the CIA and Supermarket applications in Ex 1.1 is called a risk or loss matrix. For these two matrices, 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.

Let the supermarket (S) and CIA (C) risk matrices be

$$S = \begin{pmatrix} 0 & 1 \\ 10 & 0 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}$$

For S, this means we want to minimize

$$E_{in,S} = 10E(h(\mathbf{x}) = -1, f(\mathbf{x}) = +1) + E(h(\mathbf{x}) = +1, f(\mathbf{x}) = -1)$$

And for C, we want to minimize

$$E_{in,C} = E(h(\mathbf{x}) = -1, f(\mathbf{x}) = +1) + 1000E(h(\mathbf{x}) = +1, f(\mathbf{x}) = -1)$$

Problem 1.12

There are N data points $y_1 \leq \cdots \leq y_N$ and you wish to estimate a "representative" value.

(a) *If the algorithm finds h that minimizes the in sample sum of squared deviations, show the estimate will be in the sample mean*

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

To do this, we can simply minimize $E_{in}(h)$ by setting the derivative to 0.

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

$$2\sum_{n=1}^{N} h - 2\sum_{n=1}^{N} y_n = 0$$

$$\sum_{n=1}^{N} h = \sum_{n=1}^{n} y_n$$

Dividing by N to get the mean, we see that

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

(b) *If the algorithm finds the hypothesis h that minimizes the in sample sum of absolute deviations, show that the estimate will be the in sample median h_{med}.*

The sume of absolute deviations is given by

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

For a set of N data points, let h be less than the leftmost point y_1. As h moves closer to y_1, say by ϵ each move, E_{in} is reduced by $N\epsilon$ each move. Once h passes y_1, we have a net reduction in E_{in} of $(N-2)\epsilon$ each move. When h passes y_2, this is again reduced by ϵ for each move to $(N-4)\epsilon$. This continues, and the change in E_{in} for each move of h decreases by 2ϵ for each y_n passed. Thus, once the point $y_{N/2}$ is passed, we see a *negative decrease* in the change in E_{in}; in other words, an increase. This means that the median y_n is a sort of inflection point which *minimizes* E_{in}. Thus h_{med} minimizes E_{in}.

In the case of even N, the same as above is true, but h_{med} can be any point between two middle points in the set.

(c) *Suppose y_N is perturbed to $y_n + \epsilon$, where $\epsilon \to \infty$. So the single data point y_N becomes an outlier. What happens to the two estimators above?*

For the sum of squared deviations, we will see that $h_{mean} \to \infty$. This follows directly from the defininition of h_{mean}. Since $y_N \to \infty$, the last term of the sum, $\frac{y_N}{N}$, tends to ∞, making the whole sum infinity. Thus

$$h_{mean} \to \infty.$$

In the case of the sum of absolute deviations, there is no change in h_{med}. The median measurement is tolerant of outliers, and in this case, where only the largest term becomes an outlier, the median is not affected at all. So h_{med} does not change.