Fifth Home Assignment
Data Analysis

Nikolaj Dybdahl Rathcke (rfq695)

June 13, 2015

Question 1

(1)

The VC-dimension is 3. No matter how you label the points that makes an equilateral triangle, it can be shattered. However, if you have four points, the diagonal can be the same length, in which case if you label the points in one diagonal one thing, and the other something else, they cannot be shattered. If one diagonal is larger than the other, then if these two points in that diagonal are labeled positive, it cannot be shattered, thus $d_{VC}(\mathcal{H}) = 3$.

(2)

The VC-dimension is ∞. Consider placing n points in a circle. If it is all the convex sets, we can always connect 2 points without including any other point. It does not matter what n is, so $d_{VC}(\mathcal{H}) = \infty$.

(3)

Consider points in a circle where the points make a polygon. With all convex sets, the equation (2.14) makes a very loose bound since $d_{VC} = \infty$, while it for circles in $\mathbb{R}^2$ is tighter as $d_{VC} = 3$.

(4)

The VC-dimension is 2. This case is very easy as you simply put the 2 point next to each other, and they can be shattered. However, with 3 points, we have to place them within 180 degrees of the origin as it cannot be shattered if all points are labeled the same otherwise. This means we place 3 points with some degree from the origin. If the point in middle (another point lies at a degree x which is more than that and another point lies at a degree y which is less than that) is labeled differently than the two other points, it cannot be shattered. Therefore, $d_{VC}(\mathcal{H}) = 2$.

(5)

The VC-dimension is 3. Three points that lie in a triangle, no matter what label, can be shattered. But when there are four points, if the points on a diagonal are labeled the same, but differently than the two others on the other diagonal, it cannot be shattered.

(6)

To determine this, we use equation (2.14) in [LFD] that says

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln(\frac{4((2N)^{d_{VC}} + 1)}{\delta}}$$
$$\Leftrightarrow E_{out}(g) - E_{in}(g) \leq \sqrt{\frac{8}{N} \ln(\frac{4((2N)^{d_{VC}} + 1)}{\delta}}$$

So we need to solve for N on the right side, so it is 0.05 with $\delta = 0.05$ and $d_{VC} = 10$.
This gives the answer $N = 452957$ which is the sample size needed.

Question 2

We have the following

$$X = \begin{bmatrix} 0 & 0 \\ 0 & -1 \\ -2 & 0 \end{bmatrix} \quad y = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$$

We get the inequalities

$$-b \geq 1 \qquad (i)$$
$$w_2 - b \geq 1 \qquad (ii)$$
$$-2w_1 + b \geq 1 \qquad (iii)$$

Combining (i) and (ii) gives

$$w_2 \geq 0$$

And combining (i) and (iii) gives

$$w_1 \leq -1$$

This means the following

$$(b^* = -1, w_1^* = -1, w_2^* = 0)$$

makes the optimal hyperplane. It has the margin

$$\frac{1}{\|w^*\|} = \frac{1}{\sqrt{1}} = 1$$

Question 3

(1)

One way to generate random points is to first pick an angle in $[0, 2\pi)$, then pick a random length in $[0, 1]$ and then find the corresponding x and y-coordinates. Let θ be the angle and let c be the length, then the coordinates (x, y) are given by

$$x = \frac{c \cdot sin(\theta)}{\sin(\pi/2)}$$
$$y = \sqrt{c^2 - x^2}$$

Using the law of sines to find x and then using the Pythagorean theorem to find y.

Question 4

We want to implement and apply a soft-margin SVM to the MNINST dataset. The soft SVM solution which was uploaded to absalon has been used.
First we build a classifier for distinguishing between digits "0" and "1", which we test with the first 250, 500, and 1000 occurences of the digits from the training set. Then we apply soft-margin SVM with $C = [10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100]$ and each of these we want to provide a graph of E_{in}, $||w||$, and E_{test} as a function of C. The error E_{test} is calculated with the first 1000 occurences of the numbers in the test set. This can be seen in Figure 1.

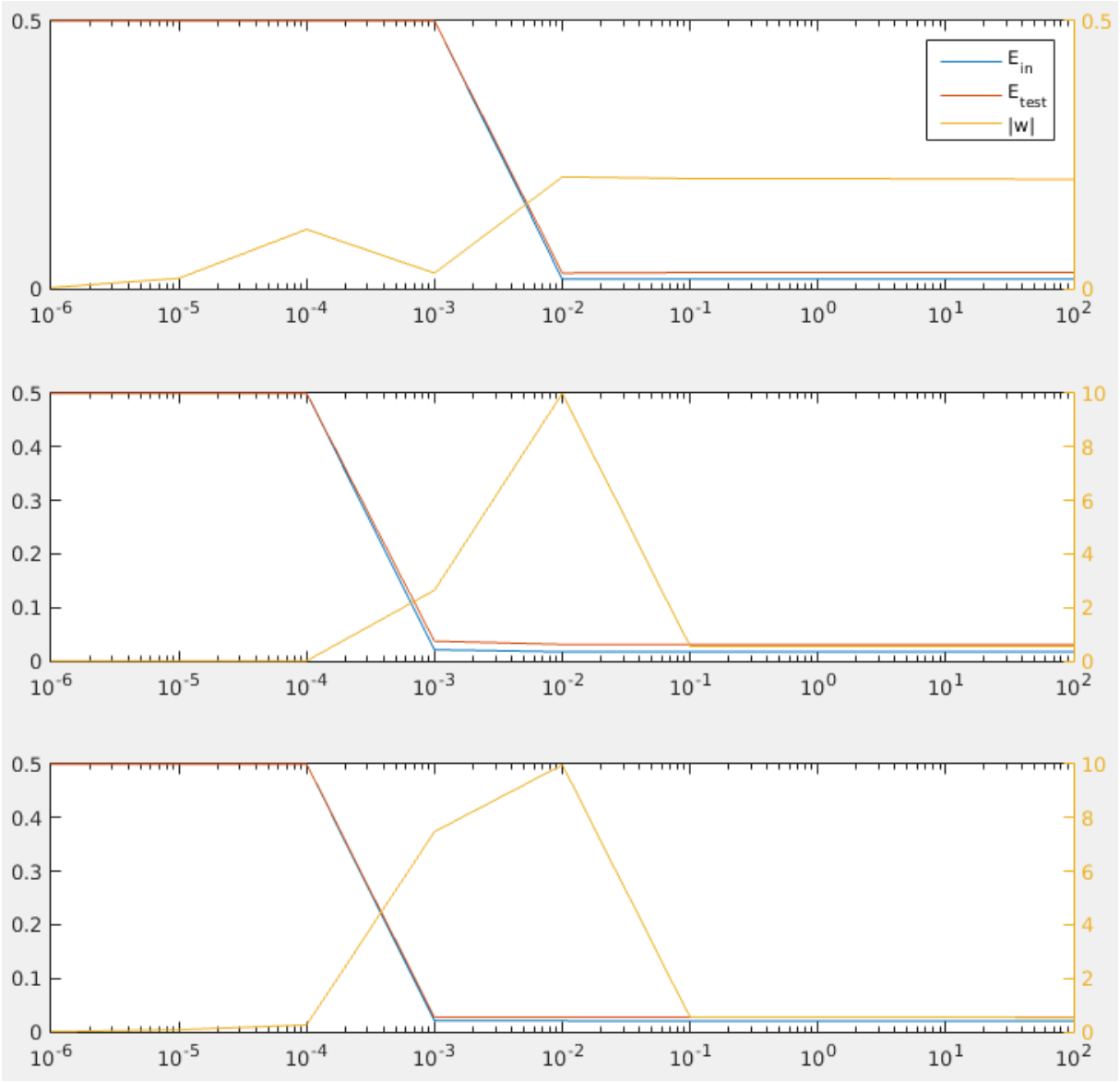

Figure 1: E_{in}, $||w||$, and E_{test} as a function of C, for the first 250, 500, and 1000 occurences of "1" and "0" in the dataset, the top graph is for 250, then 500 and at the bottom 1000.

We do the same for the digits "0" and "8", this can be seen in Figure 2.

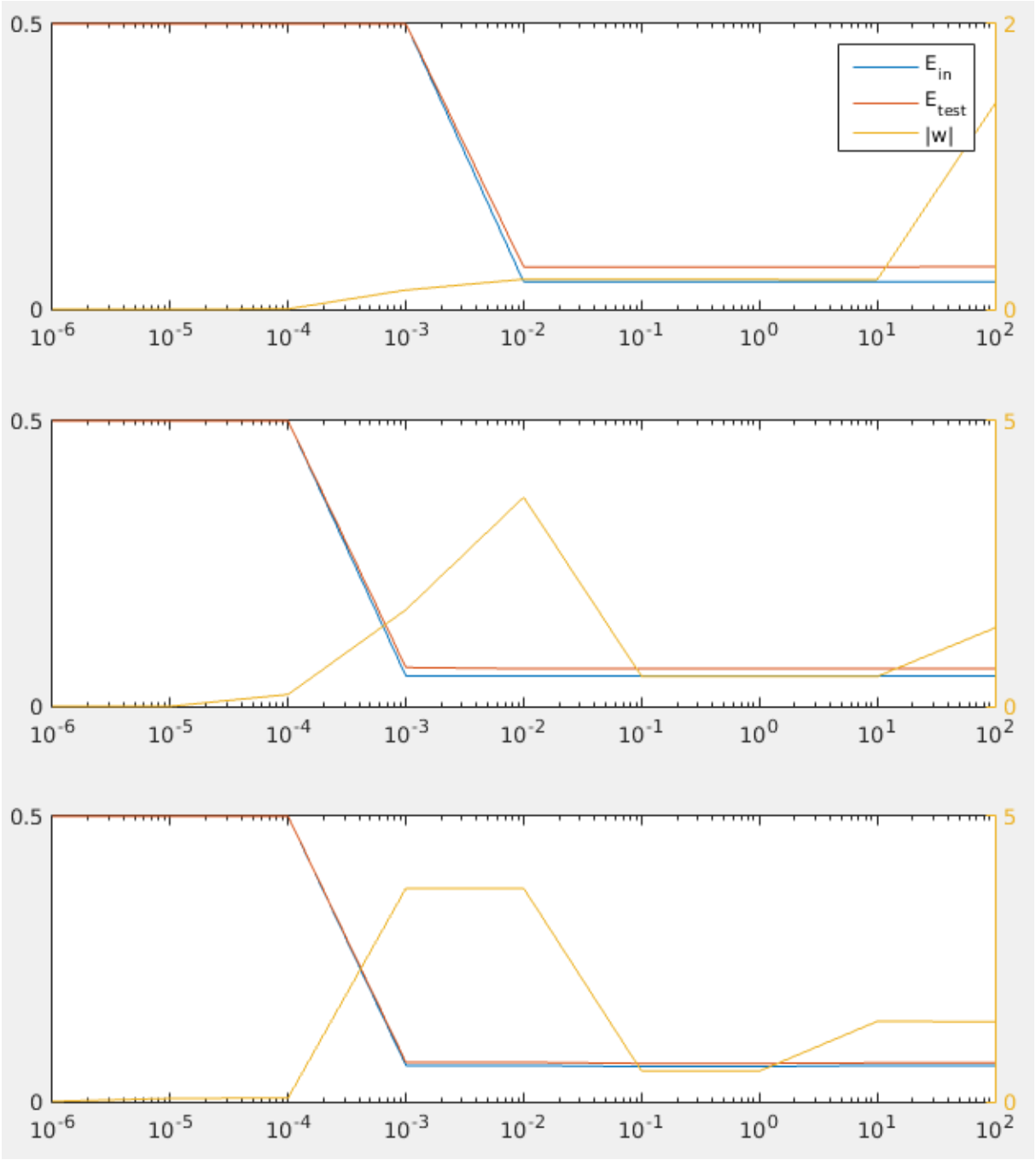

Figure 2: E_{in}, $||w||$, and E_{test} as a function of C, for the first 250, 500, and 1000 occurences of "8" and "0" in the dataset, the top graph is for 250, then 500 and at the bottom 1000.

We do the same for the digits "5" and "6", this can be seen in Figure 3.

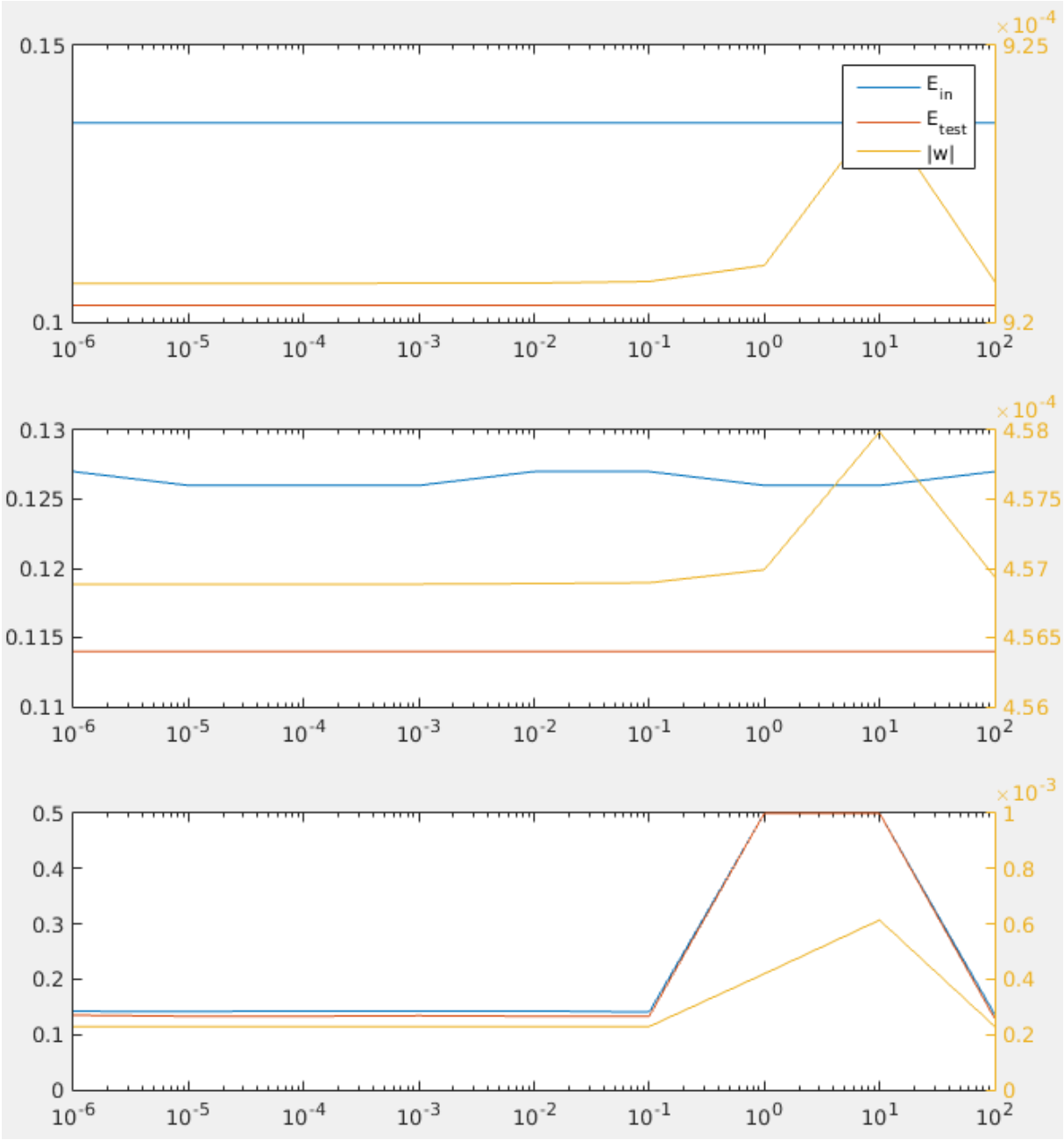

Figure 3: E_{in}, $||w||$, and E_{test} as a function of C, for the first 250, 500, and 1000 occurences of "5" and "6" in the dataset, the top graph is for 250, then 500 and at the bottom 1000.