Machine Learning Coursework 2

Jesus E. Garcia Condado

February 20, 2017

I, Jesus E. Garcia, pledge that this assignment is completely my own work, and that I did not take, borrow or steal work from any other person, and that I did not allow any other person to use, have, borrow or steal portions of my work. I understand that if I violate this honesty pledge, I am subject to disciplinary action pursuant to the appropriate sections of Imperial College London.

1 Problem 1

We aim to find a high probability bound of the test error $R(g)$ of any ERM hypothesis g selected from n i.i.d. data points. We will start by defining the empirical test error on n samples for a given hypothesis g, or training error $\hat{R}_n(g)$ as:

$$\hat{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(g(x_i) \neq f(x_i)) \tag{1}$$

Then if there exists a perfect hypothesis h^* such that $\forall i.h^*(x_i) = x_i$ then:

$$\hat{R}_n(h^*) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(h^*(x_i) \neq f(x_i)) = 0 \tag{2}$$

Then since g is an ERM hypothesis for which $g \in argmin\hat{R}_n$, and by the definition in equation 1 we have that $\hat{R}_n(h) \geq 0$, then we can conclude that:

$$\hat{R}_n(g) = 0. \tag{3}$$

This implies that there is an ideal solution and that it will be picked by the ERM.Then defining then the test error for any hypothesis h:

$$R(h) = \Pr(h(x_i) \neq f(x_i)) \tag{4}$$

For any $\epsilon > 0$ and any hypothesis h such that $R(h) > \epsilon$ we have that this h can be selected as an empirical risk minimizer with probability at most $(1-\epsilon)^n$. This follows from the fact that for the selected hypothesis h, $\hat{R}_n(h)$ must equal 0 to satisfy the proven equation 3. This means that $\forall i \in \{0..n\}.h(x_i) = f(x_i)$. Then since $R(h) > \epsilon$ we can state by equation 4 that $\Pr(h(x_i) \neq f(x_i)) > \epsilon$

and since all points x_i are i.i.d, the probability that there does not exist a point $h(x_i) \neq f(x_i)$ is at most $(1-\epsilon)^n$. This follows from the fact that the joint distribution of i.i.d. random variables is the multiplication of its marginal distributions. If there was to exists a point x_i sucht that $h(x_i) \neq f(x_i)$, then equation 3 would not hold since the training error for such hypothesis is greater than 0. We can therefore conclude that:

$$\Pr(R(h) > \epsilon) \leq (1-\epsilon)^n \tag{5}$$

We aim to bound $R(g)$ so we must take into account all possible hypothesis. Let the number of such posible hypothesis be $|\mathcal{H}|$. Then it follows from the union bound $(\Pr(a \cup b) \leq \Pr(a) + \Pr(b))$ and equation 5 that:

$$\begin{aligned}
\Pr(R(g) > \epsilon) &= \Pr(\bigcup_{h \in \mathcal{H}} R(h) > \epsilon) \\
&\leq \sum_{h \in \mathcal{H}} \Pr(R(h) > \epsilon) \\
&= |\mathcal{H}| \Pr(R(h) > \epsilon) \\
&\leq |\mathcal{H}|(1-\epsilon)^n
\end{aligned} \tag{6}$$

Let $\delta = |\mathcal{H}|(1-\epsilon)^n$. Using the bound $(1-\epsilon)^n \leq e^{-\epsilon n}$, we can define a lower bound on ϵ:

$$\delta = |\mathcal{H}|(1-\epsilon)^n \tag{7}$$

$$\delta \leq |\mathcal{H}|e^{-\epsilon n} \tag{8}$$

$$\log \delta \leq \log\big(|\mathcal{H}|(e^{-\epsilon n})\big) = log\big(|\mathcal{H}|\big) - \epsilon n log(e) \tag{9}$$

$$\frac{-1}{n} \log \frac{\delta}{|\mathcal{H}|} \geq \epsilon \tag{10}$$

$$\epsilon \leq \frac{\log \frac{|\mathcal{H}|}{\delta}}{n} \tag{11}$$

Then using equation 6, we can state that $R(g) > \epsilon$ with probability of at most δ, or taking the complement: $R(g) \leq \epsilon$ with probability of at most $1-\delta$. Using the bound of equation 11 we can conclude that with probability of at most $1-\delta$:

$$R(g) \leq \frac{\log \frac{|\mathcal{H}|}{\delta}}{n} \tag{12}$$

2 Problem 2

The VC dimension of a hypothesis family $\mathcal{H}$, is the largest number of points $\mathcal{H}$ shatters. $\mathcal{H}$ shatters n points if the maximum number of dichotomies for the classification of such n points with $\mathcal{H}$ is 2^n, which is denoted as $|\mathcal{H}(x_1, x_2, ..., x_n)| = 2^n$. This implies that if and only if $\mathcal{H}$ shatters n then $m_{\mathcal{H}(n)} = 2^n$. The growth function $m_{\mathcal{H}(n)}$ of $\mathcal{H}$ is defined as the greatest number of dichotomies possible on any n points. Therefore $m_{\mathcal{H}(n)} = \max_{x_1,x_2,...,x_n \in \mathcal{X}} |\mathcal{H}(x_1, x_2, ..., x_n)|$.

For the hypothesis family $\mathcal{H}$ defined as:

$$\mathcal{H} = \{h : \mathbb{R}^2 \to \{0,1\} : h = \mathbb{I}(x \in R(a_1, a_2, b_1, b_2)) \text{ for some } a_1 \leq a_2, b_1 \leq b_2\} \tag{13}$$

We can show that the VC dimension of $\mathcal{H}$ is 4 by showing it shatters 4 points but not 5. By the definition of shattering, it is enough to find a single set of 4 points for which there exists $2^4 = 16$ dichotomies to prove that $\mathcal{H}$ shatters 4 points. Let $\mathcal{X} = \{(1,0),(0,1),(2,1),(1,2)\}$ be the set of 4 points in the 2 dimensions x_1 and x_2. Then we can show that the 16 dichotomies are possible by finding the values of a_1, a_2, b_1, b_2 that make each dichotomy possible. Splitting these 16 dichotomies into different cases we have:

- 4 dichotomies with one point in one set and the rest in another. For any of the 4 points letting there coordinates be (x_1, x_2) the values $\{a_1 = x_1 - 0.5, a_2 = x_1 + 0.5, b_1 = x_2 - 0.5, b_2 = x_2 + 0.5\}$ obtain these dichotomies.
- 6 dichotomies with 2 points in each set. Possible values for these dichotomies are:

a_1	a_2	b_1	b_2
0	2	0.5	1.5
0.5	1.5	0	2
0	1	0	1
1	2	0	1
0	1	1	2
1	2	1	2

Table 1: Dichotomies with 2 points in each set

- 4 dichotomies with 3 points in each set. Possible values for these dichotomies are:

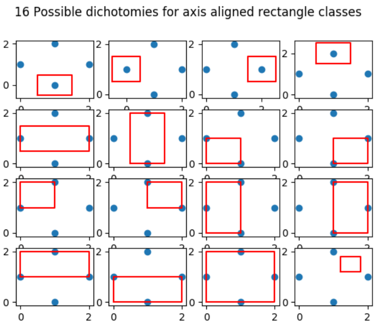


Figure 1: Graphical representation of the proof that the class of axis aligned rectangles shatters 4 points

a_1	a_2	b_1	b_2
0	1	0	2
1	2	0	2
0	2	1	2
0	2	0	1

Table 2: Dichotomies with 3 points in each set

- One dichotomy with all points included: $\{a_1 = 0, a_2 = 2, b_1 = 0, b_2 = 2\}$
- One dichotomy with no points included: $\{a_1 = 1.2, a_2 = 1.8, b_1 = 1.2, b_2 = 1.8\}$

All of possible dichotomies can be observed in figure 1

To prove that $\mathcal{H}$ does not shatter 5 points we must prove that for any possible set of 5 points there is at least one of the possible $2^4 = 32$ dichotomies which is not possible. Let x_l, x_r be the x_1 coordinate of leftmost and rightmost points and x_{top}, x_{bot} the x_2 coordinate of the highest and lowest points respectively out of all the points. We shall consider the case where 4 points are classified under the same class, which falls within the rectangle. It is therefore irrelevant if two of the points are at the same time the leftmost, rightmost, highest and lowest, since for all 4 points their coordinates x_1 and x_2 satisfy $a_1 \leq x_1 \leq a_2$ and $b_1 \leq x_2 \leq b_2$. This implies that

$$a_1 \leq x_l, x_r \leq a_2, b_1 \leq x_{bot}, x_{top} \leq a_2$$

By definition of $x_l, x_r, x_{top}, x_{bot}$ we have that the fifth point's coordinates f_1 and f_2 satisfy $x_l \leq x_1 \leq x_r$ and $x_{bot} \leq f_2 \leq x_{top}$. It therefore follows that they also satisfy $a_1 \leq f_1 \leq a_2$ and $b_1 \leq f_2 \leq b_2$. Consequently the fifth point will always be classified as the other 4. The same proof applies for any number of points greater than 5 as well, since a rightmost leftmost, lowest and highest point can always be defined, and the inner points can not be shattered. This discards one of the possible 32 dichotomies and means that $\mathcal{H}$ does not shatter 5 or more points. Since $\mathcal{H}$ does shatter 4 points we can conclude that the VC dimension of $\mathcal{H}$ is 4.

3 Problem 3

Let $X = \{x_1, x_2, ..., x_n\} \subset \mathcal{X}$ be any n sized data set with $x_1 \neq x_2$ with base set $\mathcal{X}$. Let $\mathcal{K}$ be a $k-1$ sized set of unique points in $\mathcal{X}$ and therefore $\mathcal{K} \subset \mathcal{X}$. We can then define the hypothesis class $\mathcal{H}$ as a class of classifying functions, were any point is classified as 1 if it is in $\mathcal{K}$ and -1 otherwise:

$$\mathcal{H} = \{h : x \in \mathcal{X} \to \{-1, 1\} : (x \in \mathcal{K}) \to 1 \wedge \neg(x \in \mathcal{K}) \to -1 \text{ for some } \mathcal{K} \subset \mathcal{X}, |\mathcal{K}| = k\} \tag{14}$$

The classifier therefore derives the class of a point by finding if it is one of the $k-1$ points defined in the classifier. Given that $x_1 \neq x_2$, we have that by definition of the hypothesis class, at most $k-1$ points can be classified as 1. If a single point can not be changed from a classification of 1 to a classification of -1 without altering the rest of the classifications, then not all possible dichotomies are possible. Hence we can bound the growth function as $m_{\mathcal{H}(n)} < 2^k$, since only $k-1$ points can take both values. Also the maximum number of dichotomies is the sum of all possibilities of choosing 0 to $k-1$ numbers from X, since with the $k-1$ points defined in the classifier any point can be or not be chosen up to a total of $k-1$. Since X has n points this can be expressed as:

$$|\mathcal{H}(x_1, x_2, ..., x_n)| = \sum_{i=0}^{k-1} \binom{n}{i} \tag{15}$$

4 Problem 4

4.1 Bound (a)

We wish to obtain bounds on the growth function $m_{\mathcal{H}(n)}$. We will first proof that:

$$m_{\mathcal{H}(n)} \leq n^{d_{VC}} + 1 \leq (n+1)^{d_{VC}} \tag{16}$$

The second inequality can be proven from the fact that since $(n+1)^d = n^d + n^{d-1} + n^{d-2} ... + 1$ then $(n+1)^d \geq n^d + 1$ because n and d are positive integers. To prove the first inequality we will first proof it for the case where $d_{VC} \leq n$. We will use Sauer's lemma which states that:

$$m_{\mathcal{H}(n)} \leq \sum_{i=0}^{d_{VC}} \binom{n}{i} \tag{17}$$

Hence we shall proof the first part of the inequality for the case where $d_{VC} \leq n$ by proving that for all n:

$$\forall d_{VC} \leq n. \sum_{i=0}^{d_{VC}} \binom{n}{i} \leq n^{d_{VC}} + 1 \tag{18}$$

Base case: For $n = 1$ we have that since $d_{VC} \leq n$ either $d_{VC} = 0$ and therefore $\binom{1}{0} = 1 \leq 1^0 + 1$ or $d_{VC} = 1$ and therefore $\binom{1}{0} + \binom{1}{1} = 1 + 1 \leq 1^1 + 1$.

Recursion: Using Pascal's triangle equation $\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$, the already proven inequality $(n+1)^d \geq n^d + n^{d-1} + 1$ and assuming that equation 18 holds for any n we can proof that it holds for any $n+1$:

$$\begin{aligned}
\sum_{i=0}^{d_{VC}} \binom{n+1}{i} &= 1 + \sum_{i=1}^{d_{VC}} \binom{n+1}{i} \\
&= 1 + \sum_{i=1}^{d_{VC}} \binom{n}{i} + \binom{n}{i-1} \\
&= 1 + \sum_{i=1}^{d_{VC}} \binom{n}{i} + \sum_{i=1}^{d_{VC}} \binom{n}{i-1} \\
&= \sum_{i=0}^{d_{VC}} \binom{n}{i} + \sum_{i=0}^{d_{VC}-1} \binom{n}{i-1} \\
&\leq n_{VC}^{d} + 1 + n^{d_{VC}-1} + 1 \\
&\leq (n+1)^{d_{VC}} + 1
\end{aligned} \tag{19}$$

We will now proof the case where $d_{VC} > n$. Let us start with the definition of d_{VC}:

$$d_{VC} = \max\{n : m_{\mathcal{H}}(n) = 2^n\} \tag{20}$$

Then since by definition $m_{\mathcal{H}(n)} \leq 2^n$ we have that if $d_{VC} < \infty$ then:

$$\forall n < d_{VC}. m_{\mathcal{H}}(n) < 2^n \tag{21}$$

Using the binomial theorem:

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i} \tag{22}$$

We can let $x = y = 1$ to proof that:

$$\forall n < d_{VC}. m_{\mathcal{H}}(n) < 2^n = \sum_{i=0}^{n} \binom{n}{i} \tag{23}$$

The same inductive proof used for equation 18 can now be applied to equation 23, since for the proof of equation 18 we required that the summation was over a number up to n, and we are summing up to n. We therefore have that since n is a positive integer and $d_{VC} > n$:

$$\forall d_{VC} > n . m_{\mathcal{H}}(n) < n^n + 1 \leq n^{d_{VC}} + 1 \tag{24}$$

Having provided a proof for the first inequality of equation 4.1 for both the cases where $d_{VC} > n$ and $d_{VC} \leq n$ we can conclude that the LHS holds for all n and d provided that $d_{VC} < \infty$ since this is a condition for the proof of the case $d_{VC} > n$.

4.2 Bound (b)

We will now proof a second bound on the growth function:

$$\forall n \geq d_{VC} . m_{\mathcal{H}}(n) \leq (\frac{ne}{d_{VC}})^{d_{VC}} \tag{25}$$

Given that $n \geq d$ we have that $\frac{n}{d_{VC}} \geq 1$ and therefore for any $d_{VC} - i \geq 0$ we have that $(\frac{n}{d_{VC}})^{d_{VC}-i} \geq 1$. Using Sauer's lemma shown in equation 17 and given that $d_{VC} \geq i$ we have proven the following equation:

$$m_{\mathcal{H}}(n) \leq \sum_{i=0}^{d_{VC}} \binom{n}{i} \leq \sum_{i=0}^{d_{VC}} \binom{n}{i} \left(\frac{n}{d_{VC}}\right)^{d_{VC}-i} \tag{26}$$

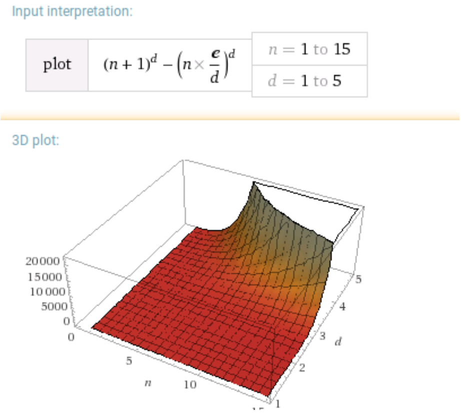
Breaking down $(\frac{n}{d_{VC}})^{d_{VC}-i}$ into $(\frac{n}{d_{VC}})^{d_{VC}}(\frac{n}{d_{VC}})^{-i} = (\frac{n}{d_{VC}})^{d_{VC}}(\frac{d_{VC}}{n})^{i}$ we can take the term $(\frac{n}{d_{VC}})^{d_{VC}}$ out of the summation since it does not contain the summation variable.

$$m_{\mathcal{H}}(n) \leq \left(\frac{n}{d_{VC}}\right)^{d_{VC}} \sum_{i=0}^{d_{VC}} \binom{n}{i} \left(\frac{d_{VC}}{n}\right)^{i} \tag{27}$$

Now we will use the binomial theorem of equation 22 with $x = (\frac{d_{VC}}{n})$ and $y = 1$. We also have that $\binom{n}{i} \geq \binom{d_{VC}}{i}$ which follows from induction on the Pascal Triangle equation and the fact that $n \geq d_{VC}$. Then:

$$\sum_{i=0}^{d_{VC}} \binom{n}{i} \left(\frac{d_{VC}}{n}\right)^{i} \leq \sum_{i=0}^{d_{VC}} \binom{d_{VC}}{i} \left(\frac{d_{VC}}{n}\right)^{i} = \left(1 + \frac{d_{VC}}{n}\right)^{d_{VC}} \tag{28}$$

Applying this to equation 27, we can continue our proof by using the fact that $(1 + \frac{1}{x})^x \leq e$. Taking logarithms of both sides, since the logarithm is a monotonically increasing function, and multiplying both sides by a variable y we


Input interpretation:

plot	$(n+1)^d - \left(n \times \frac{e}{d}\right)^d$	$n = 1$ to 15 $d = 1$ to 5

3D plot:

Figure 2: Difference in the bounds of equations 4.1 and 4.2 as a function of n and d_{VC}

have that $y \log\left((1+\frac{1}{x})^x\right) \leq \log(e)y$ and therefore $(1+\frac{1}{x})^{xy} \leq e^y$. Let $x = \frac{n}{d_{VC}}$ and $y = d_{VC}$ then $(1+\frac{d_{VC}}{n})^n \leq e^{d_{VC}}$. Therefore

$$m_{\mathcal{H}}(n) \leq \left(\frac{n}{d_{VC}}\right)^{d_{VC}} \left(1+\frac{d_{VC}}{n}\right)^{d_{VC}} \leq \left(\frac{n}{d_{VC}}\right)^{d_{VC}} e^{d_{VC}} = \left(\frac{ne}{d_{VC}}\right)^{d_{VC}} \tag{29}$$

4.3 Bound comparison

Both bounds grow exponentially with an increase in VC dimensions and linearly with an increase in n. However the second term is a tighter bound since it also has a coefficient of $\frac{1}{d_{VC}}^{d_{VC}}$. The difference in the bound becomes significant very quickly as can be seen in figure 2 which plots the difference between the two bounds.

5 Problem 5 (Structural Risk Minimization

5.1 Part a

Consider the set of classifiers $\mathcal{H}_q = \{h_p | p : \mathbb{R} \to \mathbb{R}\text{is a polynomial of degree at most}q\}$. Defining the set $\mathcal{X} = [0, 2.5] \times [-1, 2]$ we can define the transformation $\phi(X)$, where $X = (x_1, x_2) \in \mathcal{X}$, to a new set $\mathcal{Z}$ as $Z = \phi_q(X) = (x^0, x_1^1, x_1^2, ...x_1^q, x_2) \in \mathcal{Z}$. This includes adding the x_0 dimension necessary for linear classification. Then using a linear classifier h_w in $\mathcal{Z}$ such that $h_w(Z) = sign(w^T Z)$ and $w = [w_0, w_1, w_2...w_{q+1}]$ we have that $w^T Z = w_0 x^0 + w_1 x_1^1 + w_2 x_1^2, ... + w_q x_1^q + w_{q+1} x_2$. Then $sign(w^T Z) = 1$ if $x_2 \geq \frac{-1}{w_{q+1}}(w_0 x^0 + w_1 x_1^1 + w_2 x_1^2, ... + w_q x_1^q)$. We can therefore find a w such that h_w is a classifier with $h_w(Z) = 1$ if $x_2 \geq p(x_1)$ where p was defined in $\mathcal{H}_q$.

For this $\mathcal{H}_q$ we can prove that the d_{VC} is at most $q+1$. Let us define a vector w_{new} derived from the weight of our perceptron as $w_{new} = \frac{-1}{w_{q+1}}(w_0, w_1, w_2, ..., w_q)$ and a vector x^i_{pow} of the powers of any point x^i as $x^i_{pow} = (x^0, x^1_1, x^2_1, ..., x^q_1)$. Then for a point $x^{(i)}$ and a hypothesis $h_p \in \mathcal{H}_q$ we will have that $h_p(x^i) = 1$ if $x^i_2 \geq w^T_{new} x^i_{pow}$. Then if we have $q+2$ points, since the points have $||x^i_{pow}|| = q+1$ dimensions, we can express one of the $q+2$ points as a linear combination of the other $q+1$ points. Let us pick the point $x^{(min)}$ such that it has the smallest x_2 coordinate and therefore $\forall i \in \{1..q+2\} x^{min}_2 \leq x^{(i)}_2$. Therefore for some coefficients a_i we have $x^{(min)}_{pow} = \sum_{i \in 0..q+2 \wedge i \neq min} a_i x^i_{pow}$. Then if we consider the dichotomy where all the points except for $x^{(min)}$ are labeled 1, that is $\forall i \in \{0...q+2\}. \wedge i \neq min. x^i_2 \geq w^T_{new} x^i_{pow}$ we have that the $x^{(min)}$ point will be labeled 1 if the following holds:

$$x^{(min)}_2 \geq w_{new} x^{(min)}_{pow} = w_{new} \sum_{i \in 0..q+2 \wedge i \neq min} a_i x^i_{pow} \geq \sum_{i \in 0..q+2 \wedge i \neq min} a_i x^{(min)}_2 \tag{30}$$

Where the last step follows from the fact that we chose $x^{(min)}$ to have the smallest x_2 coordinate. Therefore the point will be labeled as 1 if the sum of all coefficients a_i is smaller than 1. Either if it is or it is not, the inequality is independent of w and therefore the last point can not be shattered. We can therefore conclude that for $\mathcal{H}_q$ the d_{VC} is at most $q+1$ since it can not shatter $q+2$ points.

We can then proof that the d_{VC} is $q+1$ by proving that it shatters $q+1$ points but not $q+2$. The latter has already been proved. Let x_{pow} be the matrix containing all $x^(i)_{pow}$ for any i given there are $q+1$ points. Let x_2 be another matrix of $q+1 \times 1$ containing all the second dimensions $x^(i)_2$ of all $q+1$ points. Then if we can find an w such that $x_i = w^t x_{pow}$ we can shatter $q+1$ points. This is possible since there exists a combination of linearly independent $q+1$ vectors, by making one different entry in each vector 1 out of its $q+1$ dimensions and apart from the always 1 dimension 0. Since they are linearly independent the inverse exists and we can shatter the points by choosing $w^t = x_i x^{-1}_{pow}$.

5.2 Part b

Using the transform and linear classifier described in the previous section we can now use structural risk minimization to approximate $f(x) = X(x-1)(x-2)$ from a dataset with the described noisy distribution (follows the correct labeling h_f with a probability of 90%). To do so we will have to define the complexity of our classifier $\mathcal{H}_q$. We will use the given equation equation in terms of the growth function and the bound from equation 25, since we will use $n \geq 10$ and therefore $n \geq d_{VC}$:

$$\Omega(n,\mathcal{H},\delta) = \sqrt{\frac{8}{n}\log\frac{4m_{\mathcal{H}}(2n)}{\delta}} \leq \sqrt{\frac{8}{n}\log\frac{4(\frac{2ne}{d_{VC}})^{d_{VC}}}{\delta}} = \sqrt{\frac{8d_{VC}}{n}\log\left(\frac{2ne}{d_{VC}}\right)+\frac{8}{n}\log\frac{4}{\delta}} \tag{31}$$

By then selecting the hypothesis $g = argmin_{h\in\mathcal{H}}\hat{R}_n(h)+\Omega(n,\mathcal{H}(h),w_i\delta)$ we can have a bound on our test error with probability $1-\delta$ of:

$$R(g) \leq min_i R(h^*) + 2 * \Omega(n,\mathcal{H}_i,w_i\delta) \tag{32}$$

Where $h^* = argmin_{h\in\mathcal{H}_i}R(h_i)$. This algorithm was run with equal weightings for all $\mathcal{H} = \bigcup_{q=0}^{4}\mathcal{H}_q$ and to a confidence of 90%, that is $w_i\delta = 0.2 * 0.1 = 0.02$.

To calculate the test error 1000 new data points of the same distribution were used. Since we know that our test error will be of at least 10% due to the noise, this was deemed a reasonable number. This is so because the points are uniformly distributed over the area. Therefore points are expected to be misclassified either by noise or by falling in the wrong area with a probability proportional to such area. Since there are 1000 data points this gives as an accuracy of $1/1000 = 0.001$ which is substantially more precise than the minimum expected 10% error.

Since the experiments are random they were run multiple times to take the average training and test error. This was done both for the overall SRM algorithm and for the individual ERM algorithms. The resulting training and test errors for the SRM with different number of test samples were:

n train samples	Training error	Test Error
10	15%	25.7%
100	21.3%	19.8%
10000	15.8%	15.9%

Table 3: Training and test errors of SRM

The rest of the individual ERM errors and the percentage of runs that a given hypothesis was chosen can be observed in figure 3.

The results show how SRM aims not to over-fit a data set with very few points by having the complexity penalty. This can be observed in the run with $n = 10$, since the correct order of magnitude is never chosen. While this results in a higher test error, if we were not to know the actual distribution of our data and it would be a different distribution, this pick might have been correct. If we change the weights of the complexity to benefit the correct complexity then we are incurring in data snooping. As the number of points increase the SRM results in a more complex hypothesis being chosen and the test error decreases.

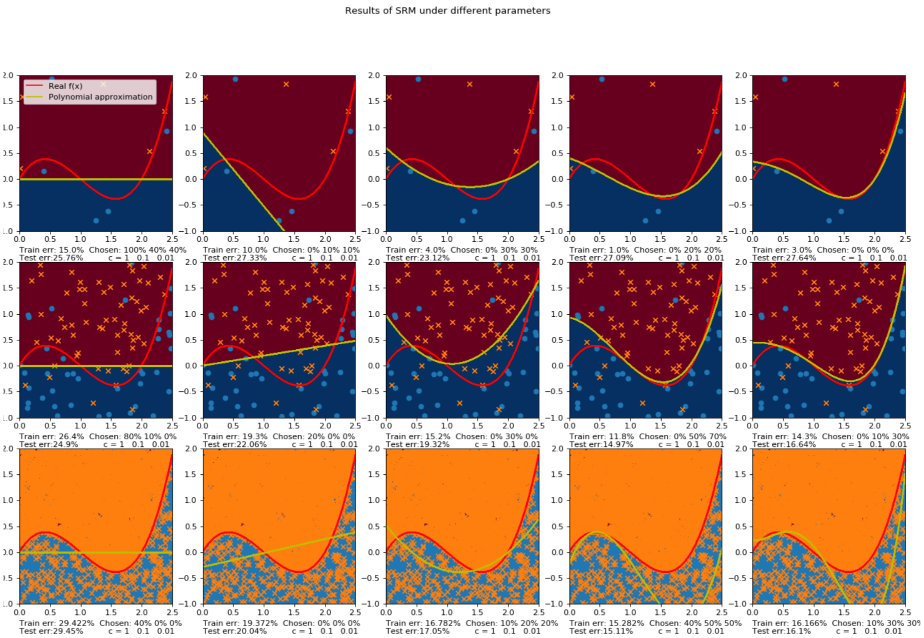


Figure 3: Output from the SRM and the implicit ERM for values of n of 10, 100, 10000 (rows) and for polynomial hypothesis classes from degree 0 to degree 4 (columns). Each subplot represents the plotted outcome of the first of 10 experiments done. The average training and test errors are taken over the 10 experiments run. The chosen percentages reflect out the percentage out of the 10 experiments for which that complexity class was chosen. The three values correspond to the different coefficients for the complexity as shown. For the last row with $n = 10000$ there seem to be more misclassified orange crosses, but this is only because the orange crosses are plotted after the blue dots. There are so many points that they overlap and given the plotting order orange predominates

However we can observe that the correct one is still not mostly chosen. Only for $n = 10000$ it is chosen and only 40% of the runs.

5.3 Part c

Since the bound in equation 5.2 is quite loose, SRM penalizes complexity in excess. This can b observed in the the runs with 10 and 100 samples. The former chooses always the lowest complexity and the latter only improves this by choosing the next complexity a 20% of the runs. By rerunning the experiments with a coefficient for the complexity of 0.1 and 0.01 we can observe improvements in the final training and test errors. This is also reflected in the fact that the correct complexity is chosen more often. A coefficient of 0.01 correctly choses the complexity in 70% of the runs for $n = 100$. 4 shows the resulting training and test errors for different coefficients.

n train samples	c	Training error	Test Error
10	1	15%	25.7%
10	0.1	1%	25.5%
10	0.01	1%	25.5%
100	1	21.3%	19.8%
100	0.1	12.3%	16.4%
100	0.01	11.4%	14.5%
10000	1	15.8%	15.9%
10000	0.1	14.10%	14.13%
10000	0.01	14.10%	14.13%

Table 4: Training and test errors of SRM for different complexity coefficients