Paul Lintilhac

Foundations of Machine Learning

Homework 2

A: Rademacher Complexity

1)

First, note that if there is only one hypothesis in H, then we can remove the sup{} operator, since sup{const} = const. Then we are starting with the following equation for the Rademacher Complexity:

$$\mathfrak{R}'(\{h\}) = \frac{1}{m} E\left[\left|\sum_{i=1}^{m} \sigma_i h(x_i)\right|\right] = \frac{1}{m} E\left[\sqrt{\left(\sum_{i=1}^{m} \sigma_i h(x_i)\right)^2}\right]$$

Since $\sqrt{}$ is a convex function, we can apply Jensens inequality to write

$$\mathfrak{R}'(\{h\}) \leq \frac{1}{m} \sqrt{E\left[\left(\sum_{i=1}^{m} \sigma_i h(x_i)\right)^2\right]}$$

Now we expand the squared sum by writing it as a double summation,

$$E\left[\left(\sum_{i=1}^{m} \sigma_i h(x_i)\right)^2\right] = E\left[\sum_{i,j=1}^{m} \sigma_i \sigma_j h(x_i) h(x_j)\right] = \sum_{i,j=1}^{m} E\left[\sigma_i \sigma_j h(x_i) h(x_j)\right]$$

Note that since σ_i are Rademacher variables, all σ_i are independent of each other when $i \neq j$ (and also independent of $h(x_i)$). Therefore $E\left[\sigma_i \sigma_j h(x_i) h(x_j)\right] = E[\sigma_i] E[\sigma_j] E\left[h(x_i) h(x_j)\right]$. Since $E[\sigma_i] = 0$, this means that all terms are zero except when $i = j$. Note that $E[{\sigma_i}^2] = 1$, since ${\sigma_i}^2$ is always equal to 1. Then we have

$$\mathfrak{R}'(\{h\}) \leq \frac{1}{m} \sqrt{\sum_{i,j=1}^{m} E[{\sigma_i}^2] E[h(x_i)^2]} = \frac{1}{m} \sqrt{\sum_{i,j=1}^{m} E[h(x)^2]} = \frac{1}{m} \sqrt{\sum_{i,j=1}^{m} E[h(x)^2]}$$

$$= \frac{1}{m} \sqrt{m E[h(x)^2]} = \sqrt{\frac{E[h(x)^2]}{m}}$$

Which is what we wanted to show.

2.

Claim: $\mathfrak{R}_m(\alpha H) = |\alpha|\mathfrak{R}_m(H)$

Proof:

First, we will derive an expression for the empirical Rademacher complexity, $\widehat{\mathfrak{R}}_m(\alpha H)$

$$\mathfrak{R}_m(\alpha H) = E_{S,\sigma}\left[\sup_{h\in H}\{\frac{\boldsymbol{\sigma}\cdot(\alpha h_S)}{m}\}\right] = |\alpha|E_{S,\sigma}\left[\sup_{h\in H}\{\frac{sgn(\alpha)*\boldsymbol{\sigma}\cdot(h_S)}{m}\}\right]$$

Because the dot product, supremum, and expectation are all linear operators. Note that we have used the identity $\alpha = |\alpha| * sgn(\alpha)$. Since the distribution of $-\boldsymbol{\sigma}$ is the same as the distribution of $\boldsymbol{\sigma}$,

$$E_{S,\sigma}\left[\sup_{h\in H}\{\frac{sgn(\alpha)*\boldsymbol{\sigma}\cdot(h_S)}{m}\}\right] = E_{S,\sigma}\left[\sup_{h\in H}\{\frac{\boldsymbol{\sigma}\cdot(h_S)}{m}\}\right]$$

Therefore, $\mathfrak{R}_m(\alpha H) = |\alpha|E_{S,\sigma}\left[\sup_{h\in H}\left\{\frac{\boldsymbol{\sigma}\cdot(h_S)}{m}\right\}\right] = |\alpha|\mathfrak{R}_m(H)$

■

Claim: $\mathfrak{R}_m(H + H') \leq \mathfrak{R}_m(H) + \mathfrak{R}_m(H')$

Proof:

$$\mathfrak{R}_m(H + H') = E_{S,\sigma}\left[\sup_{h\in H+H'}\{\frac{\boldsymbol{\sigma}\cdot(\alpha h_S)}{m}\}\right] \leq E_{S,\sigma}\left[\sup_{h\in H}\{\frac{\boldsymbol{\sigma}\cdot(\alpha h_S)}{m}\} + \sup_{h\in H'}\{\frac{\boldsymbol{\sigma}\cdot(\alpha h_S)}{m}\}\right]$$

Where we have used the sub-additivity of the sup operator. Now by the linearity of expectation we have

$$\mathfrak{R}_m(H + H') \leq E_{S,\sigma}\left[\sup_{h\in H}\left\{\frac{\boldsymbol{\sigma}\cdot(\alpha h_S)}{m}\right\}\right] + E_{S,\sigma}\left[\sup_{h\in H'}\left\{\frac{\boldsymbol{\sigma}\cdot(\alpha h_S)}{m}\right\}\right] = \mathfrak{R}_m(H) + \mathfrak{R}_m(H')$$

■

Claim: $\mathfrak{R}_m(\{\max(h, h') : h \in H, h' \in H'\}) \leq \mathfrak{R}_m(H) + \mathfrak{R}_m(H')$

Proof:

First we note as in the homework handout that $\max(h, h') = \frac{1}{2}(h + h' + |h - h'|)$

$$\mathfrak{R}_m(\{\max(h, h') : h \in H, h' \in H'\}) = \frac{1}{2}\mathfrak{R}_m(H + H' + |H - H'|) \leq \frac{1}{2}\mathfrak{R}_m(H + H') + \frac{1}{2}\mathfrak{R}_m(|H - H'|)$$
$$\leq \frac{1}{2}\mathfrak{R}_m(H) + \frac{1}{2}\mathfrak{R}_m(H') + \frac{1}{2}\mathfrak{R}_m(|H - H'|)$$

Where I have used the results from the first two proofs above. Now, the absolute value function is lipschitz-continuous with a best lipschitz constant of 1. Therefore, according to the contraction lemma, we have

$$\mathfrak{R}_m(|H - H'|) \leq \mathfrak{R}_m(H - H') \leq \mathfrak{R}_m(H) + \mathfrak{R}_m(-H') = \mathfrak{R}_m(H) + \mathfrak{R}_m(H')$$

Note that the negative sign disappears because of the fact that we must take the absolute value of a constant multiplier inside the Rademacher function, as proven above. Putting this all together, we get

$$\begin{aligned}\mathfrak{R}_m(\{\max(h, h') : h \in H, h' \in H'\}) &< \frac{1}{2}\mathfrak{R}_m(H) + \frac{1}{2}\mathfrak{R}_m(H') + \frac{1}{2}\mathfrak{R}_m(H) + \frac{1}{2}\mathfrak{R}_m(H') \\ &= \mathfrak{R}_m(H) + \mathfrak{R}_m(H')\end{aligned}$$

■

B: VC-Dimension

1) It is easy to show that the VC dimension of this set is at least three. Without showing every possibility, we can see this intuitively because the first point can either be in $[-\infty, x)$ or $[x, x+1]$, the second point can either be in$[x, x+1]$ or $(x+1, x+2)$, and the third point can either be in $(x+1, x+2)$ or $[x+2, \infty)$, meaning that we can pick whether each point is positive or negative without changing the correct rank-ordering of the 3 points. However, we can see that the VC-Dimension cannot be 4, because we cannot produce the dichotomy (+,-,+,-). Therefore, the VC-dimension is 3.

   a. I will show that the set {x,2x,3x,4x} cannot be shattered by any of the sine functions in this family by showing that the dichotomy {+,+,-,+} cannot be realized. Note that if our first point is positive, i.e. $sgn(\sin(\omega x)) = 1$, then we have $x \ mod\ \frac{2\pi}{\omega} \in (0, \frac{\pi}{\omega})$. Our second point being positive implies that $2x \ mod\ \frac{2\pi}{\omega} \in (0, \frac{\pi}{\omega})$. Since gcd $\left(\frac{2\pi}{\omega}, 2\right) = 2$, we can rewrite this as $x \ mod\ \frac{\pi}{\omega} \in (0, \frac{\pi}{2\omega})$. Combining this with the previous equation, we have that $x \ mod\ \frac{2\pi}{\omega} \in (0, \frac{\pi}{2\omega})$. The third point being negative implies that $3x \ mod\ \frac{2\pi}{\omega} \in \left(\frac{\pi}{\omega}, \frac{2\pi}{\omega}\right)$, or equivalently that $x \ mod\ \frac{2\pi}{\omega} \in \left(\frac{\pi}{3\omega}, \frac{2\pi}{3\omega}\right)$. Combining this with the previous equation, we have that $x \ mod\ \frac{2\pi}{\omega} \in \left(\frac{\pi}{3\omega}, \frac{\pi}{2\omega}\right)$. Now, multiplying both sides of the above equation by 4, we get $4x \ mod\ \frac{2\pi}{\omega} \in \left(\frac{4\pi}{3\omega}, \frac{2\pi}{\omega}\right)$. But this implies that $sgn(\sin(\omega x)) = -1$, contradicting the fact that 4x is positively labeled. Thus the dichotomy {+,+,-,+} cannot be realized for any set {x,2x,3x,4x}.

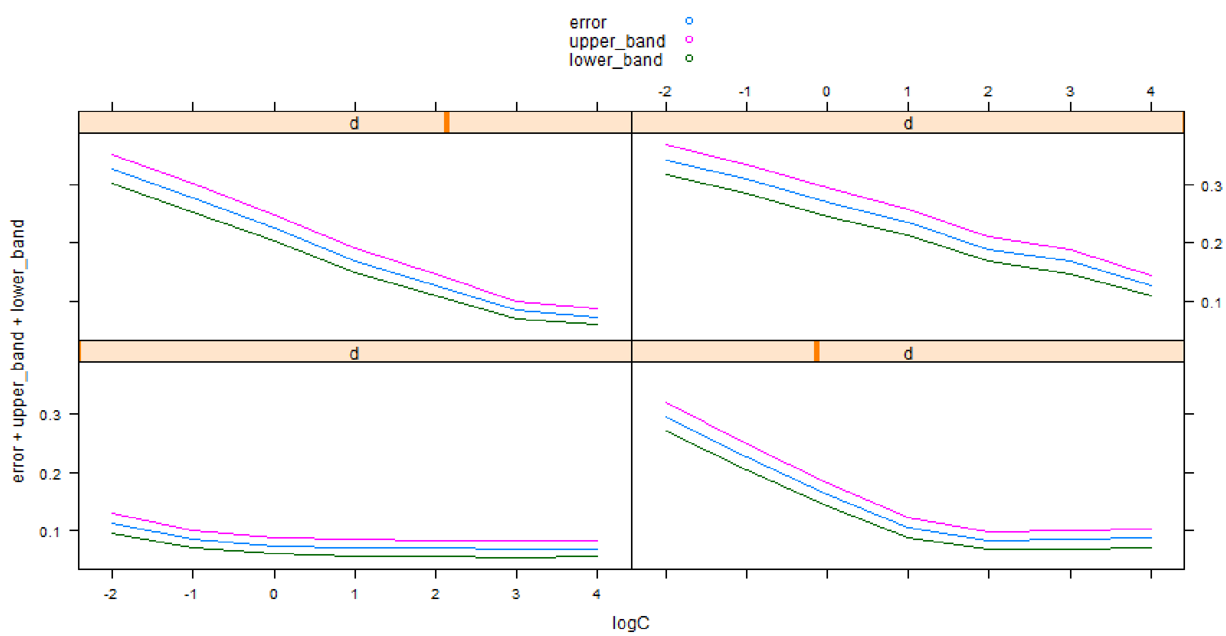
   b. In order to prove that the set $\{2^{-m} : m \in \{1, \ldots, N\}\}$ can be fully shattered by the family of functions $\sin(\omega x)$, note that for any m, we need $sign(\sin(\omega * 2^{-m})) = y_m$, where $y_m(x)$ is the label of the point at $x = 2^{-m}$. It is easy to see that $sign(\sin(\omega * 2^{-m})) = 1$ whenever $floor\left(\frac{2^{-m}\omega}{\pi}\right) = 0 \ mod\ 2$, and $sign(\sin(\omega * 2^{-m})) = -1$ whenever $floor\left(\frac{2^{-m}\omega}{\pi}\right) = 1 \ mod\ 2$. I claim that we can accomplish this if we define $\omega \equiv \pi\left(\mathbf{1}_{y_N=1}\mathbf{1}_{y_{N-1}=1} \ldots \mathbf{1}_{y_1=1}\right)$, (i.e. the binary expansion of ω given by concatenating the values of y_i, with values of -1 replaced with 0). To see this, note that $\frac{2^{-m}\omega}{\pi} = \left(2^{-m}\mathbf{1}_{y_N=1}\mathbf{1}_{y_{N-1}=1} \ldots \mathbf{1}_{y_1=1}\right) = \mathbf{1}_{y_N=1}\mathbf{1}_{y_{N-1}=1} \ldots \mathbf{1}_{y_m=1} \,.\, \mathbf{1}_{y_{m-1}=1}\mathbf{1}_{y_{m-2}=1} \ldots \mathbf{1}_{y_1=1}$. The floor value of this binary number mod 2 depends only on the units digit. If $y_m = 1$, then the units digit is 1, so $floor\left(\frac{2^{-m}\omega}{\pi}\right) = 1 \ mod\ 2$, while if $y_m = 0$, then $floor\left(\frac{2^{-m}\omega}{\pi}\right) = 0 \ mod\ 2$. Thus this definition of ω satisfies the conditions set out above for realizing any dichotomy of size N. Thus, the set $\{2^{-m} : m \in \{1, \ldots, N\}\}$ is fully shattered by $\sin\left(\pi\left(\mathbf{1}_{y_N=1}\mathbf{1}_{y_{N-1}=1} \ldots \mathbf{1}_{y_1=1}\right)x\right)$, so the VC dimension of $\sin(\omega x)$ is at least countably infinite.

C: Support Vector Machines

3)

For this problem, I tried two different approaches, once using the R extension for libsvm, and the other using the libsvm c/command line tools. Both approaches gave me results that did not match with my expectation, which was that there would be local minima for the error rate for all values of d. What I found was that actually only d=2 had a local minimum.

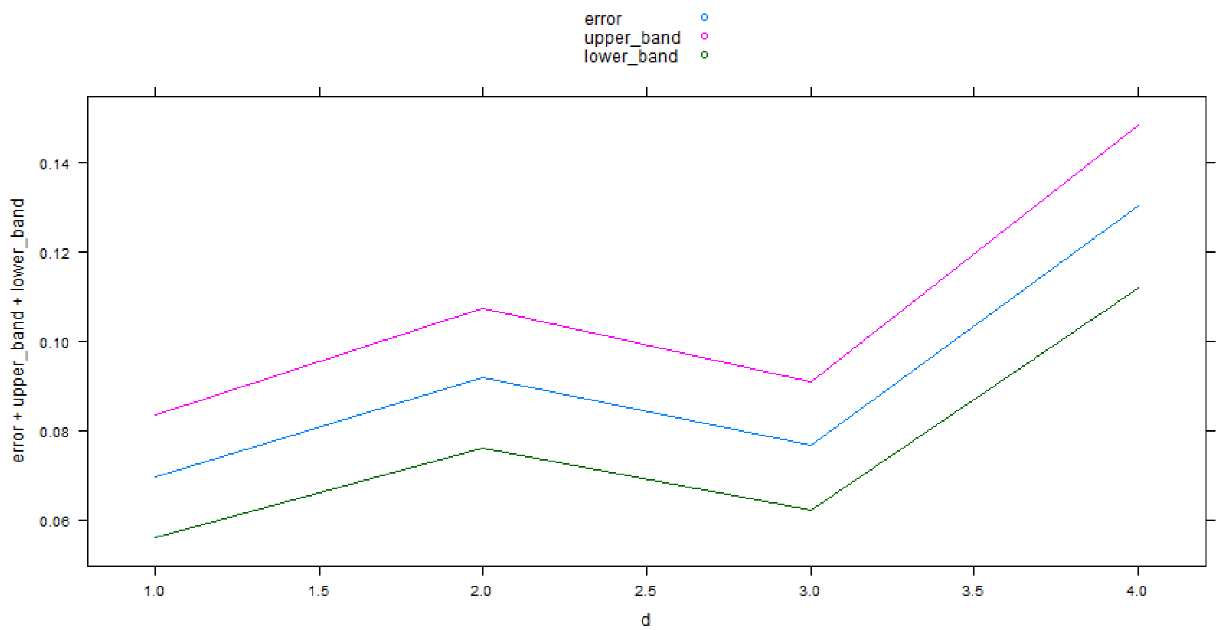
Figure 1:



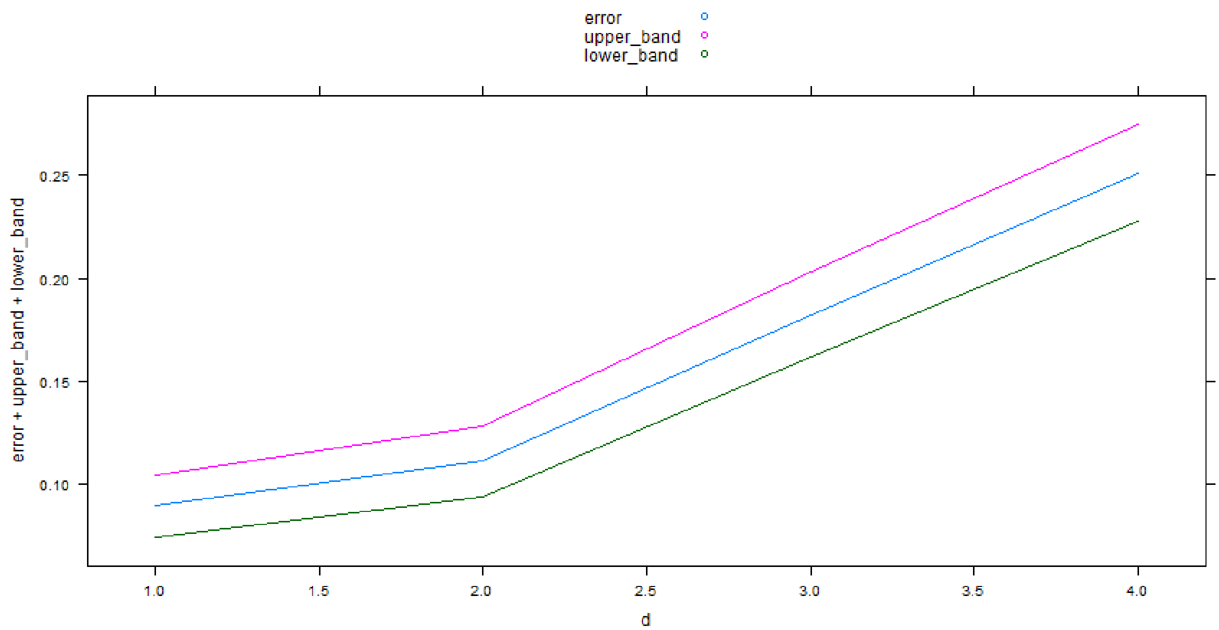
While the two different packages apparently disagree about which degree polynomial has the lowest error rate, they at least both agree on the fact that the only degree with a local minimum over C is d=2.

Therefore, for question 4 I will assume that the optimal pair of (C,d) is (2^2,2).

4.
Training (cross-validation) errors over d:



Testing errors:



One possible reason why the testing errors would increase with d is because of higher complexity in the hypothesis set, leading to the potential to overfit.

3.C)

Note that if we replace $|\boldsymbol{\xi}|_1$ $with$ $|\boldsymbol{\xi}|_\infty$ In the primal optimization problem, the lagrange equations change. The gradient with repsect to ξ_i is 0 for all $\xi_i \neq |\boldsymbol{\xi}|_\infty$, leading to the following alternative set of equations:

$$\alpha_i = \begin{cases} C - \beta_i \ when \ \xi_i = |\boldsymbol{\xi}|_\infty \\ -\beta_i \ when \ \xi_i \neq |\boldsymbol{\xi}|_\infty \end{cases}$$

Note that assuming $|\boldsymbol{\xi}|_\infty > 0$, the condition that $\xi_i\beta_i = 0$ says that $C = \alpha_{i'}$, where i' is the i such that $\xi_i = |\boldsymbol{\xi}|_\infty$. Note that because of the constraint that $\alpha_{i'} = 1 - \xi_{i'}$, this means that $\alpha_{i'}$ is a minimum over all α_i for which $\alpha_i \neq 0$. But since we have the constraint that $\beta_i \geq 0 \rightarrow \alpha_i \leq C = \alpha_{i'}$, this means that the constraint on α_i becomes $\alpha_i = 0 \cup \alpha_i = C$.