Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 2
October 25, 2017
Due: November 10, 2017

A. Growth function

Growth function of stump functions. For any $x \in \mathbb{R}$ and $\theta \in \mathbb{R}$, let ϕ_θ denote the threshold function that assigns sign $+1$ to $x \leq \theta$, -1 otherwise: $\phi_\theta(x) = 21_{x \leq \theta} - 1$. Let $\mathcal{H}$ be the family of functions mapping $\mathbb{R}^N$ to $\{-1, +1\}$ defined by

$$\mathcal{H} = \Big\{ \mathbf{x} \mapsto s\, \phi_\theta(x_i) \colon i \in [1, N], \theta \in \mathbb{R}, s \in \{-1, +1\} \Big\},$$

where x_i is the ith coordinate of $\mathbf{x} \in \mathbb{R}^N$.

1. Show that the following upper bound holds for the growth function of $\mathcal{H}$: $\Pi_m(\mathcal{H}) \leq 2mN$.

Solution: We start by noticing that, if $\mathcal{I}_1, \mathcal{I}_2$ are two families of functions mapping $\mathbb{R}^N$ to $\{-1, +1\}$, then

$$\Pi_m(\mathcal{I}_1 \cup \mathcal{I}_2) \leq \Pi_m(\mathcal{I}_1) + \Pi_m(\mathcal{I}_2). \tag{1}$$

This fact immediately follows by the definition of growth function. It follows by (1) that if

$$\mathcal{F}_i = \Big\{ \mathbf{x} \mapsto s\, \phi_\theta(x_i) \colon \theta \in \mathbb{R}, s \in \{-1, +1\} \Big\},$$

for $i = 1, \ldots, N$, then

$$\Pi_m(\mathcal{H}) \leq N \cdot \Pi_m(\mathcal{F}_1),$$

since clearly $\Pi_m(\mathcal{F}_i) = \Pi_m(\mathcal{F}_1)$ for $i = 1, \ldots, N$. Now we wanna show that $\Pi_m(\mathcal{F}_1) = 2m$. Be $\mathbf{x}^1, \ldots, \mathbf{x}^m \in \mathbb{R}^N$. W.l.o.g. we can assume that $N = 1$ (since the functions in $\mathcal{F}_1$ only depend on the first component) and that

$$x_1 \leq \cdots \leq x_m \tag{2}$$

(this is true up to reordering $x_1, \ldots, x_m$). Let's consider the set

$$\mathcal{S} = \{(s\phi_\theta(x_1), \ldots, s\phi_\theta(x_m)) \colon \theta \in \mathbb{R}, s \in \{-1, +1\}\}.$$

By (2) and by the definition of ϕ_θ it follows that

$$
\begin{aligned}
\mathcal{S} &\subseteq \left\{ \mathbf{v} = s \sum_{i=1}^{k} \mathbf{e}_i - s \sum_{i=k+1}^{m} \mathbf{e}_i \colon k = 0, \dots, m, s \in \{-1, +1\} \right\} \\
&= \left\{ \sum_{i=1}^{m} \mathbf{e}_i, -\sum_{i=1}^{m} \mathbf{e}_i \right\} \\
&\cup \left\{ \mathbf{v} = s \sum_{i=1}^{k} \mathbf{e}_i - s \sum_{i=k+1}^{m} \mathbf{e}_i \colon k = 1, \dots, m-1, s \in \{-1, +1\} \right\}
\end{aligned}
$$

where $\mathbf{e}_1, \dots, \mathbf{e}_m$ denotes the standard basis of $\mathbb{R}^m$. In particular the equality holds if $x_1 < \cdots < x_m$ and in this case $|\mathcal{S}| = 2 + 2(m-1) = 2m$. It follows that

$$
\Pi_m(\mathcal{F}_1) = 2m \quad \text{which implies that} \quad \Pi_m(\mathcal{H}) \leq 2mN.
$$

2. Let $\mathcal{H}_2$ be the family of functions mapping $\mathbb{R}^N$ to $\{-1, +1\}$ defined by

$$
\begin{aligned}
\mathcal{H}_2 = \Big\{ \mathbf{x} \mapsto s\, 1_{\phi_\theta(x_i)=+1}\, \phi_{\theta'}(x_j) + s'\, 1_{\phi_\theta(x_i)=-1}\, \phi_{\theta'}(x_j) \colon \\
i \neq j, i, j \in [1, N], \theta, \theta' \in \mathbb{R}, s, s' \in \{-1, +1\} \Big\}.
\end{aligned}
$$

Show that $\Pi_m(\mathcal{H}_2) = O(m^2 N^2)$. Give an explicit upper bound on $\Pi_m(\mathcal{H}_2)$.

Solution: We can write $\mathcal{H}_2$ as

$$
\mathcal{H}_2 = \bigcup_{i=1}^{N} \bigcup_{\substack{j=1 \\ j \neq i}}^{N} \mathcal{H}^{i,j}
$$

where

$$
\begin{aligned}
\mathcal{H}^{i,j} \doteq \Big\{ \mathbf{x} \mapsto s\, 1_{\phi_\theta(x_i)=+1}\, \phi_{\theta'}(x_j) + s'\, 1_{\phi_\theta(x_i)=-1}\, \phi_{\theta'}(x_j) \colon \\
\theta, \theta' \in \mathbb{R}, s, s' \in \{-1, +1\} \Big\}.
\end{aligned}
$$

Since $\Pi_m(\mathcal{H}^{i,j})$ does not depend on i, j it follows by (1) that

$$
\Pi_m(\mathcal{H}_2) \leq N(N-1) \Pi_m(\mathcal{H}^{1,2}).
$$

Moreover $\mathcal{H}^{1,2} = \mathcal{G}_1\mathcal{G}_2$ with

$$\mathcal{G}_1 \doteq \Big\{\mathbf{x} \mapsto s\, 1_{\phi_\theta(x_1)=+1} + s'\, 1_{\phi_\theta(x_1)=-1} \colon \theta \in \mathbb{R}, s, s' \in \{-1,+1\}\Big\}$$

and

$$\mathcal{G}_2 \doteq \Big\{\mathbf{x} \mapsto \phi_{\theta'}(x_2) \colon \theta' \in \mathbb{R}\Big\}.$$

An analysis similar to the one in Part 1. shows that $\Pi_m(\mathcal{G}_2) = m+1$. Therefore, by (**??**) (see Exercise B.) we have that

$$\Pi_m(\mathcal{H}_2) \le N(N-1)(m+1)\Pi_m(\mathcal{G}_1).$$

We now need to evaluate $\Pi_m(\mathcal{G}_1)$. Again, since the functions in $\mathcal{G}_1$ only depend on the first coordinate, we can assume w.l.o.g. that $N=1$. Be $x_1, \ldots, x_m \in \mathbb{R}$. We can assume again that $x_1 \le \cdots \le x_m$ (up to reordering them). Be

$$\mathcal{S} = \{(g(x_1), \ldots, g(x_n)) \colon g \in \mathcal{G}_1\}.$$

Then it holds that

$$\begin{aligned}
\mathcal{S} &\subseteq \Big\{\mathbf{v} = s\sum_{i=1}^{k}\mathbf{e}_i - s'\sum_{i=k+1}^{m}\mathbf{e}_i \colon k = 0,\ldots,m, s, s' \in \{-1,+1\}\Big\} \\
&= \Big\{\sum_{i=1}^{m}\mathbf{e}_i, -\sum_{i=1}^{m}\mathbf{e}_i\Big\} \\
&\cup \Big\{\mathbf{v} = s\sum_{i=1}^{k}\mathbf{e}_i - s\sum_{i=k+1}^{m}\mathbf{e}_i \colon k = 1,\ldots,m-1, s \in \{-1,+1\}\Big\}.
\end{aligned}$$

In particular the equality holds if $x_1 < \cdots < x_m$ and in this case $|\mathcal{S}| = 2m$. It follows that

$$\Pi_m(\mathcal{G}_1) = 2m \quad \text{which implies that} \quad \Pi_m(\mathcal{H}_2) \le 2N(N-1)m(m+1).$$

This is an explicit upper bound on $\Pi_m(\mathcal{H}_2)$. In particular $\Pi_m(\mathcal{H}_2) = O(m^2N^2)$.

B. VC-dimension

1. VC-dimension of circles in the plane.

   (a) Show that the VC-dimension of the circles in the planes is 3.

Solution: Be $\mathcal{C}$ the class of circles in the plane and $H = \{h : x \in \mathbb{R}^2 \mapsto \mathbb{1}\{x \in C\} : C \in \mathcal{C}\}$ (by $x \in C$ we mean x falls inside the circle). We want to evaluate $\mathrm{VCdim}(H)$. First we notice that $\Pi_H(3) = 8 = 2^3$. Indeed if we choose three points x_1, x_2, x_3 not aligned it holds that

$$|\{(h(x_1), h(x_2), h(x_3)) : h \in H\}| = 8,$$

since we can always find a circle containing exactly any subset of $\{x_1, x_2, x_3\}$. This shows that $\mathrm{VCdim}(H) \geq 3$. We now need to show that it can not be larger than 3. Indeed, consider the case of $m = 4$ sample points $x_1, \dots, x_4$. We can assume that they are not aligned, otherwise it's obvious that we can not realize all dichotomies. Then be D the convex hull of $\{x_1, x_2, x_3, x_4\}$. There are two casis:

1. D is a triangle: up to reordering, we can assume that x_1, x_2, x_3 are the vertices. Then the dichotomy $(1, 1, 1, 0)$ is not realizable.
2. D is a quadrilateral: Be (up to reordering) x_1, x_2 the two extrema of the longer diagonal of D (D has two diagonals). Since we chose the longer diagonal, the dichotomy $(1, 1, 0, 0)$ is not realizable.

This proves that $\Pi_4(H) < 2^4$, which implies that $\mathrm{VCdim}(H) = 3$.

   (b) Let H_1 and H_2 be two families of functions mapping from $\mathcal{X}$ to $\{0, 1\}$ and $H = \{h_1 h_2 \colon h_1 \in H_1, h_2 \in H_2\}$ their product. Show that

$$\Pi_H(m) \leq \Pi_{H_1}(m)\Pi_{H_2}(m).$$

Solution: Be $x_1, \dots, x_m \in \mathcal{X}$ and consider the set

$$\begin{aligned}
\mathcal{S} &\doteq \{(h_1(x_1)h_2(x_1), \dots, h_1(x_m)h_2(x_m)) \colon h_i \in H_i, i = 1, 2\} \\
&= \{(h_1(x_1), \dots, h_1(x_m)) \circ (h_2(x_1), \dots, h_2(x_m)) \colon h_1 \in H_1, h_2 \in H_2\} \\
&= \{(h_1(x_1), \dots, h_1(x_m)) \colon h_1 \in H_1\} \circ \{(h_2(x_1), \dots, h_2(x_m)) \colon h_2 \in H_2\},
\end{aligned}$$

where $\circ$ denotes the Hadamard (component-wise) product. It follows that

$$\begin{aligned}|\mathcal{S}| &\leq |\{(h_1(x_1),\dots,h_1(x_m))\colon h_1 \in H_1\}| \cdot |\{(h_2(x_1),\dots,h_2(x_m))\colon h_2 \in H_2\}| \\ &\leq \Pi_m(H_1)\Pi_m(H_2).\end{aligned}$$

Since this hold for all choices of $x_1,\dots,x_m$, it follows that

$$\Pi_H(m) \leq \Pi_{H_1}(m)\Pi_{H_2}(m).$$

(c) Give an upper bound on the VC-dimension of the family of intersections of k circles in the planes.

Solution: Be $\mathcal{C}_k$ the class of intersections of k circles in the plane and $H_k = \{h : x \in \mathbb{R}^2 \mapsto \mathbb{1}\{x \in C\} : C \in \mathcal{C}_k\}$. In particular $H_k = H^k$ (using the notations of Part (a)) and thus by Part (b) we know that

$$\Pi_{H_k}(m) \leq \Pi_H(m)^k.$$

Since by Part (a) we know that $\mathrm{VCdim}(H) = 3$, the corollary of Sauers Lemma implies that

$$\Pi_{H_k}(m) \leq \Pi_H(m)^k \leq \left(\frac{em}{3}\right)^{3k}.$$

It follows that

$$\begin{aligned}\mathrm{VCdim}(H_k) = \max\{m : \Pi_{H_k}(m) = 2^m\} &\leq \max\{m : \Pi_H(m)^k \geq 2^m\} \\ &\leq \max\left\{m : \left(\frac{em}{3}\right)^{3k} \geq 2^m\right\}.\end{aligned}$$

Therefore, if we find a $m > 1$ such that $(em/3)^{3k} < 2^m$, this will be an upper bound on $\mathrm{VCdim}(H_k)$. Equivalently, we are looking for $m > 1$ such that

$$m > 3k \log_2 \frac{em}{3}.$$

Since $\log_2(em/3) < \log_2 m$, we can look instead for

$$m > 3k \log_2 m. \tag{3}$$

The value $m = 6k \log_2 3k$ satisfies (3). Indeed $3k > 2\log_2 3k$ (since $k > 1$) and thus

$$\begin{aligned}m = 6k\log_2 3k &= 3k\log_2 3k + 3k\log_2 3k \\ &> 3k\log_2 3k + 3k\log_2 2\log_2 3k = 3k\log_2 m.\end{aligned}$$

Therefore we proved that

$$\mathrm{VCdim}(H_k) \leq 6k \log_2 3k.$$

2. VC-dimension of Decision trees. A full binary tree is a tree in which each node is either a leaf or it is an internal node and admits exactly two child nodes.

   (a) Show that that a binary tree with n internal nodes has exactly $n + 1$ leaves (*Hint*: you can proceed by induction).

Solution: We prove this proposition by induction. If $n = 1$, then it follows by the definition of full binary tree that there are $n + 1 = 2$ leaves. Suppose now that the result is true for full binary trees with $n - 1$ internal nodes. Consider a full binary tree T with n internal nodes. There exists an internal node v such that both its child nodes are leaves. If we consider the tree T' obtained by eliminating the child nodes of v from T, then T' it is a full binary tree with $n-1$ internal nodes and therefore n leaves (by the induction assumption). Going from T' to T we loose the leaf v, since it becomes an internal node, but we get two new leaves (the child nodes of v). So the number of leaves of T is $n - 1 + 2 = n + 1$. This concludes the proof.

   (b) A binary decision tree is a full binary tree with each leaf labeled with $+1$ or -1 and each internal node labeled with a question.

   A binary decision tree classifies a point as follows: starting with the root of the tree, if the internal node question applied to the point admits a positive answer, then the current node becomes the right child, otherwise it becomes the left child. This is repeated until a leaf node is reached. The label assigned to the point is then the sign of that leaf node.

   Suppose the node questions are of the form $x_i > 0$, $i \in [1, N]$. Show that the VC-dimension of the set of binary decision trees with n nodes in dimension N is at most $(2n+1)\log_2(N+2)$ (*Hint*: bound the cardinality of the set). Use that to derive an upper bound on the Rademacher complexity of that set.

Solution: Be H the class of binary decision trees with n internal nodes of the type described above. A trivial bound on the growth function of H is

$$\Pi_m(H) \leq |H|.$$

We then want to give a bound on $|H|$. Each node v of a tree in H can either be labeled with a question (out of N possible question and this would implicitly characterize the node as an internal node) or with ± 1 (and this would implicitly characterize the node as a leaf node). Therefore each node can be labelled in $N + 2$ possible different ways. It follows that the cardinality of H can be bound as

$$\Pi_m(H) \leq |H| \leq (N + 2)^{(2n+1)}.$$

Note that this bound is always strict (cause the number of internal nodes must always be n). In particular this implies that

$$\begin{aligned}
\text{VCdim}(H) &= \sup\{m \colon \Pi_m(H) = 2^m\} \\
&\leq \sup\{m \colon (N + 2)^{(2n+1)} \geq 2^m\} \\
&\leq \sup\{m \colon m \leq (2n + 1)\log_2(N + 2)\} \\
&\leq (2n + 1)\log_2(N + 2).
\end{aligned}$$

Using the result from the slides, the Rademacher complexity can be bounded as

$$\mathfrak{R}_m(H) \leq \sqrt{\frac{2\log \Pi_m(H)}{m}} \leq \sqrt{\frac{2(2n + 1)\log(N + 2)}{m}}.$$

C. Support-Vector Machines

1. Download and install the `libsvm` software library from:

   `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

2. Consider the `spambase` data set

   `http://archive.ics.uci.edu/ml/datasets/Spambase`.

   Download a shuffled version of that dataset from

   `http://www.cs.nyu.edu/~mohri/ml17/spambase.data.shuffled`

   Use the `libsvm` scaling tool to scale the features of all the data. Use the first 3000 examples for training, the last 1601 for testing. The scaling parameters should be computed only on the training data and then applied to the test data.

   Solution: I used the `sklearn` implementation of the `libsvm` library for all the below reported results, apart from the computation of the number of support vectors in Part 4. (I used `libsvm` for this). The data was re-scaled between -1 and 1 for Part 3. and 4. Nevertheless, the `sklearn` implementation was overly slow for Part 5. (customized kernel). Thus, to decrease the computational load, I re-scaled the data betweem -0.1 and 0.1 in Part 5. Clearly, this affected the performances of SVM. In particular, it implied the use of higher values of C to get results similar to Part 3. in accuracy.

3. Consider the binary classification that consists of predicting if the e-mail message if a spam using the 57 features. Use SVMs combined with polynomial kernels to tackle this binary classification problem.

   To do that, randomly split the training data into ten equal-sized disjoint sets. For each value of the polynomial degree, $d = 1, 2, 3, 4$, plot the average cross-validation error plus or minus one standard deviation as a function of C (let other parameters of polynomial kernels in `libsvm` be equal to their default values), varying C in powers of 2, starting from a small value $C = 2^{-k}$ to $C = 2^k$, for some value of k. k should be chosen so that you see a significant variation in training error, starting from a very high training error to a low training error. Expect longer training times with `libsvm` as the value of C increases.

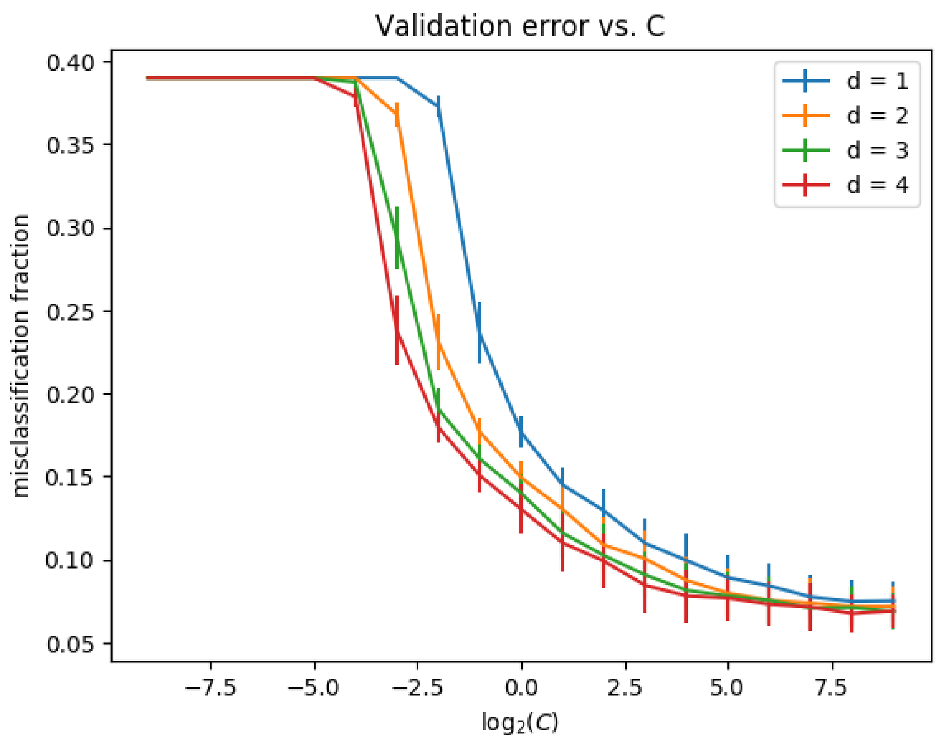

Figure 1: Average error according to 10-fold cross-validation, with error-bars indicating one standard deviation.

Solution: Figure 1 shows the average cross-validation performance as a function of the regularization parameter $C = 2^k$. As we can see, we get good results for $k \geq 1$. Polynomial kernels of higher degree seem to perform better. The optimal parameters found are $C^* = 2^8$ and $d^* = 4$.

4. Let (C^*, d^*) be the best pair found previously. Fix C to be C^*. Plot the ten-fold cross-validation error and the test errors for the hypotheses obtained as a function of d. Plot the average number of support vectors obtained as a function of d. How many of the support vectors lie on the margin hyperplanes?

Solution: The first plot in Figure 2 shows that the test error decreases (slightly) with an increase in degree. Also the test error on the held-out dataset is (slightly) optimistic when compared to the cross-validation error.

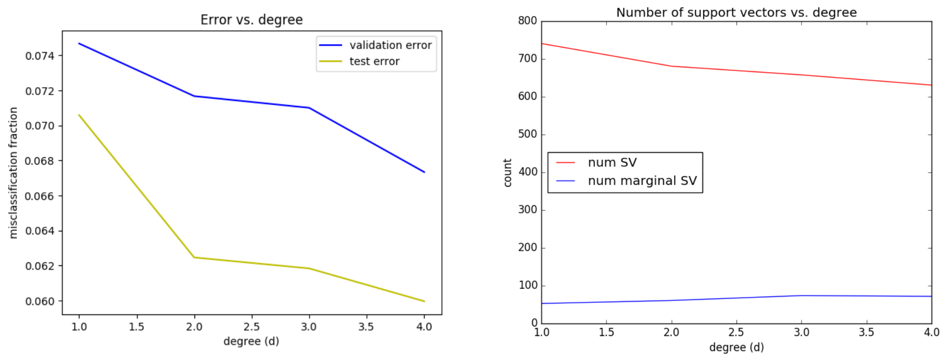

Figure 2: The test and validation error as a function of degree (left panel) as well as the number of total and marginal support vectors (right panel).

The second plot shows that the total number of marginal support vectors increases with d (as does the dimension of the feature space) while the total number of overall support vectors decreases.

5. For any d, let K_d denote the polynomial kernel of degree d. Show that for any fixed integer $u > 0$, $G_u = \frac{2}{u(u+1)} \sum_{i \leq j \leq u} K_i K_j$ is a PDS kernel. Use SVMs combined with the polynomial kernel G_4 to tackle the same binary classification problem as in the previous questions: as in the previous questions, use ten-fold cross-validation to determine the best value of C; report the ten-fold cross-validation error and the test error of the hypothesis obtained when training with G_u.

Solution: For any fixed integer $u > 0$, $G_u = \frac{2}{u(u+1)} \sum_{i \leq j \leq u} K_i K_j$ is a PDS kernel since it is a (normalized) sum of products of PDS kernels and PDS kernels are closed under sum and product.
Using SVMs combined with the polynomial kernel G_4, the optimal value of C found is $C^* = 2^{13}$. For this value of C, the cross-validation error is 0.067, while the test error on the held-out dataset is 0.058.

D. Rademacher complexity

Let $p > 2$ and let q be its conjugate: $1/p + 1/q = 1$. Let $\mathcal{H}_p$ be the family of linear functions defined over $\{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|_p \leq r_p\}$ by

$$\mathcal{H} = \Big\{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \colon \|\mathbf{w}\|_q \leq \Lambda_q\Big\},$$

for some $r_p > 0$ and $\Lambda_q > 0$. Give an upper bound on $\mathfrak{R}_m(\mathcal{H})$ in terms of Λ_q and r_p, assuming that for $p > 2$ the following inequality holds for all $z_1, \dots, z_m \in \mathbb{R}$: $\mathbb{E}_{\boldsymbol{\sigma}}\left[|\sum_{i=1}^m \sigma_i z_i|^p\right] \leq [\frac{p}{2}\sum_{i=1}^m z_i^2]^{\frac{p}{2}}$.

Solution: By definition

$$\mathfrak{R}_m(\mathcal{H}) = \frac{1}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\mathop{\mathbb{E}}_{\sigma}\Bigg[\sup_{f\in\mathcal{H}}\sum_{i=1}^m \sigma_i f(x_i)\Bigg]\Bigg] = \frac{1}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\mathop{\mathbb{E}}_{\sigma}\Bigg[\sup_{\|w\|_q\leq\Lambda_q}\sum_{i=1}^m \sigma_i\langle w, x_i\rangle\Bigg]\Bigg]$$
$$= \frac{1}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\mathop{\mathbb{E}}_{\sigma}\Bigg[\sup_{\|w\|_q\leq\Lambda_q}\langle w, \sum_{i=1}^m \sigma_i x_i\rangle\Bigg]\Bigg].$$

Then, by CauchySchwarz inequality,

$$\mathfrak{R}_m(\mathcal{H}) \leq \frac{1}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\mathop{\mathbb{E}}_{\sigma}\Bigg[\sup_{\|w\|_q\leq\Lambda_q}\|w\|_q\Big\|\sum_{i=1}^m \sigma_i x_i\Big\|_p\Bigg]\Bigg] \leq \frac{\Lambda_q}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\mathop{\mathbb{E}}_{\sigma}\Big\|\sum_{i=1}^m \sigma_i x_i\Big\|_p\Bigg].$$

By concavity of the function $y \mapsto y^{1/p}$ and the inequality stated given in the exercise, we get

$$\mathfrak{R}_m(\mathcal{H}) \leq \frac{\Lambda_q}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\mathop{\mathbb{E}}_{\sigma}\Big[\Big(\sum_{j=1}^N\Big|\sum_{i=1}^m \sigma_i x_{i,j}\Big|^p\Big)^{1/p}\Big]\Bigg] \leq \frac{\Lambda_q}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\mathop{\mathbb{E}}_{\sigma}\Big[\sum_{j=1}^N\Big|\sum_{i=1}^m \sigma_i x_{i,j}\Big|^p\Big]^{1/p}\Bigg]$$
$$= \frac{\Lambda_q}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\Big(\sum_{j=1}^N\mathop{\mathbb{E}}_{\sigma}\Big|\sum_{i=1}^m \sigma_i x_{i,j}\Big|^p\Big)^{1/p}\Bigg] \leq \frac{\Lambda_q}{m}\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\Big(\sum_{j=1}^N\big[(p/2)^{p/2}\sum_{i=1}^m x_{i,j}^2\big]^{p/2}\Big)^{1/p}\Bigg]$$
$$= \frac{\Lambda_q}{m}\sqrt{\frac{p}{2}}\cdot\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\Big(\sum_{j=1}^N\big[\sum_{i=1}^m x_{i,j}^2\big]^{p/2}\Big)^{1/p}\Bigg].$$

Since $(\sum_{i=1}^m x_{i,j}^2)^{p/2} \leq \sum_{i=1}^m |x_{i,j}|^p$, we get that

$$\mathfrak{R}_m(\mathcal{H}) \leq \frac{\Lambda_q}{m}\sqrt{\frac{p}{2}}\cdot\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\Big(\sum_{j=1}^N\sum_{i=1}^m|x_{i,j}|^p\Big)^{1/p}\Bigg] = \frac{\Lambda_q}{m}\sqrt{\frac{p}{2}}\cdot\mathop{\mathbb{E}}_{x\sim D^m}\Bigg[\Big(\sum_{i=1}^m\|x_i\|_p^p\Big)^{1/p}\Bigg]$$
$$\leq \frac{\Lambda_q}{m}\sqrt{\frac{p}{2}}\cdot\Big(\sum_{i=1}^m r_p^p\Big)^{1/p} = \frac{r_p\Lambda_q}{m}\sqrt{\frac{p}{2}}\cdot m^{1/p}$$

Thus,

$$\mathfrak{R}_m(\mathcal{H}) \leq \frac{r_p \Lambda_q}{m^{1-\frac{1}{p}}} \sqrt{\frac{p}{2}}.$$