EE3-23: Assignment 2

Spring, 2017

Instructions:

Solve all problems. The second part of Problem 5 (a) is optional, you can earn bonus marks for it.

You need to submit a single zip file on Blackboard, named `assignment2_<name>.zip` where `<name>` is your name. The file should contain a report in PDF (write your name on your solution), named `assignment2_<username>_<name>.pdf` where `<username>` is your blackboard user name, and your code and other relevant files. Your code must include a main file named `assignment2_<username>` with a possible extension (such as `.m` or `.py`) or a Makefile which produces the main file on any standard linux system: the main file should be either an executable or a script file, which generates all the data and plots required in your solution.

Justify your answers. You are allowed to consult with others (mention their names at the beginning of your solution), but you need to work out and write up your solution alone. The use of any written sources (e.g., books or the web) is permitted.

Include the following pledge into your answer:

I, <YOUR NAME>, 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.

If you don't include your pledge, you will get 0%.

Deadline: February 20, 20:00. Late submissions are not accepted.

Problem 1. Consider a binary classification problem (with binary error) over a set $\mathcal{X}$, and let $\mathcal{H}$ denote a finite hypothesis class over $\mathcal{X}$ such that there exists a perfect hypothesis h^* in $\mathcal{H}$, that is, for any $x \in \mathcal{X}$, its label is $h^*(x)$. Show that for any $\delta \in (0,1)$ and any ERM hypothesis g selected on n i.i.d. data points (i.e., $g \in \operatorname{argmin} \widehat{R}_n$), the test error of g satisfies, with probability at least $1-\delta$, $R(g) \le \frac{\log(|\mathcal{H}|/\delta)}{n}$. **[6 points]**

Hint: Show that $\widehat{R}_n(g) = 0$. For any $\varepsilon > 0$, show that for any hypothesis h with $R(h) > \varepsilon$, its probability of being selected as an empirical risk minimizer is at most $(1-\varepsilon)^n$. Use the inequality $(1-\varepsilon)^n \le e^{-\varepsilon n}$ to finish the proof. Note that the convergence–due to the perfect hypothesis assumption, also known as the *realizability assumption*–is much faster than what we can obtain from Hoeffding's inequality ($O(1/n)$ vs. $O(1/\sqrt{n})$).

Problem 2 (VC dimension of axis-aligned rectangles)**.** For any real numbers $a_1 \le a_2$ and $b_1 \le b_2$, let $R(a_1, a_2, b_1, b_2) = \{(x_1, x_2) \in \mathbb{R}^2 : a_1 \le x_1 \le a_2 \text{ and } b_1 \le x_2 \le b_2\}$ denote an axis-aligned rectangle, and consider the class of indicator functions of such rectangles:

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

Show that the VC dimension of $\mathcal{H}$ is 4. **[5 points]**

Hint: Consider the case when, out of 5 points, the leftmost, rightmost, highest and lowest points have the same label.

Problem 3. Show that for any n and k and data set $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ of size n (for some base set $\mathcal{X}$) with $x_i \neq x_j$, there exists a hypothesis class $\mathcal{H} \subset \{h : \mathcal{X} \to \{-1, +1\}\ \}$ that cannot shatter any k points from X (that is, $m_{\mathcal{H}}(k) < 2^k$) and $|\mathcal{H}(x_1, \ldots, x_n)| = \sum_{i=0}^{k-1} \binom{n}{i}$. **[3 points]**

Problem 4. Show that if $d_{VC} < \infty$ for a hypothesis class $\mathcal{H}$, then for all n

(a) $m_{\mathcal{H}}(n) \leq n^{d_{VC}} + 1 \leq (n+1)^{d_{VC}}$, **[4 points]**

(b) for all $n \geq d$, $m_{\mathcal{H}}(n) \leq \left(\frac{ne}{d_{VC}}\right)^{d_{VC}}$. **[2 points]**

Compare the above bounds. **[1 points]**

Hint: Use Sauer's lemma. For the first inequality in (a), use induction. For (b), show that $\sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i} \leq \left(\frac{n}{d}\right)^{d} \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^{i}$, and use the binomial theorem and $(1+1/x)^x \leq e$ to show that the latter sum is bounded by e^d.

Problem 5 (Structural risk minimization)**.** Let $\mathcal{X} = [0, 2.5] \times [-1, 2]$ and $f(x) = x(x-1)(x-2)$. Consider a noisy classification problem where the features (x_1, x_2) are uniformly distributed on $\mathcal{X}$ and the label of (x_1, x_2) has the following distribution: $\mathbb{P}\left(y = 1 | x_2 \geq f(x_1)\right) = \mathbb{P}\left(y = -1 | x_2 < f(x_1)\right) = 0.9$ and $\mathbb{P}\left(y = -1 | x_2 \geq f(x_1)\right) = \mathbb{P}\left(y = 1 | x_2 < f(x_1)\right) = 0.1$.

(a) For any function $g : \mathbb{R} \to \mathbb{R}$, define a classifier $h_g : \mathcal{X} \to \{+1, -1\}$ such that $h_g((x_1, x_2)) = 1$ if $x_2 \geq g(x_1)$ and -1 otherwise. That is, the true labeling function is h_f. Now consider the set of classifiers $\mathcal{H}_q = \{h_p | p : \mathbb{R} \to \mathbb{R}\ \textit{is a polynomial of degree at most}\ q\}$. Find a non-linear transformation of $\mathcal{X}$ to a new set $\mathcal{Z}$ such that any $h \in \mathcal{H}_q$ can be expressed as a linear classifier in $\mathcal{Z}$. Using this, show that the VC dimension of $\mathcal{H}_q$ is at most $q + 1$. **[4 points]**

Hint: Consider powers of x_1 as new features.

Bonus question: Show that the VC dimension of $\mathcal{H}_q$ is $q + 1$. **[3 points]**

(b) Generate 10, 100, 10000 points uniformly random from $\mathcal{X}$, as three different training sets. For each of them, use structural risk minimization (SRM) to find the best classifier from $\mathcal{H} = \cup_{q=0}^{4} \mathcal{H}_q$, where the weights w_i of the hypotheses classes $\mathcal{H}_q$ are equal. Provide the resulting training and test errors for the SRM solution, as well as for the ERM solutions for each $\mathcal{H}_q$ separately. Comment on the results. **[8 points]**

Hints: To estimate the test error, generate a large enough test set and use the empirical training error as an estimate. How many points did you chose and why? Also, do not forget that the experiments are random, so you need to repeat them a few times to analyze the behavior of the algorithms. To compute the complexity term, you may want to use the bound of Problem 4 (b). The running time for the largest dataset may be quite large; make sure you test your code on smaller data.

(c) If you find that the SRM solution does not work as well as you expected (why?), try multiplying the complexity term by 0.01 or 0.1. Can you see an improvement after this heuristic modification? Explain your observations. **[4 points]**

Hint: The VC bound can be quite loose in practice.