Homework #2

RELEASE DATE: 10/15/2015

DUE DATE: 11/02/2015 (**MONDAY**), BEFORE NOON

QUESTIONS ABOUT HOMEWORK MATERIALS ARE WELCOMED ON THE COURSERA FORUM.

Unless granted by the instructor in advance, you must turn in a printed/written copy of your solutions (without the source code) for all problems.

For problems marked with (*), please follow the guidelines on the course website and upload your source code to designated places. You are encouraged to (but not required to) include a README to help the TAs check your source code. Any programming language/platform is allowed.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

Discussions on course materials and homework solutions are encouraged. But you should write the final solutions alone and understand them fully. Books, notes, and Internet resources can be consulted, but not copied from.

Since everyone needs to write the final solutions alone, there is absolutely no need to lend your homework solutions and/or source codes to your classmates at any time. In order to maximize the level of fairness in this class, lending and borrowing homework solutions are both regarded as dishonest behaviors and will be punished according to the honesty policy.

You should write your solutions in English or Chinese with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

This homework set comes with 200 points and 20 bonus points. In general, every homework set would come with a full credit of 200 points, with some possible bonus points.

Questions 1-2 are about *noisy targets*

1. Consider the bin model for a hypothesis h that makes an error with probability μ in approximating a deterministic target function f (both h and f outputs $\{-1, +1\}$). If we use the same h to approximate a noisy version of f given by

$$P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$$

$$P(y|\mathbf{x}) = \begin{cases} \lambda & y = f(\mathbf{x}) \\ 1 - \lambda & \text{otherwise} \end{cases}$$

What is the probability of error that h makes in approximating the noisy target y? Please provide explanation of your answer.

2. Following Question 1, with what value of λ will the performance of h be independent of μ? Please provide explanation of your answer.

Questions 3-5 are about *generalization error*, and getting the feel of the bounds numerically. Please use the simple upper bound $N^{d_{\mathbf{vc}}}$ on the growth function $m_{\mathcal{H}}(N)$, assuming that $N \geq 2$ and $d_{\mathbf{vc}} \geq 2$.

3. For an $\mathcal{H}$ with $d_{\text{VC}} = 10$, if you want 95% confidence that your generalization error is at most 0.05, what is the the sample size that the VC generalization bound predicts? Please provide calculating steps of your answer, and round your answer to the closest thousand (that is, your answer should be something like 845000).

4. There are a number of bounds on the generalization error ϵ, all holding with probability at least $1 - \delta$. Fix $d_{\text{VC}} = 50$ and $\delta = 0.05$ and plot these bounds as a function of N. Use any numerical

method to calculate the (most) generalization error by the bound. Which bound is the tightest (smallest) for very large N, say $N = 10,000$? What is the generalization error calculated by the tightest bound? Note that Devroye and Parrondo & Van den Broek are implicit bounds in ϵ.

[a] Original VC bound: $\epsilon \leq \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$

[b] Variant VC bound: $\epsilon \leq \sqrt{\frac{16}{N} \ln \frac{2m_{\mathcal{H}}(N)}{\sqrt{\delta}}}$

[c] Rademacher Penalty Bound: $\epsilon \leq \sqrt{\frac{2\ln(2Nm_{\mathcal{H}}(N))}{N}} + \sqrt{\frac{2}{N} \ln \frac{1}{\delta}} + \frac{1}{N}$

[d] Parrondo and Van den Broek: $\epsilon \leq \sqrt{\frac{1}{N}(2\epsilon + \ln \frac{6m_{\mathcal{H}}(2N)}{\delta})}$

[e] Devroye: $\epsilon \leq \sqrt{\frac{1}{2N}(4\epsilon(1+\epsilon) + \ln \frac{4m_{\mathcal{H}}(N^2)}{\delta})}$

5. Continuing from Question 4, for small N, say $N = 5$, which bound is the tightest (smallest)? What is the generalization error calculated by the tightest bound?

In Questions 6-11, you are asked to play with the *growth function* or *VC-dimension* of some hypothesis sets. You should make sure your proof is rigorous and complete, as they will be carefully checked.

6. What is the growth function $m_{\mathcal{H}}(N)$ of "positive-and-negative intervals on $\mathbb{R}$"? The hypothesis set $\mathcal{H}$ of "positive-and-negative intervals" contains the functions which are $+1$ within one interval $[\ell, r]$ and -1 elsewhere, as well as the functions which are -1 within one interval $[\ell, r]$ and $+1$ elsewhere. For instance, the hypothesis $h_1(x) = \text{sign}(x(x-4))$ is a negative interval with -1 within $[0, 4]$ and $+1$ elsewhere, and hence belongs to $\mathcal{H}$. The hypothesis $h_2(x) = \text{sign}((x+1)(x)(x-1))$ contains two positive intervals in $[-1, 0]$ and $[1, \infty)$ and hence does not belong to $\mathcal{H}$. Please provide proof of your answer.

7. Continuing from the previous problem, what is the VC-dimension of the "positive-and-negative intervals on $\mathbb{R}$"? Please provide proof of your answer.

8. What is the growth function $m_{\mathcal{H}}(N)$ of "positive donuts in $\mathbb{R}^2$"? The hypothesis set $\mathcal{H}$ of "positive donuts" contains hypotheses formed by two concentric circles centered at the origin. In particular, each hypothesis is $+1$ within a "donut" region of $a^2 \leq x_1^2 + x_2^2 \leq b^2$ and -1 elsewhere. Without loss of generality, we assume $0 < a < b < \infty$. Please provide proof of your answer.

9. Consider the "polynomial discriminant" hypothesis set of degree D on $\mathbb{R}$, which is given by

$$\mathcal{H} = \left\{ h_{\mathbf{c}} \;\middle|\; h_{\mathbf{c}}(x) = \text{sign}\left(\sum_{i=0}^{D} c_i x^i\right) \right\}$$

What is the VC-Dimension of such an $\mathcal{H}$? Please provide proof of your answer.

10. Consider the "simplified decision trees" hypothesis set on $\mathbb{R}^d$, which is given by

$$\mathcal{H} = \{h_{\mathbf{t},\mathbf{S}} \mid \quad h_{\mathbf{t},\mathbf{S}}(\mathbf{x}) = 2\,[\![\mathbf{v} \in \mathbf{S}]\!] - 1, \text{where } v_i = [\![x_i > t_i]\!]\,, \\ \mathbf{S} \text{ a collection of vectors in } \{0,1\}^d, \mathbf{t} \in \mathbb{R}^d \quad \}$$

That is, each hypothesis makes a prediction by first using the d thresholds t_i to locate $\mathbf{x}$ to be within one of the 2^d hyper-rectangular regions, and looking up $\mathbf{S}$ to decide whether the region should be $+1$ or -1. What is the VC-dimension of the "simplified decision trees" hypothesis set? Please provide proof of your answer.

11. Consider the "triangle waves" hypothesis set on $\mathbb{R}$, , which is given by

$$\mathcal{H} = \{h_\alpha \mid \quad h_\alpha(x) = \text{sign}(|(\alpha x) \bmod 4 - 2| - 1), \alpha \in \mathbb{R}\}$$

Here $(z \bmod 4)$ is a number $z - 4k$ for some integer k such that $z - 4k \in [0, 4)$. For instance, $(11.26 \bmod 4)$ is 3.26, and $(-11.26 \bmod 4)$ is 0.74. What is the VC-Dimension of such an $\mathcal{H}$? Please provide proof of your answer.

In Questions 12-15, you are asked to verify some properties or bounds on the growth function and VC-dimension.

12. Is $\min_{1 \le i \le N-1} 2^i m_{\mathcal{H}}(N-i)$ an upper bound of the growth function $m_{\mathcal{H}}(N)$ for $N \ge d_{\text{VC}} \ge 2$? Please provide proof of your answer.

13. Is $2^{\lfloor \sqrt{N} \rfloor}$ a possible growth function $m_{\mathcal{H}}(N)$ for some hypothesis set? Please provide proof of your answer.

14. For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_K$ with finite, positive VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, consider the VC dimension of the **intersection** of the sets. Prove or disprove that

$$0 \ \le \ d_{\text{VC}}\left(\bigcap_{k=1}^{K} \mathcal{H}_k\right) \ \le \ \min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K}.$$

(The VC dimension of an empty set or a singleton set is taken as zero)

15. For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_K$ with finite, positive VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, consider the VC dimension of the **union** of the sets. Prove or disprove that

$$\max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K} \ \le \ d_{\text{VC}}\left(\bigcup_{k=1}^{K} \mathcal{H}_k\right) \ \le \ K - 1 + \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k).$$

For Questions 16-20, you will play with the decision stump algorithm.

In class, we taught about the learning model of "positive and negative rays" (which is simply one-dimensional perceptron) for one-dimensional data. The model contains hypotheses of the form:

$$h_{s,\theta}(x) = s \cdot \text{sign}(x - \theta).$$

The model is frequently named the "decision stump" model and is one of the simplest learning models. As shown in class, for one-dimensional data, the VC dimension of the decision stump model is 2.

In fact, the decision stump model is one of the few models that we could easily minimize E_{in} for binary classification efficiently by enumerating all possible thresholds. In particular, for N examples, there are at most $2N$ dichotomies (see page 22 of class05 slides), and thus at most $2N$ different E_{in} values. We can then easily choose the dichotomy that leads to the lowest E_{in}, where ties can be broken by randomly choosing among the lowest-E_{in} ones. The chosen dichotomy stands for a combination of some 'spot' (range of θ) and s, and commonly the median of the range is chosen as the θ that realizes the dichotomy.

In this problem, you are asked to implement such and algorithm and run your program on an artificial data set. First of all, start by generating a one-dimensional data by the procedure below:

a) Generate x by a uniform distribution in $[-1, 1]$.

b) Generate y by $\tilde{s}(x)$ + noise where $\tilde{s}(x) = \text{sign}(x)$ and the noise flips the result with 20% probability.

16. For any decision stump $h_{s,\theta}$ with $\theta \in [-1, 1]$, express $E_{\text{out}}(h_{s,\theta})$ as a function of θ and s. Please provide your derivation steps.

17. (*) Generate a data set of size 20 by the procedure above and run the one-dimensional decision stump algorithm on the data set. Record E_{in} and compute E_{out} with the formula above. Repeat the experiment (including data generation, running the decision stump algorithm, and computing E_{in} and E_{out}) 5,000 times. What is the average E_{in}? Plot a histogram for your E_{in} distribution.

18. (*) Continuing from the previous question, what is the average E_{out}? Plot a histogram for your E_{out} distribution.

Decision stumps can also work for multi-dimensional data. In particular, each decision stump now deals with a specific dimension i, as shown below.

$$h_{s,i,\theta}(\mathbf{x}) = s \cdot \text{sign}(x_i - \theta).$$

Implement the following decision stump algorithm for multi-dimensional data:

a) for each dimension $i = 1, 2, \cdots, d$, find the best decision stump $h_{s,i,\theta}$ using the one-dimensional decision stump algorithm that you have just implemented.

b) return the "best of best" decision stump in terms of E_{in}. If there is a tie, please randomly choose among the lowest-E_{in} ones.

The training data $\mathcal{D}_{train}$ is available at:

`http://www.csie.ntu.edu.tw/~htlin/course/ml14fall/hw2/hw2_train.dat`

The testing data $\mathcal{D}_{test}$ is available at:

`http://www.csie.ntu.edu.tw/~htlin/course/ml14fall/hw2/hw2_test.dat`

19. (*) Run the algorithm on the $\mathcal{D}_{train}$. What is the optimal decision stump returned by your program? What is the E_{in} of the optimal decision stump?

20. (*) Use the returned decision stump to predict the label of each example within the $\mathcal{D}_{test}$. Report an estimate of E_{out} by E_{test}.

Bonus: More on Growth Function

21. In class, we have shown that

$$B(N, k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Show that in fact the equality holds. *(Hint: there is a intuitive construction of a specific set of $\sum_{i=0}^{k-1} \binom{N}{i}$ dichotomies, where no subset of k variables can be shattered.)*