Homework #2

RELEASE DATE: 10/15/2014

DUE DATE: **10/29/2014**, 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.

There are two kinds of regular problems.

- *multiple-choice question (MCQ): There are several choices and* ***only one of them is correct****. You should choose one and only one.*
- *multiple-response question (MRQ): There are several choices and* ***none, some, or all of them are correct****. You should write down every choice that you think to be correct.*

Some problems also come with ***(+ ...)*** *that contains additional todo items.*
If there are big bonus questions (BBQ), please simply follow the problem guideline to write down your solution, if you choose to tackle them.

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. (MCQ) 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?

[a] $1 - \lambda$

[b] μ

[c] $\lambda(1 - \mu) + (1 - \lambda)\mu$

[d] $\lambda\mu + (1 - \lambda)(1 - \mu)$

[e] none of the other choices

(+ explanation of your choice)

2. (MCQ) Following Question 1, with what value of λ will the performance of h be independent of μ?

[a] 0

[b] 1

[c] 0 or 1

[d] 0.5

[e] none of the other choices

(+ explanation of your choice)

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. (MCQ) 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 closest numerical approximation of the sample size that the VC generalization bound predicts?

[a] 420,000

[b] 440,000

[c] 460,000

[d] 480,000

[e] 500,000

(+ explanation of your choice)

4. (MCQ) 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. Which bound is the tightest (smallest) for very large N, say $N = 10,000$? 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: $\epsilon \leq \sqrt{\frac{2 \ln(2N m_{\mathcal{H}}(N))}{N}} + \sqrt{\frac{2}{N} \ln \frac{1}{\delta}} + \frac{1}{N}$

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

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

(+ explanation of your choice)

5. (MCQ) Continuing from Question 4, for small N, say $N = 5$, which bound is the tightest (smallest)?

[a] Original VC bound

[b] Variant VC bound

[c] Rademacher Penalty Bound

[d] Parrondo and Van den Broek

[e] Devroye

(+ explanation of your choice)

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. (MCQ) 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 $\mathcal{H}$.

[a] $N^2 - N + 2$

[b] N^2

[c] $N^2 + 1$

[d] $N^2 + N + 2$

[e] none of the other choices

(+ proof of your choice)

7. (MCQ) Continuing from the previous problem, what is the VC-dimension of the "positive-and-negative intervals on $\mathbb{R}$"?

[a] 2

[b] 3

[c] 4

[d] 5

[e] ∞

(+ proof of your choice)

8. (MCQ) 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 $0 < a < b < \infty$.

[a] $N + 1$

[b] $\binom{N}{2} + 1$

[c] $\binom{N+1}{2} + 1$

[d] $\binom{N+1}{3} + 1$

[e] none of the other choices

(+ proof of your choice)

9. (MCQ) 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 $\mathcal{H}$?

[a] D

[b] $D + 1$

[c] $D + 2$

[d] ∞

[e] none of the other choices

(+ proof of your choice)

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

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

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?

[a] 2^d

[b] 2^{d+1}

[c] $2^{d+1} - 3$

[d] ∞

[e] none of the other choices

(+ proof of your choice)

11. (MCQ) 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 $\mathcal{H}$?

[a] 1

[b] 2

[c] 3

[d] ∞

[e] none of the other choices

(+ proof of your choice)

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

12. (MRQ) Which of the following are upper bounds of the growth function $m_\mathcal{H}(N)$ for $N \geq d_{\text{VC}} \geq 2$?

[a] $m_\mathcal{H}\left(\lfloor \frac{N}{2} \rfloor\right) \cdot m_\mathcal{H}\left(\lceil \frac{N}{2} \rceil\right)$

[b] $2^{d_{\text{VC}}}$

[c] $\min\limits_{1 \leq i \leq N-1} 2^i m_\mathcal{H}(N-i)$

[d] $N^{d_{\text{VC}}} + 1$

[e] $m_\mathcal{H}(N-1) + N \cdot d_{\text{VC}}$

(+ explanation of your choice)

13. (MRQ) Which of the following are possible growth functions $m_\mathcal{H}(N)$ for some hypothesis set?

[a] 2^N

[b] $2^{\lfloor \sqrt{N} \rfloor}$

[c] $2^{\lfloor N/2 \rfloor}$

[d] $2^{\lceil N/2 \rceil}$

[e] $1 + N + \frac{N(N-1)(N-2)}{6}$

(+ explanation of your choice)

14. (MCQ) For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_K$ with finite, positive VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, some of the following bounds are correct and some are not. Which among the correct ones is the tightest bound on the VC dimension of the **intersection** of the sets: $d_{\text{VC}}(\bigcap_{k=1}^{K} \mathcal{H}_k)$? (The VC dimension of an empty set or a singleton set is taken as zero)

[a] $0 \leq d_{\text{VC}}(\bigcap_{k=1}^{K} \mathcal{H}_k) \leq \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

[b] $0 \leq d_{\text{VC}}(\bigcap_{k=1}^{K} \mathcal{H}_k) \leq \min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K}$

[c] $0 \leq d_{\text{VC}}(\bigcap_{k=1}^{K} \mathcal{H}_k) \leq \max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K}$

[d] $\min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K} \leq d_{\text{VC}}(\bigcap_{k=1}^{K} \mathcal{H}_k) \leq \max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K}$

[e] $\min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K} \leq d_{\text{VC}}(\bigcap_{k=1}^{K} \mathcal{H}_k) \leq \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

(+ explanation of your choice)

15. (MCQ) For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_K$ with finite, positive VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, some of the following bounds are correct and some are not. Which among the correct ones is the tightest bound on the VC dimension of the **union** of the sets: $d_{\text{VC}}(\bigcup_{k=1}^{K} \mathcal{H}_k)$?

[a] $0 \leq d_{\text{VC}}(\bigcup_{k=1}^{K} \mathcal{H}_k) \leq \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

[b] $0 \leq d_{\text{VC}}(\bigcup_{k=1}^{K} \mathcal{H}_k) \leq K - 1 + \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

[c] $\min\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K} \leq d_{\text{VC}}(\bigcup_{k=1}^{K} \mathcal{H}_k) \leq \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

[d] $\max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K} \leq d_{\text{VC}}(\bigcup_{k=1}^{K} \mathcal{H}_k) \leq \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

[e] $\max\{d_{\text{VC}}(\mathcal{H}_k)\}_{k=1}^{K} \leq d_{\text{VC}}(\bigcup_{k=1}^{K} \mathcal{H}_k) \leq K - 1 + \sum_{k=1}^{K} d_{\text{VC}}(\mathcal{H}_k)$

(+ explanation of your choice)

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

16. (MCQ) 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.

[a] $0.3 + 0.5s(|\theta| - 1)$

[b] $0.3 + 0.5s(1 - |\theta|)$

[c] $0.5 + 0.3s(|\theta| - 1)$

[d] $0.5 + 0.3s(1 - |\theta|)$

[e] none of the other choices

(+ your derivation steps.)

17. (MCQ, *) 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}? Choose the closest option.

[a] 0.05

[b] 0.15

[c] 0.25

[d] 0.35

[e] 0.45

18. (MCQ, *) Continuing from the previous question, what is the average E_{out}? Choose the closest option.

[a] 0.05

[b] 0.15

[c] 0.25

[d] 0.35

[e] 0.45

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. (MCQ, *) Run the algorithm on the $\mathcal{D}_{train}$. Report the E_{in} of the optimal decision stump returned by your program. Choose the closest option.

[a] 0.05

[b] 0.15

[c] 0.25

[d] 0.35

[e] 0.45

20. (MCQ, *) 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}. Choose the closest option.

[a] 0.05

[b] 0.15

[c] 0.25

[d] 0.35

[e] 0.45

Bonus: An 'Easy' One, and More on Growth Function

21. (BBQ, 10 points) Kindly provide the TAs some data for our future final project by giving us your hand-writing of the traditional Chinese characters for the 10 Chinese digits (with their 'lower' and 'capital' faces), each face for 5 times on the attached answer sheet. By writing down the characters and getting BBQ points, you give the instructor permission for reusing the data in this class or any future versions of NTU-Coursera classes.

22. (BBQ, 10 points) For perceptron learning model in d dimensions, prove that the growth function is given by

$$m_{\mathcal{H}}(N) = 2\sum_{i=0}^{d} \binom{N-1}{i}$$

Also find the d_{VC} of d-dimensional perceptron learning algorithm. *(Small warning: this problem is supposed to be much harder.)*

Answer guidelines. First, please write down your name and school ID number.

Name:	School ID:

Then, fill in your answers for MCQ and MRQ in the table below.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20

Lastly, please write down your solution to those (+ ...) parts and bonus problems, using as many additional pages as you want.

Each problem is of 10 points.

- For Problem with (+ ...), the answer in the table is of 3 score points, and the (+ ...) part is of 7 score points. If your solution to the (+ ...) part is clearly different from your answer in the table, it is regarded as a suspicious violation of the class policy (plagiarism) and the TAs can deduct some more points based on the violation.
- For Problem without (+ ...), the problem is of 10 points by itself and the TAs can decide to give you partial credit or not as long as it is fair to the whole class.

21.

一						壹					
二						貳					
三						參					
四						肆					
五						伍					
六						陸					
七						柒					
八						捌					
九						玖					
十						拾					