Assignment 1

Due Jan 31, beginning of class

INSTRUCTIONS. Apart from correctness, clarity and conciseness of your solutions are important. Please write your solutions legibly. You may collaborate in groups of two or three, but you must write your solution on your own and should list the name of the persons with whom you collaborated. If you need to, you are welcome to come to my office hours to discuss the problems; I may give you some hints. Solutions to a number of problems here can be found in books, research papers, or elsewhere on the web. Please do not refer to these to solve the problems. If you happen to have seen a solution please mention this in your work. If you think there is an error in some problem or need a clarification, please let me know (navin001@gmail.com).

Problem 1 (10 pts)**.** *This problem asks you to prove some properties of Rademacher complexity.*

1. *Lemma 26.6 in [SSBD] shows that for any $c \in \mathbb{R}$, we have $R(cA) \leq |c|R(A)$. Should this be an equality? Prove your answer.*
2. *For sets $A, B \subseteq \mathbb{R}^m$ define their sum by $A+B := \{a+b | a \in A, b \in B\}$. Show that $R(A+B) = R(A) + R(B)$.*
3. *For a set $A \subseteq \mathbb{R}^m$, define $A - A := \{a-b | a, b \in A\}$. Show that $R(A-A)/2 = R(A)$.*
4. *If $A \subset B$, then $R(A) \leq R(B)$. Is it true that if A is a strict subset of B then $R(A) < R(B)$ (i.e. the inequality is also strict)?*

Problem 2 (10 pts)**.** *For this problem we work in the realizable setting of PAC learning (we are given $X, Y, \mathscr{H}$). Fix error ϵ, and confidence parameter $\delta_0 \in (0,1)$ and fix a positive integer m. Suppose we have a PAC-learning algorithm A such that for any realizable distribution $\mathscr{D}$ (on $Z = X \times Y$), on input $S \sim \mathscr{D}^m$, with probability at least $1-\delta_0$ algorithm A outputs a hypothesis h such that $L_{\mathscr{D}}(h) \leq \epsilon$. Suppose that the confidence parameter δ_0 is large and so the probability $1-\delta_0$ is small, and we would like to design an algorithm that has smaller confidence parameter δ. Using A as a subroutine, design another learning algorithm that can achieve any desired confidence parameter $\delta > 0$ at the price of slightly larger error $\epsilon + \epsilon'$ and using more samples. Here $\epsilon' > 0$ is any given constant. Show that this can be done with only a logarithmic price in $1/\delta$ by proving that $\mathrm{poly}(1/\delta_0, \log(1/\delta), 1/\epsilon')$ samples suffice.*

Problem 3 (5 pts)**.** *The Bayes-optimal classifier has the minimum possible generalization error $L_{\mathscr{D}}(\cdot)$ (here we assume full knowledge of $\mathscr{D}$). Please review the definition of Bayes-optimal classifier in [SSBD, page 25] or in the notes for the first lecture. These definitions are for binary classification. For multi-class classification how should the definition of Bayes-optimal classifier be amended so that it retains*

the minimum generalization error property. Can allowing the classifier to be randomized reduce its error?

Problem 4 (10 pts)**.** *Please review the contraction principle for Rademacher complexity [SSBD, Lemma 26.9]. As discussed in the proof of Lemma 26.9 one can assume without loss of generality that* $\rho = 1$*. In this problem we outline a (seemingly new) proof of the contraction principle for* $\rho = 1$ *which we discussed very briefly in class; this proof has not been independently checked. The problem asks you to prove correctness of some of the steps in the proof marked* ***Verify****. We restrict to the case when* $A = \{a^{(1)}, \ldots, a^{(N)}\} \subseteq \mathbb{R}^m$ *is finite.*

1. *We view the problem as an optimization problem (stated below) with the variables being the functions* $\phi_1, \ldots, \phi_m$ *which in this case are fully specified by their values on* A*: for* $i \in [N]$ *function* ϕ_i *is specified by its values* $\phi_i(a_i^{(1)}), \ldots, \phi_i(a_i^{(N)})$*.*
2. ***(Verify)*** *We can assume without loss of generality that* $\phi_i(0) = 0$ *for all* $i \in [N]$*.*
3. *Consider the following mathematical program:*

$$\max R(\phi \circ A)$$
$$\text{such that } \phi_i \text{ is } 1\text{-Lipschitz for all } i$$

   Note that this is not a convex program as we are maximizing a convex function.
4. ***(Verify)*** *The set of the* ϕ_i *satisfying the constraint in the above mathematical program is convex and bounded.*
5. ***(Verify)*** *The objective function* $R(\phi \circ A)$ *is a convex function of the* ϕ_i*.*
6. *Recall that the maximum of a convex function on a convex set is achieved at extreme points of the set.*
7. *Prove that the extreme points of the set of the* ϕ_i *are given by* $\phi_i(t) = t$ *or* $\phi_i(t) = -t$*.*
8. *Conclude that the maximum of the program is given by* $R(A)$*.*

Problem 5 (5 pts)**.** *Prove that for the cross-entropy loss for logistic regression the minimum does not exist for* any realizable *training sample. What happens if the training set is not realizable?*

Problem 6 (5 pts)**.** *We defined logistic regression for binary classification, and multinomial logistic regression for multi-class classification with* k *classes for* $k \geq 2$*. In the latter, if we set* $k = 2$ *then we get a binary classification model. How is this related to logistic regression?*