CIS 625, Theory of Machine Learning
Problem Set 2

Prof. Michael Kearns

Assigned November 1, 2020; Due November 17

You are free to work in groups of up to *four* people, but every student should think about every problem, and every student should participate in their group's writeup; please indicate which student wrote up each solution. You are also free to work alone. The quality of a group's work should be commensurate with its size.

Try to provide as much detail and rigor as possible in your solutions, similar to the level in lecture. If you can't solve a problem, get as far as you can, and write down what your ideas were and where you got stuck. If you feel you need to make additional assumptions to solve a problem, carefully state what they are.

1. Let $\mathcal{C}$ be a concept class of VC dimension d. Show that there exists a distribution P and constants $\alpha, \beta > 0$ (independent of d and ϵ) such that at least $\alpha \cdot d/\log(d)$ statistical queries are required to learn $\mathcal{C}$ with error ϵ, if the tolerance of each query must be $\beta \cdot \epsilon$ or greater.

2. Show that decision lists are learnable in the SQ model (and thus in the CN model).

3. Consider the variant of the SQ model in which the learning algorithm is also given access to *unlabeled* instances x drawn from the unknown distribution P. Show that any class $\mathcal{C}$ that is learnable in this modified SQ model is still learnable in the CN model (you are free to refer to the proof from class that the standard SQ model implies CN learning). Then show that rectangles in the real plane are learnable in this modified SQ model (and thus in the CN model).

4. Consider the *malicious* PAC model, which is the same as the PAC model, except the learning algorithm has access to random examples $\langle \tilde{x}, \tilde{y} \rangle$ that are generated as follows: with probability $1 - \eta$, $\tilde{x}$ is drawn i.i.d. from P, and $\tilde{y} = y = c(x)$, where $c \in \mathcal{C}$ is the target concept; but with probability η, both $\tilde{x}$ and $\tilde{y}$ are chosen arbitrarily. In other words, with probability η the learner receives a labeled example that can be chosen in an adversarial, worst-case manner. Show that the relationship $\eta \leq \frac{\epsilon}{1+\epsilon}$ must hold in order to achieve error ϵ in the malicious PAC model.

5. Suppose $\mathcal{C}$ is PAC-learnable by an algorithm whose running time and sample size are bounded by a polynomial $p(1/\epsilon, n)$. Show that there is a constant $\alpha > 0$ such that $\mathcal{C}$ is learnable in the malicious PAC model provided $\eta \leq \alpha/p(1/\epsilon, n)$.

6. The PAC model with *membership queries* is the same as the PAC model, except in addition to random examples, at each step a learning algorithm may specify any $x \in X$ and receive the correct label $y = c(x)$ for the target concept $c \in \mathcal{C}$. The goal of learning is the same — finding an h such that the error $\epsilon(h)$ is at most ϵ with respect to c and P. Show that the class of monotone DNF — boolean functions given by an expression of the form $T_1 \vee T_2 \vee \cdots \vee T_s$, where each term T_i is a conjunction over the (non-negated) boolean variables $x_1, \ldots, x_n$ — is learnable in the PAC model with membership queries, in time polynomial in $1/\epsilon$, $\log(1/\delta)$, n and the number of terms s in c. Does this result imply that (non-monotone) DNF is learnable in the PAC model with membership queries? Why or why not? (Recall that in the usual PAC model, learning monotone DNF is as hard as learning DNF.)

7. Pick any heuristic algorithm L for classification that is widely used in empirical machine learning. (For instance, L could be gradient descent on some class of models, or a decision tree learning algorithm, or Boosting, Support Vector Machines, etc.) Either (a) describe an SQ variant of this algorithm that has the "same spirit" as the original, or (b) explain why you think this algorithm is doing something that fundamentally falls outside the SQ model. In the case of (a), ideally you will precisely specify the queries of your SQ variant, and in the case of (b), ideally you will describe some functions that L can learn that cannot be learned in SQ.

8. Consider the *variable noise rate* generalization of the classification noise (CN) model, in which there is an unknown sequence $\eta_1, \eta_2, \ldots, \eta_m \in [0, 1]$ such that $(1/m)\sum_{t=1}^{m} \eta_i = \eta < 1/2$. Here m is the number of examples requested by a learning algorithm, the sequence is *fixed in advance* (i.e. it is chosen before the learning algorithm is run, but is not known to the algorithm), and the ith example given to the algorithm is corrupted by classification noise with probability η_i. Thus the standard CN model is simply the special case where $\eta_i = \eta$ for all i. Show that if $\mathcal{C}$ is learnable in the CN model, it is also learnable in the variable noise rate CN model.