Introduction to Machine Learning **Fall Semester**

Homework 3: Nov 11th, 2018

Due: Nov 25th (See the submission guidelines in the course web site)

Theory Questions

1. **(10 points) Perceptron Lower Bound.** Show that for any $0 < \gamma < 1$ there exists a number $d > 0$, vector $\mathbf{w}^* \in \mathbb{R}^d$ for which $\|\mathbf{w}^*\| \leq 1$ and a sequence of examples $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)$ such that:

   (a) $\|\mathbf{x}_i\| = 1$.

   (b) $\frac{y_i \mathbf{x}_i \cdot \mathbf{w}^*}{\|\mathbf{w}^*\|} \geq \gamma$.

   (c) Perceptron makes $\left\lfloor \frac{1}{\gamma^2} \right\rfloor$ mistakes on the sequence.

   (Hint: Choose $m = d = \left\lfloor \frac{1}{\gamma^2} \right\rfloor$ and let $\{\mathbf{x}_i\}_i$ be the standard basis of $\mathbb{R}^d$)

2. **(10 points) Halving Algorithm**. Denote by $\mathcal{A}_{Halving}$ the Halving algorithm you have seen in class. Let $d \geq 6$, $\mathcal{X} = \{1, 2, ..., d\}$ and let $\mathcal{H} = \{h_{i,j} \mid 1 \leq i < j \leq d\}$ where

$$h_{i,j}(x) = \begin{cases} 1 & (x = i) \vee (x = j) \\ 0 & otherwise \end{cases}$$

   Show that $M(\mathcal{A}_{Halving}, \mathcal{H}) = 2$.

   (Definition of mistake bound $M(\mathcal{A}, \mathcal{H})$: Let $\mathcal{H}$ be a hypothesis class and $\mathcal{A}$ an online algorithm. Given any sequence $S = (x_1, h^*(x_1)), ..., (x_m, h^*(x_m))$ where m is an integer and $h^* \in \mathcal{H}$, let $M_{\mathcal{A}}(S)$ be the number of mistakes $\mathcal{A}$ makes on the sequence S. Then $M(\mathcal{A}, \mathcal{H}) = \sup_S M_{\mathcal{A}}(S)$.)

3. **(10 points) Max of Convex Functions**. Consider m convex functions $f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})$, where $f_i : \mathbb{R}^d \to \mathbb{R}$. Now define a new function $g(\boldsymbol{x}) = \max_i f_i(\boldsymbol{x})$.

   (a) Prove that $g(\boldsymbol{x})$ is a convex function.

   (b) Show that the sub-gradient of g at point $\boldsymbol{x}$ is the gradient of f_i where i is the $argmax\{f_1(x), ..., f_m(x)\}$

4. **(15 points) ℓ^2 penalty.** Consider the following problem:

$$\begin{aligned} &\min_{\mathbf{w}, b, \boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{C}{2}\sum_{i=1}^{m} \xi_i^2 \\ &\text{s.t. } \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \qquad \forall i = 1, \ldots, m \end{aligned}$$

(a) Show that a constraint of the form $\xi_i \geq 0$ will not change the problem. meaning, Show that these non-negativity constraints can be removed. That is, show that the optimal value of the objective will be the same whether or not these constraints are present.

(b) What is the Lagrangian of this problem?

(c) Minimize the Lagrangian with respect to $\mathbf{w}, b, \boldsymbol{\xi}$ by setting the derivative with respect to these variables to 0.

(d) What is the dual problem?

5. **(10 points) Interval growth function.** The goal of this exercise is to compute the growth function of the interval hypothesis class $H = \{h_{a,b} : a < b\}$ and $\chi = \Re$. Where $h(x) = 1$ if $x \in [a, b]$ else 0. In other words, your goal is to give an explicit expression to $\Pi_H(m) = max_{C:|C|=m}|H_C|$ where H_C is the restriction of H on the set C.

6. **(15 points) Sample complexity of agnostic PAC.** Let H be a hypothesis class of functions from a domain χ to $\{0, 1\}$ and let the loss function be the $0-1$ loss. Assume that $VCdim(H) = d < \infty$. Show that if

$$n \geq \frac{64d}{\epsilon^2} ln(\frac{16d}{\epsilon^2}) + \frac{16}{\epsilon^2} \ln(\frac{1}{\delta})$$

then:

$$P[e_p(ERM(S)) - min_{h \in H} e_p(h) > \epsilon] \leq \delta$$

To prove the above claim you can use the following lemma without proving it:
Lemma: Let $a \geq 1$ and $b > 0$. Then: $x \geq 4alog(2a) + 2b \rightarrow x \geq alog(x) + b$