CS 4780/5780: Machine Learning for Intelligent Systems **(Due: 11/12/19)**

Assignment #4: SGD, Learning Theory

Instructor: Nika Haghtalab, Thorsten Joachims *Name:* Student name(s), *Netid:* NetId(s)

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- Please include your name and NetIDs on the first page. We recommend typesetting your submission in LaTeX, and an Overleaf template is linked here.
- Assignments are due by noon on the due date in PDF form on Gradescope.
- Late assignments can be submitted on Gradescope up to noon Sunday, Nov 17. This is also when the solutions will be released.
- You can do this assignment in groups of 2-3. Please submit no more than one submission per group. Collaboration across groups is not permitted.
- All sources of material outside the course must be cited. The University Academic Code of Conduct will be strictly enforced.

Submission Instructions: All group members must be added to the Gradescope submission. If you're the one submitting, add your group members on Gradescope. Otherwise, make sure you are added to the submission. Put your names on the PDF, this helps us track the groups in case there are errors on Gradescope. See this Piazza post for more information.

Problem 1: SGD

(8 + 8 + 8 + 8 = 32 points)

Consider the following optimization problem on the sample $S = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_m, y_m)\}$ and $\vec{x}_i \in \mathbb{R}^d$ for all $i = 1, \ldots, m$.

$$\min_{\vec{w}} \frac{1}{2}\|\vec{w}\|_2^2 + \frac{1}{2m}\sum_{i=1}^{m} \max(0, (\vec{w} \cdot \vec{x}_i - y_i)^2 - 1)$$

(a) Write the *gradient descent* update rule with a fixed stepsize η.

(b) Write the *stochastic gradient descent* update rule with a fixed stepsize η.

(c) If m and d are very large, what would be the advantage of using stochastic gradient descent over gradient descent to solve the optimization problem? Please limit your answer to at most 2 sentences.

(d) We now move to a more general setting than the specific reguralizer and loss function described in parts (a) - (c). Instead of taking one data point to calculate the gradient per iteration in stochastic gradient descent, one can take a mini-batch of size $1 \le m' \le m$. At each time step take a random subset S_t of the data with $|S_t| = m'$. The mini-batch update rule for a general regularizer $R(\vec{w})$ and loss $L(\vec{w} \cdot \vec{x}, y)$ is:

$$\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} - \eta_t \nabla R(\vec{w}^{(t)}) - \eta_t C \frac{1}{m'} \sum_{(\vec{x},y)\in S_t} \nabla L(\vec{w}^{(t)} \cdot \vec{x}, y)$$

What happens as we increase the mini-batch size m'? Specifically, does the per-iteration computation cost increase or decrease?

The mini-batch gradient update $\frac{1}{m'}\sum_{(\vec{x},y)\in S_t} \nabla L(\vec{w}^{(t)} \cdot \vec{x}, y)$ is a random variable. Does its variance increase or decrease as the mini-batch size m' is increased?

Problem 2: Learning Theory

(11 + 11 + 11 + 11 + 11 + 13 = 68 points)

(a) For the following hypothesis classes we will consider lower bounds on the VC-dimension.

i. **Sets of k intervals over the reals:** Consider instance space $X = \mathbb{R}$, and let $H^{k\text{-}int} = \{h_{\vec{a},\vec{b}} \mid \vec{a}, \vec{b} \in \mathbb{R}^k\}$ where $h_{\vec{a},\vec{b}} = \mathbb{1}(x \in \bigcup_{i=1}^{k}(a_i, b_i))$ represents the union of k-intervals (a_i, b_i). Show that $VCDim(H^{k\text{-}int}) \geq 2k$.

ii. **Arbitrarily centered circular hypothesis:** Consider instance space $X = \mathbb{R}^2$, and let $H^{circ} = \{h_{r,\vec{c}} \mid r \in \mathbb{R}, \vec{c} \in \mathbb{R}^2\}$ where $h_{r,\vec{c}}(\vec{x}) = \mathbb{1}\{\|\vec{x} - \vec{c}\|_2 < r\}$ represents a circle of radius r at center $\vec{c}$. Show that $VCDim(H^{circ}) \geq 3$.

(b) In this problem, you are asked to prove an upper bound on the VC dimension of a hypothesis class. Let $H^{\leq k}$ be the set of all binary functions on a finite set X that assign +1 to at most k points, i.e. $h \in H^{\leq k} \subseteq \{0,1\}^{\text{X}}$ if and only if $|\{x|h(x) = 1\}| \leq k$, for a fixed $k < |\text{X}|$. Show that $VCDim(H^{\leq k}) \leq k$.

(c.) Consider any instance space X, labels $Y = \{0, 1\}$, and hypothesis classes $H_1, H_2, ..., H_k$ of binary functions on this instance space. Now consider the hypothesis class G, such that for any $h_1 \in H_1, \ldots, h_k \in H_k, g(x) = (h_1(x) \wedge h_2(x) \wedge ... \wedge h_{k/2}) \vee (h_{k/2+1} \wedge h_{k/2} + 1 ... \wedge h_k(x))$. Which one is a valid upper bound on the growth function $G[m]$, $G[m] \leq \prod_{i=1}^{k} H_i[m]$ or $G[m] \leq \sum_{i=1}^{k} H_i[m]$? Justify in at most three sentences.

(d) Let $X = \{0,1\}^N$ be the instance space of a binary classification task. This means that an instance $\vec{x}$ consists of N binary-valued features. Let H_N be the class of all boolean functions over the input domain. What is $|H_N|$ and the VC dimension of H_N? Show all your work leading you to your answer.

($\leq$) The entire input domain is $2^N \implies VC(H_N) \leq 2^N$.

($\geq$) By definition, H_N can realize all 2^{2^N} labelings of the 2^N possible instances. Thus, $VC(H_N) \geq 2^N$.

(e) So far we have dealt with finite VC dimensions, but infinite ones are also possible.

i. Does an infinite VC dimension imply an infinite hypothesis class? Explain.

ii. Does an infinite hypothesis class imply an infinite VC dimension? Explain.

(f) Let us define an instance space $X = \{(x_1, \ldots, x_{50}) \mid x_i \in \{0,1\}\}$. We have a linear hypothesis class $H = \{sign(\vec{w} \cdot \vec{x} + b) \mid \vec{w} \in \{-1,1\}^{50}, b \in \{1, \ldots, 20\}\}$.

i. For H defined this way, let

$$\hat{h}_S = \underset{h \in H}{\text{argmin}}\ err_S(h).$$

Give the smallest m_0 that you know of, for which for all $m \geq m_0$, a set S of m i.i.d. samples from any distribution P satisfies

$$P\left(|err_S(\hat{h}_S) - err_P(\hat{h}_S)| \geq 0.2\right) \leq 0.05.$$

Show all your work.

ii. Now suppose your friend, who has not seen your sample set S or the distribution P, thinks that function $\hat{h}_{friend}$ is a good hypothesis. Give the smallest m_0 that you know of, for which for all $m \geq m_0$, a set S of m i.i.d. samples satisfies

$$P\left(|err_S(\hat{h}_{friend}) - err_P(\hat{h}_{friend})| \geq 0.2\right) \leq 0.05.$$

Show all your work.