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

Assignment #4: SGD, Learning Theory

Instructor: Nika Haghtalab, Thorsten Joachims *Name:* Martin Stein, Cole Walsh, *Netid:* ms3452, cjw295

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, Oct 13. 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 (points)

(a) $\vec{w}^{(t+1)} \leftarrow (1-\eta)\vec{w}^{(t)} - \frac{\eta}{m}\sum_{i=1}^{m}(\vec{w}^{(t)}\cdot\vec{x}_i - y_i)\vec{x}_i\mathbb{1}((\vec{w}^{(t)}\cdot\vec{x}_i - y_i)^2 \geq 1)$

(b) $\vec{w}^{(t+1)} \leftarrow (1-\eta)\vec{w}^{(t)} - \frac{\eta}{m}(\vec{w}^{(t)}\cdot\vec{x}_i - y_i)\vec{x}_i\mathbb{1}((\vec{w}^{(t)}\cdot\vec{x}_i - y_i)^2 \geq 1)$ for a random sample $(x_i,\, y_i) \sim$ S

(c) Gradient descent (GD) computes the gradient for every point in S at each iteration of the algorithm, whereas stochastic gradient descent (SGD) only computes the gradient for a single point. If m and d are large, GD will take a large number of derivatives (md) at each iteration and take much longer to converge than SGD, which only computes d derivatives at each iteration.

(d) As we increase the mini-batch size m', the per-iteration computation cost increases since we have to compute the gradient for more points. Conversely, the mini-batch update's variance decreases as we increase the mini-batch size m'.

Problem 2: Learning Theory (points)

(a)

i. Let S^{2k} be a set of $2k$ training examples ordered by x: $S = \{(x_1, y_1), ..., (x_{2k}, y_{2k}) | x_1 < ... < x_{2k}\}$. Consider any pair of ordered points, (x_i, y_i) and (x_{i+1}, y_{i+1}). This pair of points can be shattered with one interval:
If x_{i-1} or x_{i+2} are outside the range of the dataset, then any point to the left of x_1 or to the right of x_{2k}, respectively, can be chosen. We have shown that any pair of points can be shattered by one interval. We can split S^{2k} into k pairs of ordered points. Each of these smaller datasets can then be shattered with one interval. Since we have k small datasets, we can shatter S^{2k} with k intervals, thus $VCDim(H^{k-int}) \geq 2k$.

h_{a_i,b_i}	y_i	y_{i+1}
$x_i < a_i, b_i < x_{i+1}$	0	0
$x_i < a_i < x_{i+1}, \quad x_{i+1} < b_i < x_{i+2}$	0	1
$x_{i-1} < a_i < x_i, \quad x_i < b_i < x_{i+1}$	1	0
$x_{i-1} < a_i < x_i, \quad x_{i+1} < b_i < x_{i+2}$	1	1

ii. Consider the set of three points: $x_1 = (-1, 0)$, $x_2 = (0, 1)$, and $x_3 = (1, 0)$. There are $2^3 = 8$ possible classifications of these points:

$h_{r,\vec{c}}$	x_1	x_2	x_3
$r = 0.5, \quad \vec{c} = (0, 0)$	0	0	0
$r = 0.5, \quad \vec{c} = (-1, 0)$	1	0	0
$r = 0.5, \quad \vec{c} = (0, 1)$	0	1	0
$r = 0.5, \quad \vec{c} = (1, 0)$	0	0	1
$r = 1.5, \quad \vec{c} = (-1, 1)$	1	1	0
$r = 1.5, \quad \vec{c} = (0, -1)$	1	0	1
$r = 1.5, \quad \vec{c} = (1, 1)$	0	1	1
$r = 1.5, \quad \vec{c} = (0, 0)$	1	1	1

There is a consistent hypothesis $h_{r,\vec{c}}$ for each possible combination of labels. Therefore, we have shown that H shatters at least one set of 3 points, establishing a lower bound of 3 for the VC Dimension of this hypothesis class: $VCDim(H^{circ}) \geq 3$.

(b) Consider a set X with $|X| = k+1$. No hypothesis in $H^{\leq k}$ can classify all points in X as $y = +1$. Concretely, the growth function $H^{\leq k}[k+1]$ is:

$$\sum_{i=0}^{k} \binom{k+1}{i} = 2^{k+1} - \binom{k+1}{k+1} = 2^{k+1} - 1$$

Thus, $H^{\leq k}[k+1] < 2^{k+1}$; $H^{\leq k}$ does not shatter X. We then have that the $VCDim(H^{\leq k}) \leq k < k+1$.

(c) $g(x)$ will only assign a label (0 or 1) to point x if that label was given to point x from some $h_i \in H_i$ as well. Therefore, some g consistent for all points can be built from the combination of possible labels from each hypothesis class H_i. Since $H_i[m]$ is the number of different ways that hypotheses in H_i can classify m points, there are at most $\prod_{i=1}^{k} H_i[m]$ different combinations of ways that all of the binary functions can assign labels to the m points (i.e., if $H_1[m] = 3$ and $H_2[m] = 4$, then there are at most 12 different combinations of assigned labels from H_1 and H_2), so $G[m] \leq \prod_{i=1}^{k} H_i[m]$ is the valid upper bound on $G[m]$.

(e)

i Yes, an infinte VC Dimension $= \max_{m \in \mathbf{N}} \{m : H[m] = 2^m\}$ implies an infinite instance space (as m needs to be infinite), which in turn implies that we need infinitely many hypotheses to create all 2^{2^m} labelings.

ii No, consider for example the hypothesis class $H = \{h_a | h_a = sign(x - a), a \in \mathbf{R}\}$. This hypothesis class is infinite, yet the VC dimension is equal to 1.

(f)

i. For an $\hat{h} \in H$ that minimizes the empirical risk with respect to S, to ensure with probability $1 - \delta$ that the error rate is within $\frac{\varepsilon}{2}$ of the error rate on the true data distribution,

$$P(|err_S(\hat{h}_S) - err_P(\hat{h}_S)| < \frac{\varepsilon}{2}) > 1 - \delta, \tag{0.1}$$

it suffices to ensure with probability $1 - \delta$ that $\forall h \in H$:

$$P(|err_S(h) - err_P(h)| < \frac{\varepsilon}{2}) > 1 - \delta. \tag{0.2}$$

This can be rewritten as:

$$P(\exists h \in H, |err_S(h) - err_P(h)| \geq \frac{\varepsilon}{2}) \leq \delta. \tag{0.3}$$

As shown in class using Hoeffding's inequality and the union bound:

$$P(\exists h \in H, |err_S(h) - err_P(h)| \geq \frac{\varepsilon}{2}) \leq 2|H|e^{\frac{-m\varepsilon^2}{2}}. \tag{0.4}$$

Rearranging to solve for m, we have:

$$m \geq \frac{2}{\varepsilon^2}(\ln\frac{2}{\delta} + \ln|H|). \tag{0.5}$$

Now, $\frac{\varepsilon}{2} = 0.2$ and $\delta = 0.05$ here, while we can find an upper bound on $|H|$ by noting that there are 2^{50} possible $\vec{w}$ and 20 possible b in our hypothesis class. Thus, we can have no more than 20×2^{50} unique classifications of points in our instance space. Substituting these values into the above equation, we arrive at $m \geq m_0 = 517$.

ii. For any $h_i \in H$, the probability that the empirical risk of h_i with respect to S is $\frac{\varepsilon}{2}$ or greater from the error rate on the true data distribution is:

$$P(|err_S(h_i) - err_P(h_i)| \geq \frac{\varepsilon}{2}) \leq 2e^{\frac{-m\varepsilon^2}{2}}. \tag{0.6}$$

Since we again want to be 95% confident ($\delta = 0.05$) that this statement holds, we can again solve for m_0 with $\frac{\varepsilon}{2} = 0.2$ and $\delta = 0.05$ to find the minimum number of samples required to satisfy:

$$\begin{aligned} &2e^{\frac{-m\varepsilon^2}{2}} \leq \delta \\ \Rightarrow &m \geq m_0 = \frac{2}{\varepsilon^2}\ln\frac{2}{\delta} \\ \Rightarrow &m_0 = 47 \end{aligned} \tag{0.7}$$