HOMEWORK 1

>>NAME HERE<<
>>WISC ID HERE<<

Instructions: You only need to hand in a pdf answer file. There is no need to submit the latex source. Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to Canvas. Please check Piazza for updates about the homework, if any.

1 Kernel [25 pts]

Consider the following simple kernel function:

$$K(x, x') = \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise.} \end{cases}$$

Suppose the input space $\mathcal{X}$ is finite.

(a) Prove that it is a legal kernel. Specifically, describe an implicit mapping $\Phi : \mathcal{X} \to \mathbb{R}^m$ (for some value m) such that $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$.

(b) In the Φ-space, any labeling of the points in $\mathcal{X}$ will be linearly separable. So, this should be perfect for learning any target function you want to: just run a kernelized version of Perceptron or SVM.

i. Why is any assignment of labels to points linearly separable?

ii. Nonetheless, what is the problem with learning with such a kernel?

2 Probability [25 pts]

Suppose $v, \phi \in \mathbb{R}^d$. For $i \notin S$, ϕ_i's are i.i.d. random variables with $\mathbb{E}[\phi_i] = 0, \mathbb{E}[\phi_i^2] = \sigma^2$, and $\mathbb{E}[|\phi_i|^3] = \nu^3$. Prove that for any $S \subset [d]$ and $b \in \mathbb{R}$,

$$\left| \Pr\left\{ \sum_{i \in [d]} \phi_i v_i \geq b \right\} - \Pr\left\{ \sum_{i \notin S} \phi_i v_i \geq b \right\} \right| \leq O\left(\frac{|\sum_{i \in S} \phi_i v_i|}{\sqrt{\sigma^2 \sum_{i \notin S} v_i^2}} + \frac{\nu^3 \sum_{i \notin S} |v_i|^3}{(\sigma^2 \sum_{i \notin S} v_i^2)^{3/2}} \right).$$

Hint: use the Berry-Esseen Theorem.

3 Application of Probability [25 pts]

Consider a two-layer neural network:

$$f(x) = \sum_{i=1}^{m} a_i \sigma(\langle w_i, x \rangle).$$

where $\sigma(z) = \max(0, z)$ is the rectified linear unit (ReLU) activation function. Initialize each entry w_{ij} with Gaussians $\mathcal{N}(0, \sigma^2)$ independently.

Prove that for any x with unit norm $\|x\|_2 = 1$, any $\delta > 0$, with probability at least $1 - \delta$ over the random initialization of $\{w_i\}$, the following is true: for any $\tau > 0$ and any $\{\Delta_i\}$ with $\|\Delta_i\|_2 \leq \tau$, there are at least $(1 - \tau/\sigma)m - \sqrt{m \log(1/\delta)}$ neurons whose activation patterns are the same for using weights $\{w_i\}$ and $\{w_i + \Delta_i\}$, i.e.,

$$\mathbb{I}[\langle w_i, x \rangle \geq 0] = \mathbb{I}[\langle w_i + \Delta_i, x \rangle \geq 0].$$

4 Optimization [25 pts]

We say a function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies the (g, h)-proxy convexity if there exist functions $g, h : \mathbb{R}^p \to \mathbb{R}$ such that for all $w, v \in \mathbb{R}^p$:

$$\langle \nabla f(w), w - v \rangle \geq g(w) - h(v).$$

(Note that a convex function satisfies the above with $g = h = f$.)

Given a distribution $\mathcal{D}$ over z, suppose that $F(w) := \mathbb{E}_{z\sim\mathcal{D}} f(w; z)$ and $f(; z)$ satisfies the $(g(\cdot; z), h(\cdot; z))$-proxy convexity for each z. Denote $H(w) := \mathbb{E}_{z\sim\mathcal{D}} h(w; z)$ and $G(w) := \mathbb{E}_{z\sim\mathcal{D}} g(w; z)$.

Consider online stochastic gradient descent on $F(w)$:

1. Pick an arbitrary initialization w_0.
2. For $t = 0, 1, \ldots, T - 1$, sample $z_t \sim \mathcal{D}$ and let $w_{t+1} = w_t - \eta \nabla f(w_t; z_t)$ where $\eta > 0$ is the step size.

Assume there exists $L_1 > 0$ such that for all w, $\mathbb{E}_{z\sim\mathcal{D}}[\|\nabla f(w; z)\|^2] \leq L_1^2$. Prove that for any $v \in \mathbb{R}^p, \epsilon > 0$, with properly set η and T, we have

$$\min_{t<T} \; \mathbb{E}_{(z_0,z_1,\ldots,z_{t-1})\sim\mathcal{D}^t} G(w_t) \;\; \leq \;\; H(v) + \epsilon.$$