Theoretical Machine Learning - COS 511

Homework Assignment 4

Due Date: 26 Apr 2017, till 22:00

(1) Solve 4 out of the following 5 problems.

(2) Consulting other students from this course is allowed. In this case - clearly state whom you consulted with for each problem separately.

(3) Searching the internet or literature for solutions, other than the course lecture notes, is NOT allowed.

(4) All problems are weighted equally at 10 points each. Indicate on your problem set which four problems you choose to solve. Feel free to write down solutions for the other two as well, but your homework grade will only depend upon the four you mark to be graded.

Ex. 1:

Given a convex set $\mathcal{K}$ let

$$\Pi_{\mathcal{K}}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathcal{K}} \|\mathbf{x}-\mathbf{y}\|$$

(1) Prove that for every $\mathbf{z} \in \mathcal{K}$:

$$\|\mathbf{y}-\mathbf{z}\| \geq \|\mathbf{y}-\Pi_{\mathcal{K}}(\mathbf{y})\|$$

(2) Show that the projection onto a convex set is well defined. i.e. for every $\mathbf{x}$, there exists a unique $\mathbf{y}$ such that $\mathbf{y} = \arg\min_{\mathbf{y}\in\mathcal{K}} \|\mathbf{x}-\mathbf{y}\|$

Ex. 2:

Repeat the analysis of SGD only this time without the projection step (alternatively assume

$\mathcal{K} = \mathbb{R}^n$ the whole space). Show that with a fixed step size $\eta_t = \frac{D}{G\sqrt{T}}$ We obtain also:

$$f(\bar{\mathbf{x}}_T) \leq \min_{\mathbf{x}^* \in \mathcal{K}} + O(\frac{DG}{\sqrt{T}})$$

Ex. 3:

Prove that the kernel function $k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = (\mathbf{x}^{(1)} \cdot \mathbf{x}^{(2)})^d$ indeed defines a kernel. (If you use any non-trivial statement not used in class, you need to prove it. Try to avoid using such statements!!!).

Ex. 4:

(1) Let $F = \{f_1, \ldots, f_r\}$ be a finite set of feature functions (i.e. each $f_i : \mathcal{X} \to \mathbb{R}$). Let $\Delta = \{\mu : \sum_{i=1}^m \mu(i) = 1,\ \mu > 0\}$ be the class of distributions over $[r]$ which we will identify as distributions over features (according to $\mu(f_i) = \mu(i)$). A distribution in $\mu \in \Delta$ operates over a point $x \in \mathcal{X}$ according to

$$\mu(x) = \mathop{\mathbb{E}}_{f_i \sim \mu}[f_i(x)] = \sum \mu(i) f_i(x).$$

(2) Assume that F is the class of all monomials of degree d over $\mathcal{X} = \mathbb{R}^n$. Show that for any convex L-Lipschitz loss function the problem $(\mathcal{X}, \Delta, \ell)$ is learnable (not necessarily efficiently) with sample complexity $O\left(L\sqrt{\frac{2d \log n}{m}}\right)$.

Ex. 5:

Let $k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = e^{-\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|^2/\sigma}$. Show that for a finite sample $S = ((\mathbf{x}^{(1)}, y_1), \ldots, (\mathbf{x}^{(m)}, y_m)$, for sufficintly small σ (may depend on the sample): For any vector $\mathbf{y}$ there exists $f \in H$ such that

$$\sum_{i=1}^m \frac{1}{m} \|f(\mathbf{x}^{(i)}) - y_i\|^2 \leq \epsilon$$

Prove that for some $\mathbf{y}$ we must have $\|f\| = \Omega(\sqrt{m})$

Ex. 6:

Use hardness results for agnostic learning of half-spaces to show the following: Let $\mathcal{X} = \mathbb{R}^n$ be a domain, H an RKHS such that for every $\|\mathbf{w}\| < 1$ there exists $f_{\mathbf{w}} \in H$ such that

$$\sup_{x \in \mathcal{X}} |f(\mathbf{x}) - \text{sgn}(\mathbf{w}^\top \mathbf{x})| < \epsilon$$

and let $c > 0$ be a constant. Then there exist some distribution D and vector $\|\mathbf{w}^*\| < 1$ such that:

$$\|f_{\mathbf{w}^*}\|_H = \Omega(\text{poly}(d, 1/\epsilon, \log 1/\delta)^c)$$