Mehryar Mohri
Foundations of Machine Learning 2024
Courant Institute of Mathematical Sciences
Homework assignment 3
November 5, 2024
Due: November 19, 2024

A Kernels

1. For $\alpha \geq 0$, the kernel $K_\alpha : (x, x') \mapsto \sum_{k=1}^{N} \min(|x_k|^\alpha, |x'_k|^\alpha)$ over $\mathbb{R}^N \times \mathbb{R}^N$ is used in image classification. Show that K_α is PDS. To do that, you can proceed as follows.

   (a) Use the fact that $(f, g) \mapsto \int_{t=0}^{+\infty} f(t)g(t)dt$ is an inner product over the set of measurable functions over $[0, +\infty)$ to show that $(x, x') \mapsto \min(x, x')$ is a PDS kernel (hint: associate an indicator function to x and another one to x').

   (b) Use the previous question to show that K_1 is PDS and similarly K_α with other values of α.

B Boosting

1. In class, we showed that AdaBoost can be viewed as coordinate descent applied to a convex upper bound on the empirical error. Here, we consider instead an algorithm seeking to minimize the empirical margin loss. For any $0 \leq \rho < 1$, using the same notation as in class, let $\widehat{R}_\rho(f) = \frac{1}{m} \sum_{i=1}^{m} 1_{y_i f(x_i) \leq \rho}$ denote the empirical margin loss of a function f of the form $f = \frac{\sum_{t=1}^{T} \alpha_t h_t}{\sum_{t=1}^{T} \alpha_t}$ for a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$.

   (a) Prove the upper bound $\widehat{R}_\rho(f) \leq \exp\left(\sum_{t=1}^{T} \alpha_t \rho\right) \prod_{t=1}^{T} Z_t$, where the normalization factors Z_t are defined as in the case of AdaBoost in class.

   (b) Give the expression of Z_t as a function of ρ and ϵ_t, where the weighted error ϵ_t are defined as in the case of AdaBoost in class. Use that to prove the following upper bound

$$\widehat{R}_\rho(f) \leq \exp\left(- \sum_{t=1}^{T} \mathsf{D}\Big(\frac{1-\rho}{2} \Big\| \epsilon_t\Big)\right),$$

   where $\mathsf{D}(p\|q)$ denotes the binary relative entropy of p and q: $\mathsf{D}(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$, for any $p, q \in [0, 1]$.

   (c) Assume that for all $t \in [1, T]$, $\frac{1-\rho}{2} - \epsilon_t > \gamma > 0$. Use the result of the previous question to show that

$$\widehat{R}_\rho(f) \leq \exp\big(- 2\gamma^2 T\big).$$

   (hint: you can use Pinsker's inequality: $\mathsf{D}(p\|q) \geq 2(p-q)^2$ for all $p, q \in [0, 1]$). Show that for $T > \frac{\log m}{2\gamma^2}$, all points have margin at least ρ.

C Maxent

1. Derive optimization problem of L_2-regularized Maxent with Mahalonobis distance (counterpart of Maxent with relative entropy). Show that it is a convex optimization problem.

2. Give the general form of the solution (it might be useful to use Lagrange function and representer theorem).

3. Derive dual problem and equivalence (Lagrange duality).