Mehryar Mohri
Foundations of Machine Learning 2020
Courant Institute of Mathematical Sciences
Homework assignment 3
Nov 14, 2020
Due: Nov 24, 2020 [before class starts].

A. Kernel methods

1. Graph kernel. Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. $\mathcal{V}$ could represent a set of documents or biosequences and E the set of connections between them. Let $w[e] \in \mathbb{R}$ denote the weight assigned to edge $e \in \mathcal{E}$. The weight of a path is the product of the weights of its constituent edges. Show that the kernel K over $\mathcal{V} \times \mathcal{V}$ where $K(p, q)$ is the sum of the weights of all paths of length two between p and q is PDS (*Hint*: you could introduce the matrix $W = (W_{pq})$, where $W_{pq} = 0$ when there is no edge between p and q, W_{pq} equal to the weight of the edge between p and q otherwise).

2. Pixel kernel. This problem consists of showing that a kernel useful in applications is PDS.

   (a) Show that $S\colon (z, z') \mapsto \int_0^{+\infty} 1_{t\in[0,z]} 1_{t\in[0,z']}\, dt$ is a PDS kernel defined over $\mathbb{R} \times \mathbb{R}$.

   (b) Use the previous question to prove that the kernel $K_\mu\colon (\mathbf{x}, \mathbf{x}') \mapsto \prod_{k=1}^N e^{\min(|x_k|^\mu, |x'_k|^\mu)}$, where x_k is the kth coordinate of $\mathbf{x}$, is a PDS kernel defined over $\mathbb{R}^N \times \mathbb{R}^N$, for any $\mu \geq 0$.

B. Boosting

1. Logistic loss boosting.

   (a) Show that $\Phi\colon u \mapsto \log_2(1 + e^{-u})$ defines a convex and decreasing function upper-bounding $u \mapsto 1_{u\leq 0}$.

   (b) Let $\mathcal{H} = \{h_1, \ldots, h_N\}$ be a finite hypothesis set serving as a family of base predictors taking values in $\{-1, +1\}$. Define the objective $F(\boldsymbol{\alpha})$ for a boosting-type algorithm using Φ, instead of the exponential function used in AdaBoost, for a given sample of size m and show that it is a convex function of $\boldsymbol{\alpha}$.

(c) Determine the best direction at iteration t if you apply coordinate descent to F. You should adopt a notation similar to the one used in class and define D_t, Z_t, and $\epsilon_{t,k}$ for $t \in [T]$ and $k \in [N]$. What should be the boosting condition on $\epsilon_{t,k}$ for the best direction k?

(d) Show that for any $(u, v) \in \mathbb{R}^2$, the following inequality holds:

$$\Phi(u + v) - \Phi(u) \leq -\Phi'(u)(e^{-v} - 1).$$

(e) Use that to show the following:

$$F(\alpha_{t-1} + \eta \mathbf{e}_k) - F(\alpha_{t-1}) \leq \frac{1}{m} \sum_{i=1}^{m} D_t(i) Z_t [e^{-\eta y_i h_k(x_i)} - 1].$$

(f) To determine the best step η for a given direction k, we can minimize the upper bound of the previous question. Show that this is syntactically the same minimization as the one for finding the best step in AdaBoost. Give the expression of the step at iteration t.

(g) Give the full pseudocode of the algorithm.

(h) Give a margin-based generalization bound for this algorithm.

(i) Compare empirically AdaBoost and this logistic loss boosting algorithm on the same dataset as the one used in the previous homework by choosing T via cross-validation as multiples of 100.