Homework 9

Due April 17 at 11pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. Show that the function $(x, x') \in \mathbb{R}^+ \times \mathbb{R}^+ \mapsto \min(x, x')$ is a kernel. Hint: use that
$$\min(x, x') = \int_0^\infty \mathbf{1}\{s \leq \min(x, x')\}\mathrm{d}s = \int_0^\infty \mathbf{1}\{s \leq x\}\mathbf{1}\{s \leq x'\}\mathrm{d}s.$$

2. Find an example of a (semi-definite) kernel, but such that we do not have $k(x, x') \geq 0$ for all x, x'. Find an example of a function k such that $k(x, x') \geq 0$ for all x, x', but k is not a (semi-definite) kernel.

3. (Sparse Gram matrix approximation) Let $(\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_n}, \mathbf{y_n})$ be a training sample of n observations with $\mathbf{x_i} \in \mathcal{X}$ and $\mathbf{y_i} \in \mathbb{R}$. Let k be a kernel on $\mathcal{X}$, let $\mathbf{G}$ be the Gram matrix associated with the observations

and let $\mathbf{Y} = (\mathbf{y_1}, \dots, \mathbf{y_n})^\top \in \mathbb{R}^n$. Given a positive parameter λ, the predictor given by kernel ridge regression is $x \mapsto \sum_{i=1}^n \hat{a}_i k(x, \mathbf{x_i})$, where

$$\hat{a} = (\mathbf{G} + \lambda n \mathrm{Id}_n)^{-1} \mathbf{Y}.$$

Computing $\hat{a}$ requires to store $\mathbf{G}$ (n^2 entries) and to inverse a $n \times n$ matrix (n^3 complexity). To speed up the process, we are going to consider a low-rank approximation of $\mathbf{G}$.

(a) Let $\Phi : \mathcal{X} \to \mathcal{H}$ be a feature map associated with k. Let $0 \leq m \leq n$ and let B be a $n \times m$ matrix. We approximate the vector $\Phi(\mathbf{x_i}) \in \mathcal{H}$ using only the m first vectors $\Phi(\mathbf{x_1}), \dots, \Phi(\mathbf{x_m})$ by defining $\tilde{\Phi}_B(\mathbf{x_i}) = \sum_{j=1}^m B_{ij} \Phi(\mathbf{x_j})$. Define the reconstruction error $\mathrm{Err}(B) := \sum_{i=1}^n \|\tilde{\Phi}_B(\mathbf{x_i}) - \Phi(\mathbf{x_i})\|_{\mathcal{H}}^2$. Show that $\mathrm{Err}(B)$ is equal to

$$\sum_{i=1}^n \left(\mathbf{G}_{ii} - 2 \sum_{j=1}^m B_{ij} G_{i,j} + \sum_{1 \leq j, j' \leq m} B_{ij} B_{ij'} G_{jj'} \right).$$

(b) Let $\mathbf{G}^{nm}$ be the $n \times m$ matrix obtained by taking the m first columns of $\mathbf{G}$. Also, let $\mathbf{G}^{mm}$ be the $m \times m$ matrix obtained by taking the first m rows and first columns of m. Assume that $\mathbf{G}^{mm}$ is invertible. Show that $\mathrm{Err}(B)$ is minimized for $B = \mathbf{G}^{nm}(\mathbf{G}^{mm})^{-1}$. Hint: compute the partial derivatives of $\mathrm{Err}(B)$ with respect to each of the entry $B_{i_0 j_0}$ for $1 \leq i_0 \leq n$ and $1 \leq j_0 \leq m$. The minimum is attained when the gradient is zero.

(c) Consider the feature map $\tilde{\Phi}_B$ with $B = \mathbf{G}^{nm}(\mathbf{G}^{mm})^{-1}$. Show that, for this feature map, the Gram matrix associated with the observations $(\mathbf{x_1}, \dots, \mathbf{x_n})$ is equal to

$$\tilde{\mathbf{G}} = \mathbf{G}^{nm}(\mathbf{G}^{mm})^{-1}(\mathbf{G}^{nm})^\top.$$

(d) Let U be a $n \times m$ matrix, V a $m \times m$ invertible matrix and W be a $m \times n$ matrix. Assume that $(\mathrm{Id}_n + UVW)$ is invertible. The Sherman-Woodbury-Morrison formula states that

$$(\mathrm{Id}_n + UVW)^{-1} = \mathrm{Id}_n - U(V^{-1} + WU)^{-1} W.$$

Assume that multiplying a $m \times n$ matrix by a $n \times p$ matrix requires $O(mnp)$ operations, and that inversing a $m \times m$ matrix requires

$O(m^3)$ operations. Using those different elements, show that kernel ridge regression with feature map $\tilde{\Phi}_B$ can be computed using $O(nm^2)$ operations. What is the spatial complexity required to store $\tilde{\mathbf{G}}$?