## Section 07: Kernels

Let  $\phi \colon \mathbb{R}^d \to \mathbb{R}^k$  be a feature map, define K as the kernel function, and define G to be the kernel matrix of  $\phi$ .

- (a) The kernel matrix is symmetric, that is, show  $G_{i,j} = G_{j,i}$ .
- (b) The kernel matrix G is positive semi-definite, that is, for any column vector x,  $x^{\top}Gx \geq 0$ .
- (c) *Mercer's* theorem: A function  $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$  is a valid kernel if and only if the corresponding kernel matrix G is symmetric and positive definite.

## 1. Kernelized Linear Regression

Recall that the definition of a kernel is the following:

**Definition 1.** A function  $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$  is a *kernel* for a map  $\phi$  if  $K(x, x') = \phi(x) \cdot \phi(x') = \langle \phi(x), \phi(x') \rangle$  for all x, x'.

Consider regularized linear regression (without a bias, for simplicity). Our objective to find the optimal parameters  $\hat{w} = \arg\min_{w} L(w)$  for a dataset  $(x_i, y_i)_{i=1}^n$  that minimize the following loss function:

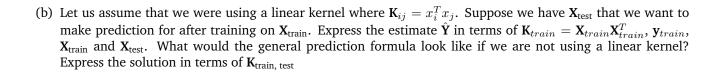
$$L(w) = \sum_{i=1}^{n} (w^{T} x_{i} - y_{i})^{2} + \lambda ||w||_{2}^{2}$$

Note that from class, we know there is an optimal  $\hat{w}$  that lies in the span of the datapoints. Concretely, there exist  $\alpha_1,...,\alpha_n \in \mathbb{R}$  such that  $\hat{w} = \sum_i^n \alpha_i x_i$ . Also recall from lecture that the expression of our loss function L(w) in terms of the kernel is:

$$L(w) = ||\mathbf{y} - \mathbf{K}\alpha||_2^2 + \lambda \alpha^T \mathbf{K}\alpha$$

This derivation can be seen here on slide 53.

(a) Solve for the optimal  $\hat{\alpha}$ .



## **2.** Proving $\hat{w} \in \text{Span}(x_1, ..., x_n)$

We will prove this through contradiction. Assume  $\hat{w} \notin \operatorname{span}(x_1,...,x_n)$  solves  $\operatorname{arg\,min}_w L(w)$ . Then, there exists a component of  $\hat{w}$  that is perpendicular to the span, which we will call  $w^{\perp}$ . Concretely,

$$\hat{w} = \bar{w} + w^{\perp}$$

Where  $\bar{w} = \sum_{i=1}^{n} \alpha_{i} x_{i}$  is the component of  $\hat{w}$  in the span of the datapoints.

To show that  $w^{\perp}$  is part of our optimal parameters, we need to consider both the error term and the regularization term of L(w). Since  $\bar{w}$  and  $w^{\perp}$  are perpendicular to each other, their contribution to L(w) can be minimized independently. Let us split the error and regularization terms into their  $\bar{w}$  and  $w^{\perp}$  components.

(a) First, we will find the optimal hyperparameter selection for the error term of our loss function in terms of  $\bar{w}$  and  $w^{\perp}$ . Show that  $\hat{w} \cdot x_i = \bar{w} \cdot x_i$ , for every  $x_i$ . (Hint: what is the relationship of  $w^{\perp}$  and  $x_i$ )

(b) We have shown that for the optimal solution, the error term relies only on  $\operatorname{Span}(x_1,...x_n)$ . Let us find the regularization term in terms of  $\bar{w}$  and  $w^{\perp}$  and the range of values it can take. Now, show that  $||\hat{w}||_2^2 \ge ||\bar{w}||_2^2$ .

