Section 07: Kernels

1. Kernelized Linear Regression

Consider regularized linear regression (without a bias, for simplicity). We want to find the optimal parameters $\hat{w} = \arg\min_w L(w)$ for k-featured dataset $(x_i, y_i)_{i=1}^n$ (i.e. $X \in \mathbb{R}^{n \times k}$) that minimizes the following loss function:

$$L(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda ||w||_2^2$$

From class, we know there is an optimal $\hat{w}$ that lies in the span of the datapoints. Concretely, there exists $\alpha_1, ..., \alpha_n \in \mathbb{R}$ such that $\hat{w} = \sum_i^n \alpha_i x_i$.

(a) Let $\alpha \in \mathbb{R}^n$, $\mathbf{K} \in \mathbb{R}^{n \times n}$, $\mathbf{y} \in \mathbb{R}^n$. Further, let us assume that we are using a linear kernel where $\mathbf{K}_{ij} = x_i^T x_j$. Express the loss function $L(w)$ in terms of $\mathbf{K}$ and α as $L(\alpha)$.

(b) Assuming that $\mathbf{K}$ is invertible, solve for the optimal $\hat{\alpha}$.

(c) Suppose after training our model on $\mathbf{X}_{\text{train}}$ we seek to make predictions on $\mathbf{X}_{\text{test}}$. Express these predictions $\hat{\mathbf{Y}}$ in terms of $\mathbf{K}_{train} = \mathbf{X}_{train}\mathbf{X}_{train}^T$, $\mathbf{y}_{train}$, $\mathbf{X}_{\text{train}}$, and $\mathbf{X}_{\text{test}}$. What would the general prediction formula look like if we are not using a linear kernel? Express the solution in terms of $\mathbf{K}_{\text{train, test}} = \Phi(\mathbf{X}_{test})\Phi(\mathbf{X}_{train}^T)$ and $\hat{\alpha}$, where for arbitrary d, we have $\Phi : \mathbb{R}^k \to \mathbb{R}^d$.

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

We will prove this through contradiction. Assume $\hat{w} \notin Span(x_1, ..., x_n)$ solves $\arg\min_w L(w)$ where $L(w)$ is defined above. 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 $\bar{w} = \sum_i^n \alpha_i x_i$ is the component of $\hat{w}$ in the span of the datapoints.

(a) 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) Now, show that $||\hat{w}||_2^2 \geq ||\bar{w}||_2^2$. (Hint: Use the Pythagorean Theorem)

(c) Finally, show that $\hat{w} \in Span(x_1, ..., x_n)$. (Hint: Think about the regularization term)

3. Kernel Proofs

Let $\phi : d \to k$ be a feature map, and define K to be the kernel matrix of ϕ.

(a) Prove that the kernel matrix is symmetric. That is, show $K_{i,j} = K_{j,i}$.

(b) Recall that a matrix M is positive semi-definite if $x^T M x \geq 0, \forall x \in \mathbb{R}^n$.
Show that K is positive semi-definite. (Hint: consider the matrix B where the i^{th} column of B is $\phi(x_i)$.