

Section 07: Solutions

1. Properties of Kernels

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

- (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 , $x^\top G x \geq 0$.
- (c) Mercer's theorem: A function $K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a valid kernel if and only if the corresponding kernel matrix G is symmetric and positive definite.

Solution:

Let $\phi(x_i)$ and $\phi(x_j)$ be the feature maps for x_i and x_j , respectively. Then $K_{i,j} = \phi(x_i)^\top \phi(x_j) = \phi(x_j)^\top \phi(x_i) = K_{j,i}$.

Alternatively, as the kernel itself represents a dot product and a dot product is a symmetric operation we can conclude that the kernel matrix is symmetric.

2. Kernelized Linear Regression

Recall that the definition of a kernel is the following:

Definition 1. A function $K: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a *kernel* for a map ϕ 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:

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

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

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

This derivation can be seen here on slide 53.

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

Solution:

Setting gradient of $L(w)$ with respect to α equal to 0:

$$\nabla_{\alpha} L(w) = 0$$

$$-2\mathbf{K}(\mathbf{y} - \mathbf{K}\alpha) + 2\lambda\mathbf{K}\alpha = 0$$

$$-\mathbf{K}(\mathbf{y} - \mathbf{K}\alpha) + \lambda\mathbf{K}\alpha = 0$$

$$\mathbf{K}(\mathbf{K}\alpha - \mathbf{y} + \lambda\alpha) = 0$$

$$\mathbf{K}((\mathbf{K} + \lambda I)\alpha - \mathbf{y}) = 0$$

$$\mathbf{K}(\mathbf{K} + \lambda I)\alpha = \mathbf{K}\mathbf{y}$$

$$\hat{\alpha} = (\mathbf{K} + \lambda I)^{-1}\mathbf{y}$$

- (b) Let us assume that we were using a linear kernel where $\mathbf{K}_{ij} = x_i^T x_j$. Suppose we have \mathbf{X}_{test} that we want to make prediction for after training on $\mathbf{X}_{\text{train}}$. Express the estimate $\hat{\mathbf{Y}}$ in terms of $\mathbf{K}_{\text{train}} = \mathbf{X}_{\text{train}}\mathbf{X}_{\text{train}}^T$, $\mathbf{y}_{\text{train}}$, $\mathbf{X}_{\text{train}}$ and \mathbf{X}_{test} . What would the general prediction formula look like if we are not using a linear kernel? Express the solution in terms of $\mathbf{K}_{\text{train, test}}$.

Solution:

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}_{\text{test}}\hat{w} \\ &= \mathbf{X}_{\text{test}}\mathbf{X}_{\text{train}}^T\hat{\alpha} \\ &= \mathbf{X}_{\text{test}}\mathbf{X}_{\text{train}}^T(\mathbf{K}_{\text{train}} + \lambda I)^{-1}\mathbf{y}_{\text{train}}\end{aligned}$$

General Solution for Kernel Ridge

$$\hat{\mathbf{Y}} = \mathbf{K}_{\text{train, test}}\hat{\alpha}$$

Where $\mathbf{K}_{\text{train, test}} = \mathbf{X}_{\text{test}}\mathbf{X}_{\text{train}}^T$

3. Proving $\hat{w} \in \text{Span}(x_1, \dots, x_n)$

We will prove this through contradiction. Assume $\hat{w} \notin \text{span}(x_1, \dots, x_n)$ solves $\arg \min_w L(w)$. 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 $\bar{w} = \sum_i^n \alpha_i x_i$ 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)

Solution:

$$\begin{aligned}
\hat{w} \cdot x_i &= (\bar{w} + w^\perp) \cdot x_i \\
&= \bar{w} \cdot x_i + w^\perp \cdot x_i \\
&= \bar{w} \cdot x_i + 0 && w^\perp \text{ is perpendicular to each } x_i \\
&= \bar{w} \cdot x_i
\end{aligned}$$

(b) We have shown that for the optimal solution, the error term relies only on $\text{Span}(x_1, \dots, x_n)$. 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 $\|\hat{w}\|_2^2 \geq \|\bar{w}\|_2^2$.

Solution:

$$\begin{aligned}
\|\hat{w}\|_2^2 &= \|\bar{w} + w^\perp\|_2^2 \\
&= (\bar{w} + w^\perp)^T (\bar{w} + w^\perp) \\
&= \bar{w}^T \bar{w} + 2\bar{w}^T w^\perp + (w^\perp)^T w^\perp \\
&= \|\bar{w}\|_2^2 + \|w^\perp\|_2^2 && \text{as } \bar{w}^T w^\perp = \langle \bar{w}, w^\perp \rangle = 0 \\
&\geq \|\bar{w}\|_2^2
\end{aligned}$$

(c) We now know the minimum value of the regularization term and what it is equal to with respect to \hat{w} and w^\perp . Finally, show that $\hat{w} \in \text{Span}(x_1, \dots, x_n)$. (Hint: Think about the regularization term. What is w^\perp when the regularization term is minimized?)

Solution:

Note that in the loss function, we're trying to minimize the magnitude of w (with the regularization term $\lambda \|w\|_2^2$). Now note that if $\forall_i \hat{w}^T x_i = \bar{w}^T x_i$, and $\|\hat{w}\|_2^2 \geq \|\bar{w}\|_2^2$, then our optimization will always choose $w^\perp = 0$ (as we favor smaller solutions), meaning that $\hat{w} = \bar{w}$ and $\hat{w} \in \text{Span}(x_1, \dots, x_n)$, which completes the contradiction.

Remark For running Jupyter Notebook locally, the following are required:

- Jupyter Notebook [Install] [Document]
- ipywidgets
- pytorch
- matplotlib
- numpy