Section 07: Solutions

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 $L(\alpha)$.

Solution:

$$\begin{aligned}
L(\alpha) &= \sum_i^n \left(\sum_j^n \left(\alpha_j x_j^T \right) x_i - y_i \right)^2 + \lambda || \sum_i^n \alpha_i x_i ||^2 \\
&= \sum_i^n \left(\sum_j^n \left(\alpha_j x_j^T \right) x_i - y_i \right)^2 + \lambda (\sum_i^n \alpha_i x_i)^T (\sum_i^n \alpha_i x_i) \\
&= \sum_i^n \left(\sum_j^n \alpha_j \langle x_j, x_i \rangle - y_i \right)^2 + \lambda \sum_i^n \sum_j^n \alpha_i \alpha_j x_i^T x_j \\
&= \sum_i^n \left(\sum_j^n \alpha_j K(x_j, x_i) - y_i \right)^2 + \lambda \sum_i^n \sum_j^n \alpha_i \alpha_j K(x_i, x_j) \\
&= ||\mathbf{y} - \mathbf{K}\alpha||_2^2 + \lambda \alpha^T \mathbf{K} \alpha
\end{aligned}$$

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

Solution:

Set the gradient of $L(\alpha) = 0$:

$$\nabla L(\alpha) = 0$$

So then,

$$\begin{aligned}
-2\mathbf{K}(\mathbf{y} - \mathbf{K}\alpha) + 2\lambda \mathbf{K}\alpha &= 0 \\
-\mathbf{K}(\mathbf{y} - \mathbf{K}\alpha) + \lambda \mathbf{K}\alpha &= 0 \\
\mathbf{K}(\mathbf{K}\alpha - \mathbf{y} + \lambda\alpha) &= 0 \\
\mathbf{K}((\mathbf{K} + \lambda I)\alpha - \mathbf{y}) &= 0 \\
\mathbf{K}(\mathbf{K} + \lambda I)\alpha &= \mathbf{K}\mathbf{y} \\
\hat{\alpha} &= (\mathbf{K} + \lambda I)^{-1} \mathbf{y}
\end{aligned}$$

(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 $\Phi : \mathbb{R}^k \to \mathbb{R}^d$.

Solution:

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

General Solution for Kernel Ridge

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

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)
Solution:

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

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

$$\begin{aligned}||\hat{w}||_2^2 &= ||\bar{w} + w^\perp||_2^2 \\ &= ||\bar{w}||_2^2 + ||w^\perp||_2^2 && \text{Pythagorean Theorem} \\ &\geq ||\bar{w}||_2^2\end{aligned}$$

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

Note that in the loss function we're trying to minimize the magnitude of w (with the regularization term $\lambda||w||_2^2$). 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 $w^\perp = 0$, meaning that $\hat{w} = \bar{w}$ and $\hat{w} \in Span(x_1, ..., x_n)$, which completes the contradiction.

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 $K_{i,j} = K_{j,i}$.
Solution:

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

(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 $\phi(x_i)$.
Solution:

Recall that $K_{i,j} = \phi(x_i)^T \phi(x_j)$. Observe that $K = B^T B$, as $(B^T B)_{i,j} = \phi(x_i)^T \phi(x_j)$. Now consider an arbitrary vector y. To show K is PSD it suffices to show $y^T K y$ is non-negative. We have:

$$y^T K y = y^T B^T B y = (By)^T (By) = ||By||_2^2 \geq 0$$