Section 06: Solutions

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 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:

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

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 14.

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

Solution:

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

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

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

(b) Let us assume that we were using a linear kernel where $\mathbf{K}_{ij} = x_i^T x_j$. Suppose we have $\mathbf{X}_{\text{test}}$ that we want to make prediction for after training on $\mathbf{X}_{\text{train}}$. Express the estimate $\hat{\mathbf{Y}}$ in terms of $\mathbf{K}_{train} = \mathbf{X}_{train}\mathbf{X}_{train}^T$, $\mathbf{y}_{train}$, $\mathbf{X}_{\text{train}}$ 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 $\mathbf{K}_{\text{train, test}}$ **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

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

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

2. Kernel Proofs

Let $\phi\colon \mathbb{R}^d \to \mathbb{R}^k$ 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 $K_{i,j} = \phi(x_i)^T\phi(x_j) = \phi(x_j)^T\phi(x_i) = K_{j,i}$.

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.

(b) Recall that a matrix M is positive semi-definite if $x^TMx \geq 0, \forall x \in \mathbb{R}^n$.
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 $K = B^TB$, as $(B^TB)_{i,j} = \phi(x_i)^T\phi(x_j)$. Now consider an arbitrary vector y. To show K is PSD it suffices to show y^TKy is non-negative. We have:

$$y^TKy = y^TB^TBy = (By)^T(By) = ||By||_2^2 \geq 0$$

3. 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 $\arg\min_w L(w)$. 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:

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

(b) 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) 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 $\lambda||w||_2^2$). 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 $\hat{w} \in Span(x_1, ..., x_n)$, which completes the contradiction.