2. Kernel Matrices

Problem 2.1. Consider a set of vectors $S = \{x_1, \ldots, x_m\}$. Let X denote the matrix whose rows are these vectors. Form the Gram matrix $K = XX^T$. Show that knowing K is equivalent to knowing the set of pairwise distances among the vectors in S as well as the vector lengths.

Solution From the definition of the Gram matrix, knowing K is equivalent to knowing $\langle x, x' \rangle \ \forall \ x, x' \in S$. Now let $x, x' \in S$ and observe that

$$\begin{aligned}
\langle x, x' \rangle &= \langle x' + (x - x'), x + (x' - x) \rangle \\
&= \langle x', x \rangle + \langle x', x' \rangle + \langle x', -x \rangle + \langle x, x \rangle + \langle -x', x \rangle + \langle x - x', x' - x \rangle \\
&= \|x\|^2 + \|x'\|^2 - \|x - x', x - x'\|^2 - \langle x', x \rangle \\
&= \frac{\|x\|^2 + \|x'\|^2 - \|x - x', x - x'\|^2}{2}
\end{aligned}$$

Thus it is possible to express elements of K in terms of the distance between vectors in S as well as the vector lengths. We can conclude that knowing K is equivalent to knowing the set of pairwise distances among the vectors in S as well as the vector lengths.

3. Kernel Ridge Regression

Problem 3.1. Show that for w to be a minimizer of $J(w)$, we must have $X^TXw + \lambda Iw = X^Ty$. Show that the minimizer of $J(w)$ is $w = (X^TX + \lambda I)^{-1}X^Ty$. Justify that the matrix $X^TX + \lambda I$ is invertible, for $\lambda > 0$.

Solution

1. For w to be a minimizer of $J(w)$ we must have

$$\begin{aligned} J'(w) &= 2(Xw - y)^TX + 2\lambda w^T \\ &= 2(Xw)^TX - 2y^TX + 2\lambda w^T \\ &= 2w^TX^TX - 2y^TX + 2\lambda w^T \\ &= 0 \end{aligned}$$

So then

$$y^TX = w^TX^TX + \lambda w^T$$

And by taking the transpose of both sides we get

$$X^Ty = X^TXw + \lambda w$$

2. Solving for w we get

$$\begin{aligned} X^Ty &= (X^TX + \lambda I)w \\ w &= (X^TX + \lambda I)^{-1}X^Ty \end{aligned}$$

3. Note that X^TX is positive semi-definite. Let $v \in \mathbf{R}^d$ and $\lambda \in \mathbf{R}$ and observe that

$$\begin{aligned} v^T(X^TX + \lambda I)v &= v^TX^TXv + \lambda v^TX^TXv \\ &= v^TX^TXv + \lambda v^Tv \end{aligned}$$

But $v^TX^TXv \geq 0$ since X^TX is positive semi-definite and if $v \neq 0$ then $\lambda v^Tv > 0$. Thus $v^T(X^TX + \lambda I)v > 0$ and so $X^TX + \lambda I$ is symmetric positive definite. As a result it is also inevitable.

Problem 3.2. Rewrite $X^TXw + \lambda Iw = X^Ty$ as $w = \frac{1}{\lambda}(X^Ty - X^TXw)$. Based on this, show that we can write $w = X^T\alpha$ for some α, and give an expression for α.

Solution Note that

$$w = \frac{1}{\lambda}(X^Ty - X^TXw)$$
$$w = \frac{1}{\lambda}X^T(y - Xw)$$

Then setting $\alpha = \frac{1}{\lambda}(y - Xw)$ we have

$$w = X^T\alpha$$

Problem 3.3. Based on the fact that $w = X^T\alpha$, explain why we say w is "in the span of the data."

Solution Since $w = X^T\alpha$, we can also express w as $w = \sum_{i=1}^{n} x_i\alpha_i$. Thus w is a linear combination of the training data X and as such is in the span of the training data.

Problem 3.4. Show that $\alpha = (\lambda I + XX^T)^{-1}y$. Note that XX^T is the kernel matrix for the standard vector dot product.

Solution Observe that

$$\begin{aligned}
\alpha &= \frac{1}{\lambda}(y - Xw) \\
\alpha &= \frac{1}{\lambda}(y - XX^T\alpha) \\
\lambda\alpha &= y - XX^T\alpha \\
\lambda\alpha + XX^T\alpha &= y \\
(\lambda I + XX^T)\alpha &= y \\
\alpha &= (\lambda I + XX^T)^{-1}y
\end{aligned}$$

Problem 3.5. Give a kernelized expression for the Xw, the predicted values on the training points.

Solution Recall that

$$\alpha = (\lambda I + XX^T)^{-1}y$$

Then

$$\begin{aligned} Xw &= XX^T\alpha \\ &= XX^T(\lambda I + XX^T)^{-1}y \\ &= K(\lambda I + K)^{-1}y \end{aligned}$$

Problem 3.6. Give an expression for the prediction $f(x) = x^T w^*$ for a new point x, not in the training set. The expression should only involve x via inner products with other x's. [Hint: It is often convenient to define the column vector

$$k_x = \begin{pmatrix} x^T x_1 \\ \vdots \\ x^T x_n \end{pmatrix}$$

to simplify the expression.]

Solution Recall that $w^* = X^T \alpha^*$ and that $\alpha^* = (\lambda I + XX^T)^{-1} y$. Also note that

$$x^T X^T = \begin{pmatrix} x^T x_1 \\ \vdots \\ x^T x_n \end{pmatrix} = k_x$$

Then

$$\begin{aligned} f(x) &= x^T w^* \\ &= x^T X^T \alpha^* \\ &= x^T X^T (\lambda I + XX^T) y \\ &= k_x (\lambda I + K) y \end{aligned}$$

4. Pegasos and SSGD for ℓ_2-regularized ERM

Problem 4.1. For each $i = 1, \ldots, n$, let $g_i(w)$ be a subgradient of $J_i(w)$ at $w \in \mathbf{R}^d$. Let $v_i(w)$ be a subgradient of $\ell_i(w)$ at w. Give an expression for $g_i(w)$ in terms of w and $v_i(w)$

Solution

$$g_i(w) = \lambda w + v_i(w)$$

Problem . Show that $\mathbb{E}g_i(w) \in \partial J(w)$, where the expectation is over the randomly selected $i \in 1, \ldots, n$.

Solution Let $g(w)$ be a subgradient of $J(w)$ and $v_i(w)$ a subgradient of $l_i(w)$ Now observe that

$$\begin{aligned} g(w) &= \lambda w + \frac{1}{n}\sum_{i=1}^{n} v_i(w) \\ &= \lambda w + \mathbb{E}v_i(w) \end{aligned}$$

And that

$$\begin{aligned} \mathbb{E}g_i(w) &= \mathbb{E}\big(\lambda w + v_i(w)\big) \\ &= \lambda w + \mathbb{E}v_i(w) \end{aligned}$$

Then $\mathbb{E}g_i(w) = g(w) \in \partial J(w)$

Problem 4.3. Now suppose we are carrying out SSGD with the Pegasos step-size $\eta^{(t)} = 1/(\lambda t)$, $t = 1, 2, \ldots$, starting from $w^{(1)} = 0$. In the t'th step, suppose we select the ith point and thus take the step $w^{(t+1)} = w^{(t)} - \eta^{(t)} g_i(w^{(t)})$. Let's write $v^{(t)} = v_i(w^{(t)})$, which is the subgradient of the loss part of $J_i(w^{(t)})$ that is used in step t. Show that

$$w^{(t+1)} = -\frac{1}{\lambda t} \sum_{\tau=1}^{t} v^{(\tau)}$$

Solution

1. Base case $t = 1$
   Observe that

$$\begin{aligned} w^{(2)} = w^{(1)} &= \frac{1}{\lambda} * \lambda w^{(1)} - \frac{1}{\lambda} v^{(1)} \\ &= -\frac{1}{\lambda} v^{(1)} \\ &= -\frac{1}{\lambda \cdot 1} \sum_{\tau=1}^{1} v^{(\tau)} \\ &= -\frac{1}{\lambda t} \sum_{\tau=1}^{t} v^{(\tau)} \end{aligned}$$

2. Induction step.
   Assume that $w^{(t)} = -\frac{1}{\lambda(t-1)} \sum_{\tau=1}^{t-1} v^{(\tau)}$. Now observe that

$$\begin{aligned} w^{(t+1)} &= w^{(t)} - \frac{1}{t} w^{(t)} - \frac{1}{\lambda t} v^{(t)} \\ &= \frac{t-1}{t} w^{(t)} - \frac{1}{\lambda t} v^{(t)} \\ &= -\frac{1}{\lambda t} \sum_{\tau=1}^{t-1} v^{(\tau)} - \frac{1}{\lambda t} v^{(t)} \\ &= -\frac{1}{\lambda t} \sum_{\tau=1}^{t} v^{(\tau)} \end{aligned}$$

Problem 4.a. Explain how Algorithm 1 can be implemented so that, if x_j has s nonzero entries, then we only need to do $O(s)$ memory accesses in every pass through the loop.

Solution If we implement the algorithm with sparse matrices or dictionaries, then when computing $y_j\langle w^{(t)}, x_j\rangle$ and $y_j x_j$ we only need to access the s nonzero elements of x_j and their corresponding elements in y_j and $w^{(t)}$. The remaining operations are $O(1)$ and so in total we only need to do $O(s)$ memory accesses.

5. Kernelized Pegasos

Problem . Kernelize the expression for the margin. That is, show that $y_j \left\langle w^{(t)}, x_j \right\rangle = y_j K_{j\cdot} \alpha^{(t)}$, where $k(x_i, x_j) = \langle x_i, x_j \rangle$ and $K_{j\cdot}$ denotes the jth row of the kernel matrix K corresponding to kernel k.

Solution Observe that

$$\begin{aligned}
\langle w^{(t)}, x_j \rangle &= \langle \sum_{i=1}^{n} \alpha_i^{(t)} x_i, x_j \rangle \\
&= \sum_{i=1}^{n} \alpha_i^{(t)} \langle x_i, x_j \rangle \\
&= \begin{bmatrix} \langle x_1, x_j \rangle & \dots & \langle x_n, x_j \rangle \end{bmatrix} \alpha^{(t)} \\
&= \begin{bmatrix} k(x_1, x_j) & \dots & k(x_n, x_j) \end{bmatrix} \alpha^{(t)} \\
&= \begin{bmatrix} k(x_j, x_1) & \dots & k(x_j, x_n) \end{bmatrix} \alpha^{(t)} \\
&= K_j \alpha^{(t)}
\end{aligned}$$

Thus $y_j \left\langle w^{(t)}, x_j \right\rangle = y_j K_{j\cdot} \alpha^{(t)}$

Problem 5.2. Suppose that $w^{(t)} = \sum_{i=1}^{n} \alpha_i^{(t)} x_i$ and for the next step we have selected a point (x_j, y_j) that does not have a margin violation. Give an update expression for $\alpha^{(t+1)}$ so that $w^{(t+1)} = \sum_{i=1}^{n} \alpha_i^{(t+1)} x_i$.

Solution When (x_j, y_j) does not result in a margin violation, $w^{(t+1)} = (1 - \eta^{(t)}\lambda)w^{(t)}$. Then in order to have $w^{(t+1)} = \sum_{i=1}^{n} \alpha_i^{(t+1)} x_i$, we must also have

$$(1 - \eta^{(t)}\lambda)w^{(t)} = \sum_{i=1}^{n} \alpha_i^{(t+1)} x_i$$

$$(1 - \eta^{(t)}\lambda) \sum_{i=1}^{n} \alpha_i^{(t)} x_i = \sum_{i=1}^{n} \alpha_i^{(t+1)} x_i$$

Now setting $\alpha^{(t+1)} = (1 - \eta^{(t)}\lambda)\alpha^{(t)}$ preserves the equality above. Thus this is a valid update rule.

Problem 5.3. Repeat the previous problem, but for the case that (x_j, y_j) has a margin violation. Then give the full pseudocode for kernelized Pegasos. You may assume that you receive the kernel matrix K as input, along with the labels $y_1, \ldots, y_n \in \{-1, 1\}$.

Solution Note that when we have a margin error,

$$\begin{aligned} w^{(t+1)} &= (1 - \eta^{(t)}\lambda)w^{(t)} + \eta^{(t)} y_j x_j \\ &= (1 - \eta^{(t)}\lambda) \sum_{i=1}^{n} \alpha_i^{(t)} x_i + \eta^{(t)} y_j x_j \\ &= \sum_{i=1}^{n} \big((1 - \eta^{(t)}\lambda)\alpha_i^{(t)} + \mathbb{1}[i = j]\eta^{(t)} y_j\big) x_i \end{aligned}$$

Thus an appropriate update rule for α is

$$\alpha_i^{(t+1)} = (1 - \eta^{(t)}\lambda)\alpha_i^{(t)} + \mathbb{1}[i = j]\eta^{(t)} y_j$$

Problem 5.4. While the direct implementation of the original Pegasos required updating all entries of w in every step, a direct kernelization of Algorithm 2, as we have done above, leads to updating all entries of α in every step. Give a version of the kernelized Pegasos algorithm that does not suffer from this inefficiency. You may try splitting the scale and direction similar to the approach of the previous problem set, or you may use a decomposition based on Algorithm 1 from the optional problem 4 above.

Solution

7. Representer Theorem

Problem 7.1. Let M be a closed subspace of a Hilbert space $\mathcal{H}$. For any $x \in \mathcal{H}$, let $m_0 = \text{Proj}_M x$ be the projection of x onto M. By the Projection Theorem, we know that $(x - m_0) \perp M$. Then by the Pythagorean Theorem, we know $\|x\|^2 = \|m_0\|^2 + \|x - m_0\|^2$. From this we concluded in lecture that $\|m_0\| \leq \|x\|$. Show that we have $\|m_0\| = \|x\|$ only when $m_0 = x$.

Solution First, assume that $m_0 = x$. Then clearly $\|m_0\| = \|x\|$.
Now it remains to show that $\|m_0\| \neq \|x\|$ when $m_0 \neq x$.

Assume that $m_0 \neq x$. Also note that since the inner product is positive-definite, $\|m_0\|^2 \neq \|x\|^2$ iff $\|m_0\| \neq \|x\|$. Then

$$\begin{aligned} \|m_0\|^2 &= \|x - (x - m_0)\|^2 \\ &= \|x\|^2 - \|x - m_0\|^2 && \text{Pythagorean Theorem} \\ &= \|x\|^2 - \langle x - m_0, x - m_0 \rangle \end{aligned}$$

Now since $x \neq m_0$, $x - m_0 \neq 0$. Then since the inner product is positive-definite, $\langle x - m_0, x - m_0 \rangle > 0$. Thus $\|m_0\|^2 \neq \|x\|^2$ and so $\|m_0\| \neq \|x\|$.

Problem 7.2. Give the proof of the Representer Theorem in the case that R is strictly increasing. That is, show that if R is strictly increasing, then all minimizers have this form claimed.

Solution Proof by contradiction:

Assume R is strictly increasing and w^* is a minimizer of $J(w)$. Let $w_M^* = Proj_M w$ where $M = span(x_1, \ldots, x_n)$. Note that for any $m \in M$, $\exists\ \alpha$ such that $m = \sum_{i=1}^n \alpha_i \psi(x_i)$. By way of contradiction also assume that $w^* \notin M$.

Since $w^* \notin M$, $w^* \neq w_M^*$. Furthermore, $\|w_M^*\| < \|w^*\|$ since projections reduce norms. Since R is strictly increasing, $R\big(\|w_M^*\|\big) < R\big(\|w^*\|\big)$. Now let $x \in M$ and observe that

$$\begin{aligned} \langle w^*, x\rangle &= \langle w_M^* + w^* - w_M^*, x\rangle \\ &= \langle w_M^*, x\rangle + \langle w^* - w_M^*, x\rangle \\ &= \langle w_M^*, x\rangle \qquad\qquad \text{since } w^* - w_M^* \perp M \end{aligned}$$

Thus

$$L\big(\langle w^*, \psi(x_1)\rangle, \ldots, \langle w^*, \psi(x_n)\rangle\big) = L\big(\langle w_M^*, \psi(x_1)\rangle, \ldots, \langle w_M^*, \psi(x_n)\rangle\big)$$

Therefore $J(w_M^*) < J(w^*)$ and so w^* is not a minimizer of $J(w)$.

We have shown that when R is strictly increasing, we cannot have any minimizers outside of M. Thus all minimizers are elements of M and can be written as $w^* = \sum_{i=1}^n \alpha_i \psi(x_i)$.

Problem 7.3. Suppose that $R : \mathbf{R}^{\geq 0} \to \mathbf{R}$ and $L : \mathbf{R}^n \to \mathbf{R}$ are both convex functions. Use properties of convex functions to **show that** $w \mapsto L\left(\langle w, \psi(x_1)\rangle, \ldots, \langle w, \psi(x_n)\rangle\right)$ is a convex function of w, and then that $J(w)$ is also a convex function of w. For simplicity, you may assume that our feature space is $\mathbf{R}^d$, rather than a generic Hilbert space. You may also use the fact that the composition of a convex function and an affine function is convex. That is, suppose $f : \mathbf{R}^n \to \mathbf{R}$, $A \in \mathbf{R}^{n\times m}$ and $b \in \mathbf{R}^n$. Define $g : \mathbf{R}^m \to \mathbf{R}$ by $g(x) = f\left(Ax + b\right)$. Then if f is convex, then so is g. From this exercise, **we can conclude** that if L and R are convex, then J does have a minimizer of the form $w^* = \sum_{i=1}^n \alpha_i \psi(x_i)$, and if R is also strictly increasing, then all minimizers of J have this form.

Solution Note that in $\mathbf{R}^n$, $\langle x, y\rangle = x^T y = y^T x$. Also note that since the sum of convex functions is also convex, if $L\left(\langle w, \psi(x_1)\rangle, \ldots, \langle w, \psi(x_n)\rangle\right)$ is convex then $J(w)$ is convex.
Now observe that

$$\begin{bmatrix} \langle w, \psi(x_1)\rangle \\ \vdots \\ \langle w, \psi(x_n)\rangle \end{bmatrix} = \begin{bmatrix} w^T\psi(x_1) \\ \vdots \\ w^T\psi(x_n) \end{bmatrix} = \begin{bmatrix} — & \psi(x_1) & — \\ & \vdots & \\ — & \psi(x_n) & — \end{bmatrix} w$$

Thus $\begin{bmatrix} \langle w, \psi(x_1)\rangle \\ \vdots \\ \langle w, \psi(x_n)\rangle \end{bmatrix}$ can be represented as an affine function of w.

Since the composition of a convex function and an affine function is also convex,
$L\left(\langle w, \psi(x_1)\rangle, \ldots, \langle w, \psi(x_n)\rangle\right)$ is a convex function of w. Therefore $J(w)$ is convex.

8. Ivanov and Tikhonov Regularization

Problem 8.1. Suppose that for some $\lambda \geq 0$ we have the Tikhonov regularization solution

$$f^* \in \arg\min_{f \in \mathcal{F}} \left[\phi(f) + \lambda\Omega(f)\right]. \tag{1}$$

Show that f^* is also an Ivanov solution. That is, $\exists r \geq 0$ such that

$$f^* \in \arg\min_{f \in \mathcal{F}} \phi(f) \text{ subject to } \Omega(f) \leq r. \tag{2}$$

Solution Suppose $f^* \in \arg\min_{f \in \mathcal{F}} \left[\phi(f) + \lambda\Omega(f)\right]$ and let $r = \Omega(f^*)$. By way of contradiction, suppose that the solution to the Ivanov form is $f' \neq f^*$. Then we have

$$\Omega(f') \leq r = \Omega(f^*)$$

and

$$\phi(f') < \phi(f^*).$$

Thus

$$\phi(f') + \lambda\Omega(f') < \phi(f^*) + \lambda\Omega(f^*).$$

This contradicts our original assumption that f^* is a minimizer of the Tikhonov problem. Thus $f' = f^*$ And so f^* is a solution to both the Tikhonov and the Ivanov form of the problem.

Problem 8.2.1. Write the Lagrangian $L(w, \lambda)$ for the Ivanov optimization problem.

Solution

$$L(w, \lambda) = \phi(w) + \lambda(\Omega(w) - r)$$

Problem 8.2.2. Write the dual optimization problem in terms of the dual objective function $g(\lambda)$, and give an expression for $g(\lambda)$.

Solution The dual optimization problem is to find

$$d^* = \sup_{\lambda \succ 0} \inf_w L(w, \lambda) = \sup_{\lambda \succ 0} g(\lambda)$$

Where $g(\lambda) = \inf_w \big(\phi(w) + \lambda(\Omega(w) - r)\big)$

Problem 8.2.3. We assumed that the dual solution is attained, so let $\lambda^* \in \arg\max_{\lambda \geq 0} g(\lambda)$. We also assumed strong duality, which implies $\phi(w^*) = g(\lambda^*)$. Show that the minimum in the expression for $g(\lambda^*)$ is attained at w^*. **Conclude the proof** by showing that for the choice of $\lambda = \lambda^*$, we have $w^* \in \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda^* \Omega(w)\right]$.

Solution Observe that

$$\begin{aligned}
\phi(w^*) &= g(\lambda^*) \\
&= \inf_w L(w, \lambda^*) \\
&\leq L(w^*, \lambda^*) \\
&= \phi(w^*) + \lambda^*(\Omega(w^*) - r) \\
&\leq \phi(w^*) && \text{since } \Omega(w^*) - r \leq 0
\end{aligned}$$

Thus we have equality throughout the expression above and so $\inf_w L(w, \lambda^*) = L(w^*, \lambda^*)$. We can now conclude that, since $\phi(w^*) = \inf_w \left[\phi(w^*) + \lambda^*(\Omega(w^*) - r)\right]$,

$$\begin{aligned}
w^* &\in arg\, min_w\ \phi(w) + \lambda^*(\Omega(w) - r) \\
&= arg\, min_w\ \phi(w) + \lambda^* \Omega(w).
\end{aligned}$$

Problem 8.3. Show that the Ivanov form of ridge regression

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \\ \text{subject to} \quad & w^T w \leq r. \end{aligned}$$

is a convex optimization problem with a strictly feasible point, so long as $r > 0$. (Thus implying the Ivanov and Tikhonov forms of ridge regression are equivalent when $r > 0$.)

Solution Assume that $r > 0$.
To show that Slater's condition holds, we need to find a w such that

$$w^T w - r < 0$$

Let $w = 0$, then since $r > 0$, we have $w^T w - r < 0$. Thus a solution has been found and Slater's condition holds.