CSC411H1 L0101, Fall 2018
Assignment 7

Name: JingXi, Hao
Student Number: 1000654188

Due Date: 5th December, 2018 11:59pm

1. **Representer Theorem.** In this question, you'll prove and apply a simplified version of the Representer Theorem, which is the basis for a lot of kernelized algorithms. Consider a linear model:

$$z = \mathbf{w}^T \psi(\mathbf{x})$$
$$y = g(z)$$

, where ψ is a feature map and g is some function (e.g. identity, logistic, etc.). We are given a training set $\{\mathbf{x}^{(i)}, t^{(i)}\}_{i=1}^N$. We are interested in minimizing the expected loss plus an L_2 regularization term:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^{(i)}, t^{(i)}) + \frac{\lambda}{2} ||\mathbf{w}||^2$$

, where $\mathcal{L}$ is some loss function. Let $\mathbf{\Psi}$ denote the feature matrix

$$\mathbf{\Psi} = \begin{bmatrix} \psi(\mathbf{x^{(1)}})^T \\ \vdots \\ \psi(\mathbf{x^{(N)}})^T \end{bmatrix}$$

Observe that this formulation captures a lot of the models we've covered in this course, including linear regression, logistic regression, and SVMs.

(a) Show that the optimal weights must lie in the row space of $\mathbf{\Psi}$

Hint: Given a subspace $\mathbf{S}$, a vector $\mathbf{v}$ can be decomposed as $\mathbf{v} = \mathbf{v}_S + \mathbf{v}_\perp$, where $\mathbf{v}_S$ is the projection of $\mathbf{v}$ onto $\mathbf{S}$, and $\mathbf{v}_\perp$ is orthogonal to $\mathbf{S}$. (You may assume this fact without proof, but you can review it here3.) Apply this decomposition to $\mathbf{w}$ and see if you can show something about one of the two components.

Solution:
Given a subspace, $\mathbf{S}$, which is the row space of $\mathbf{\Psi}$, let $\mathbf{w}$ be the optimal weights, where $\mathbf{w} = \mathbf{w}_S + \mathbf{w}_\perp$ by hint. Then, we rewrite our cost function which is

$$\begin{aligned}\mathcal{J}(\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y^{(i)}, t^{(i)}) + \frac{\lambda}{2}||\mathbf{w}||^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(g(\mathbf{w}^T\psi(\mathbf{x}^{(i)})),\ t^{(i)}) + \frac{\lambda}{2}||\mathbf{w}||^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(g((\mathbf{w}_S + \mathbf{w}_\perp)^T\psi(\mathbf{x}^{(i)})),\ t^{(i)}) + \frac{\lambda}{2}||\mathbf{w}||^2\end{aligned}$$

. Then, let $\mathbf{w}^*$ be the optimal weights, where by hint $\mathbf{w}^* = \mathbf{w}_S^* + \mathbf{w}_\perp^*$. In order to compute the optimal weights, we need to minimize the cost function, $\mathcal{J}(\mathbf{w})$. In which case, the decomposition component $\mathbf{w}_\perp^*$ is a zero vector. Therefore, this implies that $\mathbf{w}^* = \mathbf{w}_S^*$. We know that $\mathbf{w}_S^*$ is the projection of $\mathbf{w}^*$ onto $\mathbf{S}$. Hence, $\mathbf{w}_S^*$ lies in the row space of $\mathbf{\Psi}$, which implies that the optimal weights, $\mathbf{w}^*$, lies in the row space of $\mathbf{\Psi}$.

(b) Another way of stating the result from part (a) is that $\mathbf{w} = \mathbf{\Psi}^T\alpha$ for some vector α. Hence, instead of solving for $\mathbf{w}$, we can solve for α. Consider the vectorized form of the L_2 regularized linear regression cost function:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2N}||\mathbf{t} - \mathbf{\Psi}\mathbf{w}||^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

Substitute in $\mathbf{w} = \mathbf{\Psi}^T\alpha$, to write the cost function as a function of α. Determine the optimal value of α. Your answer should be an expression involving λ, $\mathbf{t}$, and the Gram matrix $\mathbf{K} = \mathbf{\Psi}\mathbf{\Psi}^T$. For simplicity, you may assume that $\mathbf{K}$ is positive definite. (The algorithm still works if $\mathbf{K}$ is merely PSD, it's just a bit more work to derive.)

Hint: the cost function $\mathcal{J}(\alpha)$ a quadratic function. Simplify the formula into the following form:

$$\frac{1}{2}\alpha^T\mathbf{A}\alpha + \mathbf{b}^T\alpha + c$$

for some positive definite matrix $\mathbf{A}$, vector $\mathbf{b}$, and constant c (which can be ignored). You may assume without proof that the minimum of such a quadratic function is given by $\alpha = -\mathbf{A}^{-1}\mathbf{b}$.

Solution:

First, we substitute $\mathbf{w} = \mathbf{\Psi}^T\alpha$ into $\mathcal{J}(\mathbf{w})$ in order to obtain $\mathcal{J}(\alpha)$. Thus, we have that

$$
\begin{aligned}
\mathcal{J}(\mathbf{w}) &= \frac{1}{2N}||\mathbf{t}-\mathbf{\Psi w}||^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \\
\mathcal{J}(\alpha) &= \frac{1}{2N}||\mathbf{t}-\mathbf{\Psi\Psi}^T\alpha||^2 + \frac{\lambda}{2}||\mathbf{\Psi}^T\alpha||^2 \\
&= \frac{1}{2N}(\mathbf{t}-\mathbf{\Psi\Psi}^T\alpha)^T(\mathbf{t}-\mathbf{\Psi\Psi}^T\alpha) + \frac{\lambda}{2}(\mathbf{\Psi}^T\alpha)^T(\mathbf{\Psi}^T\alpha) \\
&= \frac{1}{2N}(\mathbf{t}^T\mathbf{t}-\mathbf{t}^T\mathbf{\Psi\Psi}^T\alpha-\alpha^T\mathbf{\Psi\Psi}^T\mathbf{t}+\alpha^T\mathbf{\Psi\Psi}^T\mathbf{\Psi\Psi}^T\alpha) + \frac{\lambda}{2}(\mathbf{\Psi}^T\alpha)^T(\mathbf{\Psi}^T\alpha) \\
&= \frac{1}{2N}(\mathbf{t}^T\mathbf{t}-\mathbf{t}^T\mathbf{\Psi\Psi}^T\alpha-\alpha^T\mathbf{\Psi\Psi}^T\mathbf{t}+\alpha^T\mathbf{\Psi\Psi}^T\mathbf{\Psi\Psi}^T\alpha) + \frac{\lambda}{2}(\alpha^T\mathbf{\Psi\Psi}^T\alpha) \\
&= \frac{1}{2N}(\mathbf{t}^T\mathbf{t}-\mathbf{t}^T\mathbf{K}\alpha-\alpha^T\mathbf{K}\mathbf{t}+\alpha^T\mathbf{K}\mathbf{K}\alpha) + \frac{\lambda}{2}(\alpha^T\mathbf{K}\alpha) \quad \#By\ \mathbf{K}=\mathbf{\Psi\Psi}^T \\
&= \frac{1}{2N}(\mathbf{t}^T\mathbf{t}-\mathbf{t}^T\mathbf{K}\alpha-\mathbf{t}^T\mathbf{K}\alpha+\alpha^T\mathbf{K}\mathbf{K}\alpha) + \frac{\lambda}{2}(\alpha^T\mathbf{K}\alpha) \quad \#\mathbf{K}\ is\ a\ Gram\ matrix,\ square\ and\ symmetric \\
&= \frac{1}{2N}(\mathbf{t}^T\mathbf{t}-2\mathbf{t}^T\mathbf{K}\alpha+\alpha^T\mathbf{K}\mathbf{K}\alpha) + \frac{\lambda}{2}(\alpha^T\mathbf{K}\alpha) \\
&= \frac{1}{2N}\mathbf{t}^T\mathbf{t} - \frac{1}{N}\mathbf{t}^T\mathbf{K}\alpha + \frac{1}{2N}\alpha^T\mathbf{K}\mathbf{K}\alpha + \frac{\lambda}{2}(\alpha^T\mathbf{K}\alpha) \\
&= \frac{1}{2}(\frac{1}{N}\alpha^T\mathbf{K}\mathbf{K}\alpha+\lambda\alpha^T\mathbf{K}\alpha) - \frac{1}{N}\mathbf{t}^T\mathbf{K}\alpha + \frac{1}{2N}\mathbf{t}^T\mathbf{t} \\
&= \frac{1}{2}[\alpha^T(\frac{1}{N}\mathbf{K}\mathbf{K}+\lambda\mathbf{K})\alpha] - \frac{1}{N}\mathbf{t}^T\mathbf{K}\alpha + \frac{1}{2N}\mathbf{t}^T\mathbf{t}
\end{aligned}
$$

Then, let $\mathbf{A} = \frac{1}{N}\mathbf{K}\mathbf{K}+\lambda\mathbf{K}$ since $\mathbf{K}$ is positive definite and $\mathbf{b} = (-\frac{1}{N}\mathbf{t}^T\mathbf{K})^T = -\frac{1}{N}\mathbf{K}\mathbf{t}$ since $\mathbf{K}$ is a Gram matrix. Then, by hint, we are able to find the optimal α, which is

$$
\begin{aligned}
\alpha &= -\mathbf{A}^{-1}\mathbf{b} \\
&= -(\frac{1}{N}\mathbf{K}\mathbf{K}+\lambda\mathbf{K})^{-1}(-\frac{1}{N}\mathbf{K}\mathbf{t}) \\
&= (\frac{1}{N}\mathbf{K}\mathbf{K}+\lambda\mathbf{K})^{-1}(\frac{1}{N}\mathbf{K}\mathbf{t})
\end{aligned}
$$

Therefore, the optimal α is $\alpha = (\frac{1}{N}\mathbf{K}\mathbf{K}+\lambda\mathbf{K})^{-1}(\frac{1}{N}\mathbf{K}\mathbf{t})$.

2. **Compositional Kernels.** One of the most useful facts about kernels is that they can be composed using addition and multiplication. I.e., the sum of two kernels is a kernel, and the product of two kernels is a kernel. We'll show this in the case of kernels which represent dot products between finite feature vectors.

   (a) Suppose $k_1(x,x') = \psi_1(x)^T\psi_1(x')$ and $k_2(x,x') = \psi_2(x)^T\psi_2(x')$. Let k_S be the sum kernel $k_S(x,x') = k_1(x,x') + k_2(x,x')$. Find a feature map ψ_S such that $k_S(x,x') = \psi_S(x)^T\psi_S(x')$.

   Solution:

   We find the feature space for $k_S(x,x')$ that is

$$
\begin{aligned}
k_S(x, x') &= k_1(x, x') + k_2(x, x') \\
&= \psi_1(x)^T \psi_1(x') + \psi_2(x)^T \psi_2(x') \\
&= [\psi_1(x)^T \ \psi_2(x)^T] \begin{bmatrix} \psi_1(x') \\ \psi_2(x') \end{bmatrix} \\
&= \begin{bmatrix} \psi_1(x) \\ \psi_2(x) \end{bmatrix}^T \begin{bmatrix} \psi_1(x') \\ \psi_2(x') \end{bmatrix}
\end{aligned}
$$

Therefore, $\psi_S = \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix}$ such that $k_S(x, x') = \psi_S(x)^T \psi_S(x')$.

(b) Suppose $k_1(x, x') = \psi_1(x)^T \psi_1(x')$ and $k_2(x, x') = \psi_2(x)^T \psi_2(x')$. Let k_P be the product kernel $k_P(x, x') = k_1(x, x') k_2(x, x')$. Find a feature map ψ_P such that $k_P(x, x') = \psi_P(x)^T \psi_P(x')$.

Hint: For inspiration, consider the quadratic kernel from Lecture 20, Slide 11.

Solution:

We find the feature space for $k_P(x, x')$ that is

$$
\begin{aligned}
k_P(x, x') &= k_1(x, x') k_2(x, x') \\
&= \psi_1(x)^T \psi_1(x') \psi_2(x)^T \psi_2(x') \\
&= (\sum_{i=1}^{N_1} \psi_1(x)_i \psi_1(x')_i)(\sum_{j=1}^{N_2} \psi_2(x)_j \psi_2(x')_j) \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \psi_1(x)_i \psi_1(x')_i \psi_2(x)_j \psi_2(x')_j \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} [\psi_1(x)_i \psi_2(x)_j][\psi_1(x')_i \psi_2(x')_j] \\
&= \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \psi(x)_{ij} \psi(x')_{ij} \quad \#Let\ \psi(x)_{ij} = \psi_1(x)_i \psi_2(x)_j \\
&= \sum_{k=1}^{N_1 N_2} \phi(x)_k \phi(x')_k \quad \#Let\ \phi(x)_k = \psi(x)_{ij}\ where\ k = N_1(i-1) + j,\ then\ \phi(x) = \psi_1(x) \times \psi_2(x) \\
&= \phi(x)^T \phi(x')
\end{aligned}
$$

Therefore, the feature map $\psi_P = \phi = \psi_1 \times \psi_2$, which is a Cartesian product, such that $k_P(x, x') = \psi_P(x)^T \psi_P(x')$.