CMSC 25400: Machine Learning **Winter 2019**

Recitation 6: Kernel PCA, Ridge Regression

1 PCA

Recall in principal components analysis, we are interested in the following maximization problem. Given data: $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \in \mathbb{R}^d$ that has been centered (each index of the data vectors has mean 0), the first principal component is the unit vector $\boldsymbol{w}$ whose projection onto the data is maximized:

$$\boldsymbol{w_1} = \underset{||\boldsymbol{w}||=1}{\arg\max} \frac{1}{m} \sum_{i}^{m} (\boldsymbol{x_i}^\top \boldsymbol{w})^2$$

As we've seen in lecture, this is an eigenvalue problem in disguise. Let $\boldsymbol{X} = \begin{bmatrix} | & | & & | \\ \boldsymbol{x_1} & \boldsymbol{x_2} & \ldots & \boldsymbol{x_m} \\ | & | & & | \end{bmatrix}$ and rewrite the maximization problem in the form of an inner product of this data matrix.

$$\begin{aligned}
\boldsymbol{w_1} &= \underset{||\boldsymbol{w}||=1}{\arg\max} \frac{1}{m} ||\boldsymbol{X}^\top w||^2 \\
&= \underset{||\boldsymbol{w}||=1}{\arg\max} \frac{1}{m} \langle \boldsymbol{X}^\top \boldsymbol{w}, \boldsymbol{X}^\top \boldsymbol{w} \rangle \\
&= \underset{||\boldsymbol{w}||=1}{\arg\max} \frac{1}{m} \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}^\top \boldsymbol{w}
\end{aligned}$$

This is exactly the Rayleigh Quotient that we saw in homework 1. So the $\boldsymbol{w}$ that maximizes $\boldsymbol{w}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w}$ is the eigenvector with largest eigenvalue of $\boldsymbol{X}\boldsymbol{X}^\top$.

1.1 Kernel PCA

Now instead suppose our data comes from some set $\mathcal{X}$. We have some positive semidefinite kernel on this set $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, with the corresponding induced feature mapping $\phi : \mathcal{X} \to \mathbb{R}^d$. We're interested in the same sort of maximization above but in the feature space given by ϕ.

Let $\boldsymbol{x_1}, ..., \boldsymbol{x_m} \in \mathcal{X}$ and $\boldsymbol{\Phi} = \begin{bmatrix} | & | & & | \\ \phi(\boldsymbol{x_1}) & \phi(\boldsymbol{x_2}) & \ldots & \phi(\boldsymbol{x_m}) \\ | & | & & | \end{bmatrix}$.

$$\begin{aligned}
\boldsymbol{v_1} &= \underset{||\boldsymbol{v}||=1}{\arg\max} \sum_{i=1}^{m} \langle \phi(\boldsymbol{x_i}), \boldsymbol{v} \rangle^2 \\
&= \underset{||\boldsymbol{v}||=1}{\arg\max} ||\boldsymbol{\Phi}^\top \boldsymbol{v}||^2 \\
&= \underset{||\boldsymbol{v}||=1}{\arg\max} \boldsymbol{v}^\top \boldsymbol{\Phi} \boldsymbol{\Phi}^\top \boldsymbol{v}
\end{aligned}$$

So $\boldsymbol{v_1}$ is the top eigenvector of $\boldsymbol{\Phi\Phi}^\top$. We can say a bit more about the top principal component vector $\boldsymbol{v_1}$ if we consider the form it must take with respect to the feature mappings of the input data.

Lemma 1 *The $\boldsymbol{v_1}$ that maximizes this sum of square projections onto the $\phi(\boldsymbol{x_i})$'s will be a linear combination of the $\phi(\boldsymbol{x_i})$'s:*

$$\boldsymbol{v_1} = \sum_{i=1}^{m} \alpha_i \phi(\boldsymbol{x_i})$$

for some set of $\alpha_i \in \mathbb{R}$. This can be equivalently written as: $\boldsymbol{v_1} = \boldsymbol{\Phi\alpha}$, for $\boldsymbol{\alpha} \in \mathbb{R}^m$.

Proof: Suppose $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{\Phi\Phi}^\top = \sum_{i=1}^m \phi(\boldsymbol{x_i})\phi(\boldsymbol{x_i})^\top$:

$$\begin{aligned}
\boldsymbol{\Phi\Phi}^\top \boldsymbol{v} &= \sum_{i=1}^{n} \phi(\boldsymbol{x_i})\phi(\boldsymbol{x_i})^\top \boldsymbol{v} \\
\lambda \boldsymbol{v} &= \sum_{i=1}^{n} \phi(\boldsymbol{x_i}) \underbrace{\phi(\boldsymbol{x_i})^\top \boldsymbol{v}}_{\in \mathbb{R}} \\
\boldsymbol{v} &= \frac{1}{\lambda} \sum_{i=1}^{n} (\phi(\boldsymbol{x_i})^\top \boldsymbol{v}) \phi(\boldsymbol{x_i})
\end{aligned}$$

Let $\alpha_i = \frac{\phi(\boldsymbol{x_i})^\top \boldsymbol{v}}{\lambda}$ and we have our desired result. ∎

Now let's use this explicit form $\boldsymbol{v_1} = \boldsymbol{\Phi\alpha}$ in our computations and see what pops out:

$$\begin{aligned}
\boldsymbol{\Phi\Phi}^\top \boldsymbol{v_1} &= \boldsymbol{\Phi\Phi}^\top \boldsymbol{\Phi\alpha} \\
\lambda \boldsymbol{v_1} &= \boldsymbol{\Phi K \alpha}
\end{aligned}$$

where $\boldsymbol{K} = \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \in \mathbb{R}^{m \times m}$ is the Gram matrix of our input data: $\boldsymbol{K}_{i,j} = \phi(\boldsymbol{x_i})^\top \phi(\boldsymbol{x_j}) = k(\boldsymbol{x_i}, \boldsymbol{x_j})$.

Hit both sides by $\boldsymbol{\Phi}^\top$:

$$\begin{aligned}
\boldsymbol{\Phi}^\top \lambda \boldsymbol{v_1} &= \boldsymbol{\Phi}^\top \boldsymbol{\Phi K \alpha} \\
\lambda \boldsymbol{\Phi}^\top \boldsymbol{\Phi \alpha} &= \boldsymbol{\Phi}^\top \boldsymbol{\Phi K \alpha} \\
\lambda \boldsymbol{K \alpha} &= \boldsymbol{K K \alpha}
\end{aligned}$$

Eliminating one $\boldsymbol{K}$ term from both sides gives us: $\boldsymbol{K\alpha} = \lambda \boldsymbol{\alpha}$, which tells us that the coefficient vector $\boldsymbol{\alpha}$ of $\boldsymbol{v_1}$, is in fact the top eigenvector of $\boldsymbol{K}$. Note that $\boldsymbol{K}$ might not be full rank, but this will only be an issue for the zero-eigenvalued eigenvectors, which will not be a top principal component in the first place.

An equivalent way of getting to this result is by directly plugging in $\boldsymbol{v_1} = \boldsymbol{\Phi\alpha}$ into our maximization problem:

$$\begin{aligned}
\max_{||v||=1} \boldsymbol{v_1}^\top \boldsymbol{\Phi\Phi}^\top \boldsymbol{v_1} &= \max_{||v||=1} \boldsymbol{\alpha}^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi\Phi}^\top \boldsymbol{\Phi\alpha} \\
&= \max_{||v||=1} \boldsymbol{\alpha}^\top \boldsymbol{KK\alpha}
\end{aligned}$$

Thus, $\boldsymbol{\alpha}$ must be the top eigenvector of $\boldsymbol{K}^2$. The eigenvectors of $\boldsymbol{K}^2$ are the same as the eigenvectors for $\boldsymbol{K}$. So computing the principal component of our data just boils down to finding the top eigenvector of the Gram matrix $\boldsymbol{K}$.

2 Kernel Ridge Regression

Using the same notation as in the previous section, suppose we have some base set $\mathcal{X}$, a positive semidefinite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and the corresponding Reproducing Kernel Hilbert Space $\mathcal{H}_k$ induced by by our kernel k. Given $\boldsymbol{x_1}, ...\boldsymbol{x_m} \in \mathcal{X}$, with corresponding target values $y_1, ..., y_m \in \mathbb{R}$, we have the Gram matrix $\boldsymbol{K}$, with entries: $\boldsymbol{K}_{i,j} = k(\boldsymbol{x_i}, \boldsymbol{x_j})$. We'd like to learn some function $f \in \mathcal{H}_k$ that fits our data subject to an additional regularization term:

$$\hat{f} = \underset{f \in \mathcal{H}_k}{\arg\min} \underbrace{\Big[\sum_{i=1}^{m} (f(\boldsymbol{x_i}) - y_i)^2 + \lambda ||f||_{\mathcal{H}_k} \Big]}_{\mathcal{R}[f]}$$

The Representer theorem tells us that $\hat{f}$ must be of the form:

$$\hat{f}(\cdot) = \sum_{j=1}^{m} \alpha_j k(\cdot, \boldsymbol{x_j})$$

Plug this form of $\hat{f}$ into the minimization expression $\mathcal{R}[f]$. First let's consider the datafitting term of $\mathcal{R}[f]$:

$$\begin{aligned} \sum_{i=1}^{m} (\hat{f}(\boldsymbol{x_i}) - y_i)^2 &= \sum_{i=1}^{m} (\sum_{j=1}^{m} \alpha_j k(\boldsymbol{x_i}, \boldsymbol{x_j}) - y_i)^2 && (1) \\ &= ||\boldsymbol{K\alpha} - \boldsymbol{y}||^2 && (2) \end{aligned}$$

where $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$, $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}$. The inner sum: $\sum_{j=1}^{m} \alpha_j k(\boldsymbol{x_i}, \boldsymbol{x_j})$ is the same as taking the dot product of the ith row vector of $\boldsymbol{K}$ with $\boldsymbol{\alpha}$.

Now expanding the regularization term:

$$\begin{aligned} \lambda ||f||_{\mathcal{H}_k} &= \lambda \langle \sum_{i=1}^{m} \alpha_i k(\cdot, \boldsymbol{x_i}), \sum_{j=1}^{m} \alpha_j k(\cdot, \boldsymbol{x_j}) \rangle && (3) \\ &= \lambda \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(\boldsymbol{x_i}, \boldsymbol{x_j}) && (4) \\ &= \lambda \boldsymbol{\alpha}^\top \boldsymbol{K\alpha} && (5) \end{aligned}$$

where we use the definition of the inner product of $\mathcal{H}_k$: $\langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y)$ and the linearity of the inner product to go from eq(3) to eq(4). Plugging eq(2) and eq(5) back into $\mathcal{R}[f]$:

$$\begin{aligned} \mathcal{R}[f] &= ||\boldsymbol{K\alpha} - \boldsymbol{y}||^2 + \lambda \boldsymbol{\alpha}^\top \boldsymbol{K\alpha} \\ &= \boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{K\alpha} - 2\boldsymbol{\alpha}^\top \boldsymbol{K}^\top \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{y} + \lambda \boldsymbol{\alpha}^\top \boldsymbol{K\alpha} \end{aligned}$$

Taking the gradient of $\mathcal{R}[f]$ with respect to $\boldsymbol{\alpha}$ and setting it to $\mathbf{0}$:

$$\begin{aligned}
\nabla_{\boldsymbol{\alpha}}\mathcal{R}[f] &= 2\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{\alpha} - 2\boldsymbol{K}^\top\boldsymbol{y} + 2\lambda\boldsymbol{K}\boldsymbol{\alpha} \\
\mathbf{0} &= 2\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{\alpha} - 2\boldsymbol{K}^\top\boldsymbol{y} + 2\lambda\boldsymbol{K}\boldsymbol{\alpha} \\
\boldsymbol{K}^2\boldsymbol{\alpha} + \lambda\boldsymbol{K}\boldsymbol{\alpha} &= \boldsymbol{K}\boldsymbol{y} \\
(\boldsymbol{K}^2 + \lambda\boldsymbol{K})\boldsymbol{\alpha} &= \boldsymbol{K}\boldsymbol{y} \\
\boldsymbol{\alpha} &= (\boldsymbol{K}^2 + \lambda\boldsymbol{K})^{-1}\boldsymbol{K}\boldsymbol{y} \\
&= (\boldsymbol{K}(\boldsymbol{K} + \lambda\boldsymbol{I}))^{-1}\boldsymbol{K}\boldsymbol{y} \\
&= (\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\boldsymbol{K}^{-1}\boldsymbol{K}\boldsymbol{y} \\
&= (\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}
\end{aligned}$$

$\boldsymbol{K}$ is symmetric so we can replace $\boldsymbol{K}^\top$ with $\boldsymbol{K}$ in the equations above. Evaluating $\hat{f}$ on some new point $\boldsymbol{z} \in \mathcal{X}$ then boils down to computing the inner product between $\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$ with the vector of kernel evaluations: $\boldsymbol{k_z} = \begin{bmatrix} k(\boldsymbol{z}, \boldsymbol{x_1}) \\ \vdots \\ k(\boldsymbol{z}, \boldsymbol{x_m}) \end{bmatrix}$

$$\hat{f}(\boldsymbol{z}) = \sum_{i=1}^{m} \alpha_i k(\boldsymbol{z}, \boldsymbol{x_i}) = \boldsymbol{\alpha}^\top \boldsymbol{k_z}$$

Something to think about: how does the penalty term $\lambda||f||_{\mathcal{H}_k}$ affect the regularized risk minimization problem? How does our solution, $\hat{f}$, change as we increase or decrease the parameter λ?