Chapter 9

stevenjin8

April 14, 2021

Comments and Proofs

4.4 Kernel PCA

It took me a while to understand this section. The idea is to leverage the Mercer property of kernels to map the data to a larger (potentially infinite) dimensional feature space and to compute the principal components over said feature space. Given that, we first compute the Gram matrix:

$$\begin{aligned}\mathbf{K} &= \mathbf{\Phi}\mathbf{\Phi}^T \\ k_{i,j} &= \kappa(\mathbf{x}_i, \mathbf{x}_j).\end{aligned}$$

Using the eigenvalue/eigenvector trick presented earlier we find the formula for $\mathbf{V}_{kpca}$. Thus the kpca embedding of a data point $\mathbf{x}_*$ is $\phi(\mathbf{x}_*)\mathbf{\Phi}^T\mathbf{U}\mathbf{\Lambda}^{-\frac{1}{2}}$ (note that equation 14.40 is missing a transpose).

I still don't understand algorithm 14.2. Given some new data $\mathbf{X}_*$, the vectorized equation for $\tilde{\mathbf{K}}_*$ should be

$$\begin{aligned}\tilde{\mathbf{K}}_* &= (\mathbf{\Phi}_* - \frac{1}{N}\sum \phi_i)\mathbf{\Phi}^T\mathbf{U}_{:,1:z}\mathbf{\Lambda}_{:,1:z} \\ &= \left(\mathbf{K}_* - \mathbf{1}_{N_*}\overline{\mathbf{k}}^T - \overline{\mathbf{k}}_*\mathbf{1}_N^T + \overline{k}\mathbf{1}_{N_*}\mathbf{1}_N^T\right)\mathbf{U}_{:,1:z}\mathbf{\Lambda}_{:,1:z},\end{aligned}$$

where $\mathbf{K}_* = \mathbf{\Phi}_*\mathbf{\Phi}^T$ contains the pairwise kernel between the new data and the training data; $\overline{\mathbf{k}}$ is the row-wise mean for $\mathbf{K}$; $\overline{\mathbf{k}}_*$ is the row-wise mean of $\mathbf{K}_*$; and $\overline{k}$ is the mean of all values in $\mathbf{K}$.

Regardless, line 8 of the equation cannot be correct since both $\mathbf{O}_*$ and $\mathbf{K}_*$ are $N_* \times N$.

Something that I found really interesting is that we do not normalize the columns of $\mathbf{\Phi}$. It makes sense, however, the whole idea of KPCA is centered around the kernel function and dimensions in the feature space that have more extreme values are going to have a larger impact on the latent values.

Exercises

Exercise 1

a. The plane that separates $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_2)$ with the largest margin is perpendicular to $\phi(\mathbf{x}_2) - \phi(\mathbf{x}_1)$. So, $\mathbf{w} \parallel \phi(\mathbf{x}_2) - \phi(\mathbf{x}_1) = \langle 0, 2, 2 \rangle$.
b. The value of the margin is $\sqrt{2}$: half the distance between $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_2)$.
c.

$$\mathbf{w} = \left\langle 0, \frac{1}{2}\frac{1}{2} \right\rangle.$$

d. Plugging in our values, we have

$$\begin{aligned} -w_0 &> 1 \\ 2 + w_0 &> 1. \end{aligned}$$

Thus, $w_0 = -1$.
e.

$$f(x) = -1 + \frac{\sqrt{2}}{2}x + \frac{1}{2}x^2.$$

Exercise 2

The resulting decision boundary is guaranteed to separate the classes. At a high level, this is a result of the fact that we are regularizing $\|\mathbf{w}\|$ and not w_0.

By definition, there exists $\mathbf{w}$ and w_0 such that

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + w_0) > 0$$

for all i. However, we can scale $\mathbf{w}$ and w_0 by any $a > 0$ while preserving the above inequality:

$$\begin{aligned} y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + w_0) &> 0 \\ ay_i(\mathbf{w}^T\phi(\mathbf{x}_i) + w_0) &> 0 \\ y_i((a\mathbf{w})^T\phi(\mathbf{x}_i) + aw_0) &> 0. \end{aligned}$$

in other words we can scale $\mathbf{w}$ arbitrarily while having f perfectly classify the data. Thus, the regularization loss is meaningless.