Brief Article

The Author

October 9, 2013

1 6.1

Originally, we optimize $J(\boldsymbol{w}) = 1/2(\boldsymbol{\Phi w} - \boldsymbol{t})^T(\boldsymbol{\Phi w} - \mathbf{t})$. Now we set $\boldsymbol{w} = \boldsymbol{\Phi}^T \boldsymbol{a}$ (since the optimal $\boldsymbol{w}$ must take this form) and instead optimize $\boldsymbol{a}$. That is (6.7) in PRML pp. 293. The solution of (6.7) takes the form $K(K + \lambda I)\boldsymbol{a} = K\boldsymbol{t}$, and we can see the solution of $\boldsymbol{a}$ is $(K+\lambda I)^{-1}(t+N(K)\boldsymbol{\lambda})$. Every one of them is fine. We can split $\boldsymbol{a}$ into a part in the span of K and the other part in the null space of K, and we can check that by only preserving the span part, $J(\boldsymbol{a}$ is the same. So we have $\boldsymbol{a} = vect\Phi\boldsymbol{u}$, and see the solution manual. Actually, I think it's useless to talk about the rank of matrix $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$. The point is, if we have a solution to $\boldsymbol{alpha}$, we can use it to get a solution of $\boldsymbol{w} = \boldsymbol{\Phi}^T\boldsymbol{alpha}$.

2 6.2

Certain we can see that $\boldsymbol{w} = \sum_n a_n t_n \phi(\boldsymbol{x}_n)$, where a_i is the number of addition of this signed (normalized) sample. The learning rule is, whenever we classify $\phi(\boldsymbol{x}_n)$ wrongly, we make a_n bigger by one. In prediciton, we use $\boldsymbol{w} = \boldsymbol{\Phi}^T\boldsymbol{\alpha}$ to get $\boldsymbol{w}^T\phi(\boldsymbol{x}) = \boldsymbol{\alpha}^T\boldsymbol{\Phi}\phi(\boldsymbol{x})$. Clearly the feature vector enters only in the form of the kernel function.