Tutorial 5 - Solutions

Sunday 21st February, 2016

Problem 1. Consider the matrix V whose columns are the vectors $\phi(x_1)$,$\phi(x_2)\ldots,\phi(x_n)$. Then, one can see that $\mathcal{K}=V^TV$. Now, for any $y\in\mathbb{R}^n$, $y^T\mathcal{K}y=y^TV^TVy=\|Vy\|^2\geq 0$ and hence, every gram matrix is positive semi-definite. (Solution also on page 6 of `https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-13-annotated.pdf`)

Problem 2.

1. $\cos(x_1-x_2)=\cos x_1\cos x_2+\sin x_1\sin x_1$. If one defines $\phi(x)$ as $[\cos x\ \ \sin x]^T$, then $\cos(x_1-x_2)=\phi^T(x_1)\phi(x_2)$. Use the property proved in Problem 1.

2. Since K_1 and K_2 are valid kernels, for any $n\times n$ kernel matrices $\mathcal{K}_1$ and $\mathcal{K}_2$ defined on K_1 and K_2 and any $y\in R^n$, we will have $y^T\mathcal{K}_1y\geq 0$ and $y^T\mathcal{K}_2y\geq 0$. Adding them, we get $y^T\mathcal{K}_1y+y^T\mathcal{K}_2y=y^T(\mathcal{K}_1+\mathcal{K}_2)y=y^T\mathcal{K}y\geq 0$. Since, our choice of y and n is arbitrary, all the kernel matrices $\mathcal{K}$ defined on K are positive-definite and hence, K is also positive semi-definite.

3. Clearly, $K(x_1,x_2)=(<x_1,x_2>+c)^d$ is a polynomial with positive coefficients in $<x_1,x_2>$,i.e. a sum of monomials with positive coefficients in $<x_1,x_2>$. If we prove that each monomial in $<x_1,x_2>$ induces a positive-semi definite matrix then using the result of previous sub-problem we are done. Consider $K'(x_1,x_2)=<x_1,x_2>^m$ where m is some constant. Let $\phi(x)$ be a vector whose entries are of the form $x(1)^{i_1}x(2)^{i_2}\ldots x(n)^{i_n}$ such that $\sum_{j=1}^n i_j=m$ and $i_j\geq 0$ for all $j\in\{1,2,...,,n\}$. Then, $K'(x_1,x_2)=\phi^T(x_1)\phi(x_2)$. Using property proved in Problem 1, K' is a positive-semi definite kernel and hence, K is positive-semi definite too.

4. $e^{<x_1,x_2>}=\sum_{i=1}^{\infty}\frac{<x_1,x_2>^i}{i!}$. Clearly, each term in the summation is of the form $(<x_1,x_2>+c)^d$ times some positive coefficient and hence, induces a positive semi-definite matrix (using previous sub-problem).Now, use result of sub-problem 2 to see that summation of these terms is a positive-semi definite kernel as well.

Problem 3. Please take a look at the following link. Equation 6 is exactly the same as discussed in class. And in Section-3 (eqn 8-15) the closed from expression for SMO is explained.

`http://link.springer.com/article/10.1023%2FA%3A1012474916001`

Problem 4.

- We have already stated the equivalence of Lasso formulations in (3) and (1). For this problem, we will go with the formulation in (3)

$$\mathbf{w}^*=\underset{\mathbf{w}}{\mathbf{argmin}}\|\phi\mathbf{w}-\mathbf{y}\|^{\mathbf{2}}+\lambda\|\mathbf{w}\|_{\mathbf{1}}\tag{1}$$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\mathbf{argmin}} \|\phi\mathbf{w} - \mathbf{y}\|^{\mathbf{2}} \;\; \mathbf{s.t.} \; \|\mathbf{w}\|_{\mathbf{1}} \leq \eta, \tag{2}$$

where

$$\|\mathbf{w}\|_1 = \Big(\sum_{i=1}^{n} |w_i|\Big) \tag{3}$$

- Since $\|\mathbf{w}\|_1$ is not differentiable, one can express (3) as a set of constraints

$$\sum_{i=1}^{n} \xi_i \leq \eta, \;\; w_i \leq \xi_i, \;\; -w_i \leq \xi_i$$

- The resulting problem is a linearly constrained Quadratic optimization problem (LCQP):

$$\mathbf{w}^* = \underset{\mathbf{w}}{\mathbf{argmin}} \|\phi\mathbf{w} - \mathbf{y}\|^{\mathbf{2}} \;\; \mathbf{s.t.} \; \sum_{\mathbf{i=1}}^{\mathbf{n}} \xi_{\mathbf{i}} \leq \eta, \;\; \mathbf{w_i} \leq \xi_{\mathbf{i}}, \;\; -\mathbf{w_i} \leq \xi_{\mathbf{i}} \tag{4}$$

- Lagrangian is

$$\|\phi\mathbf{w} - \mathbf{y}\|^2 \;\; + \beta(\sum_{i=1}^{n} \xi_i - \eta) + \sum_{i=1}^{n} \big(\theta_i(\; w_i - \xi_i) + \lambda_i(-w_i - \xi_i)\big)$$

- KKT conditions: Setting gradient wrt $\mathbf{w}$ to $\mathbf{0}$:

$$2(\phi^T\phi)\mathbf{w} - \mathbf{2}\phi^{\mathbf{T}}\mathbf{y} + (\theta - \lambda) = \mathbf{0}$$

  Setting gradient wrt ξ_i to 0:

$$\beta - \theta_i - \lambda_i = 0$$

- Substituting for $\mathbf{w}$ and θ_1 and λ_i from the necessary and sufficient conditions above in the Lagrangian, we get the Langrage dual optimization problem

$$\underset{\theta_i,\lambda_i}{\mathbf{argmin}} \; \big(\phi^T y + \tfrac{1}{2}(\lambda - \theta)\big)^T (\phi^T\phi)^{-1}\phi^T\phi(\phi^T\phi)^{-1} \big(\phi^T y + \tfrac{1}{2}(\lambda - \theta)\big) - 2\mathbf{y}^{\mathbf{T}}\phi(\phi^{\mathbf{T}}\phi)^{-\mathbf{1}} \big(\phi^{\mathbf{T}}\mathbf{y} + \tfrac{\mathbf{1}}{\mathbf{2}}(\lambda - \theta)\big)$$
$$+\mathbf{y}^{\mathbf{T}}\mathbf{y} + \beta\eta + (\theta - \lambda)^{\mathbf{T}} (\phi^{\mathbf{T}}\phi)^{-\mathbf{1}} \big(\phi^{\mathbf{T}}\mathbf{y} + \tfrac{\mathbf{1}}{\mathbf{2}}(\lambda - \theta)\big)$$
$$= \underset{\theta_i,\lambda_i}{\mathbf{argmin}} \; - \mathbf{y}^{\mathbf{T}}\phi(\phi^{\mathbf{T}}\phi)^{-\mathbf{1}}\phi^{\mathbf{T}}\mathbf{y} + \mathbf{y}^{\mathbf{T}}\mathbf{y} + \beta\eta + \tfrac{\mathbf{1}}{\mathbf{2}}\,(\theta - \lambda)^{\mathbf{T}} (\phi^{\mathbf{T}}\phi)^{-\mathbf{1}}(\lambda - \theta)$$

- Note that $\phi^T\phi$ is not a kernel (gram) matrix where $\phi\phi^T$ is (see page 11 of `https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-12-annotated.pdf`)

- Even if using the identities on pages 12-14 of `https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-12-annotated.pdf` that was used for deriving the "kernelized dual" of ridge regression, we were to kernelize the first term above, the last term will remain $\frac{1}{2}\,(\theta - \lambda)^T (\phi^T\phi)^{-1}(\lambda - \theta)$ which cannot be kernelized.

- Thus, Lasso does not have purely kernelized dual