Feature Selection with Kernel Class Separability (Appendix Only)

Lei Wang

Manuscript received ; revised .

L. Wang is with Research School of Information Sciences and Engineering, The Australian National University, Australia.

APPENDIX I

THE RELATIONSHIP TO THE RADIUS-MARGIN BOUND

Recall that the optimal $\|\mathbf{w}\|^2$ can be computed as

$$\begin{aligned}&\tfrac{1}{2}\|\mathbf{w}\|^2 = \max_{\boldsymbol{\alpha}\in\mathbb{R}^n}\left[\sum_{i=1}^n \alpha_i - \tfrac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)\right]\\ &subject\ to: \ \sum_{i=1}^n \alpha_i y_i = 0;\ \alpha_i \geq 0\end{aligned} \tag{1}$$

Let us define

$$\widetilde{\alpha}_i = \begin{cases} 1/n_1 & \text{when } \mathbf{x}_i \in \mathcal{D}_1 \\ 1/n_2 & \text{when } \mathbf{x}_i \in \mathcal{D}_2 \end{cases}. \tag{2}$$

Please note that $\widetilde{\alpha}_i$ is within the feasible region of the maximization problem in Eq.(1) because it satisfies $\sum_{i=1}^n \widetilde{\alpha}_i y_i = 0$ and $\widetilde{\alpha}_i \geq 0$. Taking the $\widetilde{\alpha}_i$ as the (sub-optimal) solution of (1) leads to

$$\sum_{i=1}^n \widetilde{\alpha}_i - \tfrac{1}{2}\sum_{i,j=1}^n \widetilde{\alpha}_i\widetilde{\alpha}_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) = 2 - \tfrac{1}{2}\sum_{i,j=1}^n \widetilde{\alpha}_i\widetilde{\alpha}_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) \leq \tfrac{1}{2}\|\mathbf{w}\|^2 \tag{3}$$

The inequality is because $\frac{1}{2}\|\mathbf{w}\|^2$ is defined as the maximum value of the object function in Eq.(1).

Furthermore, it can be shown that

$$\begin{aligned} &\sum_{i,j=1}^n \widetilde{\alpha}_i\widetilde{\alpha}_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)\\ =\ & \left(\sum_{\mathbf{x}_i\in\mathcal{D}_1,\mathbf{x}_j\in\mathcal{D}_1} + 2\sum_{\mathbf{x}_i\in\mathcal{D}_1,\mathbf{x}_j\in\mathcal{D}_2} + \sum_{\mathbf{x}_i\in\mathcal{D}_2,\mathbf{x}_j\in\mathcal{D}_2}\right)\widetilde{\alpha}_i\widetilde{\alpha}_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)\\ =\ & \left(\frac{1}{n_1^2}\sum_{\mathcal{D}_1,\mathcal{D}_1} - 2\frac{1}{n_1 n_2}\sum_{\mathcal{D}_1,\mathcal{D}_2} + \frac{1}{n_2^2}\sum_{\mathcal{D}_2,\mathcal{D}_2}\right) k(\mathbf{x}_i,\mathbf{x}_j)\\ =\ & \left[\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_1,\mathcal{D}_1})}{n_1^2} - 2\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_1,\mathcal{D}_2})}{n_1 n_2} + \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_2,\mathcal{D}_2})}{n_2^2}\right]\\ =\ & \left(\frac{n_1+n_2}{n_1 n_2}\right)\left[\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_1,\mathcal{D}_1})}{n_1} + \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_2,\mathcal{D}_2})}{n_2} - \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n_1+n_2}\right]\\ =\ & \left(\frac{n_1+n_2}{n_1 n_2}\right)\mathrm{tr}(\mathbf{S}_B^{\phi}) \end{aligned} \tag{4}$$

Combing the results in Eq.(3) and (4) and noting that $\gamma = 1/\|\mathbf{w}\|$, it can be obtained that

$$\gamma^2 \leq \frac{1}{4 - \left(\frac{n_1+n_2}{n_1 n_2}\right) \mathrm{tr}(\mathbf{S}_B^\phi)} \tag{5}$$

Please note that $4 - \left(\frac{n_1+n_2}{n_1 n_2}\right) \mathrm{tr}(\mathbf{S}_B^\phi)$ is always non-negative for a kernel which maps the input data onto a unit hypersphere, including all stationary kernels and the normalized kernels[1]. The result in (5) indicates that (i) γ^2 is upper bounded by a function of $\mathrm{tr}(\mathbf{S}_B^\phi)$ and (ii) to allow γ^2 to be maximized, the $\mathrm{tr}(\mathbf{S}_B^\phi)$ needs to be maximized too.

Similarly, the optimal R^2 is obtained by solving

$$\begin{aligned} &R^2 = \max_{\boldsymbol{\beta}\in\mathbb{R}^n} \left[\sum_{i=1}^n \beta_i k_{ii} - \sum_{i,j=1}^n \beta_i\beta_j k_{ij}\right] \\ &subject\ to: \ \ \sum_{i=1}^n \beta_i = 1;\ \ \beta_i \geq 0 \end{aligned} \tag{6}$$

Similarly, let us define $\widetilde{\beta}_i = 1/(n_1+n_2)$ and $\widetilde{\beta}_i$ is also within the feasible region of the maximization problem in Eq.(6). Taking $\widetilde{\beta}_i$ as the (sub-optimal) solution of Eq.(6) leads to

$$\sum_{i=1}^n \widetilde{\beta}_i k(\mathbf{x}_i,\mathbf{x}_i) - \sum_{i,j=1}^n \widetilde{\beta}_i\widetilde{\beta}_j k(\mathbf{x}_i,\mathbf{x}_j) = \frac{\mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{(n_1+n_2)} - \sum_{i,j=1}^n \widetilde{\beta}_i\widetilde{\beta}_j k(\mathbf{x}_i,\mathbf{x}_j) \leq R^2 \tag{7}$$

The inequality is due to that R^2 is defined as the maximum value of the object function in Eq.(6).

[1]It can be proven that $\mathrm{tr}(\mathbf{S}_B^\phi) = \left(\frac{n_1 n_2}{n_1+n_2}\right)\|\mathbf{m}_1^\phi - \mathbf{m}_2^\phi\|^2$, where $\mathbf{m}_i^\phi$ is the mean vector of class i in the kernel space. Thus, $\left[4 - \left(\frac{n_1+n_2}{n_1 n_2}\right)\mathrm{tr}(\mathbf{S}_B^\phi)\right]$ can be rewritten as $4 - \|\mathbf{m}_1^\phi - \mathbf{m}_2^\phi\|^2$. Since $\mathbf{m}_i^\phi$ is a convex combination of all the samples, $\phi(\mathbf{x})$, in class i, it must lie inside the unit hypersphere when a stationary or normalized kernel is used. Hence, $\|\mathbf{m}_1^\phi - \mathbf{m}_2^\phi\|$ must be less than 2, the length of the diameter. For a Gaussian RBF kernel, $\|\mathbf{m}_1^\phi - \mathbf{m}_2^\phi\|$ is even less than $\sqrt{2}$ because $\langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$ is always positive.

Moreover, it can be shown that

$$\begin{aligned} \frac{\mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{(n_1+n_2)} - \sum_{i,j=1}^{n} \widetilde{\beta}_i \widetilde{\beta}_j k(\mathbf{x}_i, \mathbf{x}_j) &= \frac{\mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{(n_1+n_2)} - \frac{1}{(n_1+n_2)^2} \sum_{\mathbf{x}_i \in \mathcal{D}, \mathbf{x}_j \in \mathcal{D}} k(\mathbf{x}_i, \mathbf{x}_j) \\ &= \frac{1}{(n_1+n_2)} \left[\mathrm{tr}(\mathbf{K}_{\mathcal{D},\mathcal{D}}) - \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n_1+n_2} \right] \\ &= \frac{1}{(n_1+n_2)} \mathrm{tr}(\mathbf{S}_T^{\phi}) \end{aligned} \tag{8}$$

Combining the results in Eq.(7) and (8), it can be obtained that

$$R^2 \geq \frac{1}{(n_1+n_2)} \mathrm{tr}(\mathbf{S}_T^{\phi}) \tag{9}$$

Hence, $\frac{1}{(n_1+n_2)}\mathrm{tr}(\mathbf{S}_T^{\phi})$ is a lower bound of R^2 and, to allow R^2 to be minimized, the $\mathrm{tr}(\mathbf{S}_T^{\phi})$ needs to be minimized.

APPENDIX II
THE RELATIONSHIP TO THE KERNEL ALIGNMENT

$$\begin{aligned} & \mathrm{tr}(\mathbf{S}_B^{\phi}) \\ = \;& \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_1,\mathcal{D}_1})}{n_1} + \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_2,\mathcal{D}_2})}{n_2} - \frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n_1+n_2} \\ = \;& \frac{S_{11}}{n_1} + \frac{S_{22}}{n_2} - \frac{S_{11}+S_{22}+2S_{12}}{n_1+n_2} \\ = \;& (n_1+n_2)^{-1} \left[\frac{n_2}{n_1} S_{11} + \frac{n_1}{n_2} S_{22} - 2S_{12} \right] \\ = \;& (n_1+n_2)^{-1} \left[S_{11} + S_{22} - 2S_{12} \right] \quad (\text{when } n_1 = n_2) \\ = \;& (n_1+n_2)^{-1} \langle \mathbf{K}, \mathbf{y}\mathbf{y}^{\top} \rangle \end{aligned} \tag{10}$$

For a Gaussian kernel (and a part of normalized kernels), there is $k(\mathbf{x}_i, \mathbf{x}_j) \in (0, 1]$ and thus $k^2(\mathbf{x}_i, \mathbf{x}_j) \leq k(\mathbf{x}_i, \mathbf{x}_j)$. Hence, it can be obtained that

$$\langle \mathbf{K}, \mathbf{K} \rangle = \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}} k^2(\mathbf{x}_i, \mathbf{x}_j) \leq \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}} k(\mathbf{x}_i, \mathbf{x}_j) \tag{11}$$

Recall that $\mathbf{m}^\phi$ denote the mean of all training samples in the kernel space. It can be shown that

$$\begin{aligned}\sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{D}} k(\mathbf{x}_i,\mathbf{x}_j) &= \sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{D}} \left[1-\tfrac{1}{2}\|\phi(\mathbf{x}_i)-\phi(\mathbf{x}_j)\|^2\right] \\ &= (n_1+n_2)^2 - \tfrac{1}{2}\sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{D}} \|\phi(\mathbf{x}_i)-\phi(\mathbf{x}_j)\|^2 \\ &= (n_1+n_2)^2 - \tfrac{1}{2}\sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{D}} \|(\phi(\mathbf{x}_i)-\mathbf{m}^\phi)-(\phi(\mathbf{x}_j)-\mathbf{m}^\phi)\|^2 \\ &(\because \sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{D}}(\phi(\mathbf{x}_i)-\mathbf{m}^\phi)^\top(\phi(\mathbf{x}_j)-\mathbf{m}^\phi)=0 \ ) \\ &= (n_1+n_2)^2 - (n_1+n_2)\sum_{\mathbf{x}_i\in\mathcal{D}} \|\phi(\mathbf{x}_i)-\mathbf{m}^\phi\|^2 \\ &= (n_1+n_2)^2 - (n_1+n_2)\mathrm{tr}(\mathbf{S}_T^\phi)\end{aligned} \tag{12}$$

Therefore,

$$\langle\mathbf{K},\mathbf{K}\rangle \le (n_1+n_2)\left[(n_1+n_2)-\mathrm{tr}(\mathbf{S}_T^\phi)\right] \tag{13}$$

APPENDIX III

THE RELATIONSHIP TO THE KFDA

According to the definitions of $\mathbf{S}_B^\phi$ and $\mathbf{S}_T^\phi$, it is known that both of them are PSD (Positive Semi-Definite). Following the property of Rayleigh Quotient, it can be obtained that

$$\begin{aligned} 0 \le \frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}}{\mathbf{w}^\top\mathbf{w}} \le \lambda_{max}(\mathbf{S}_B^\phi) \\ 0 \le \frac{\mathbf{w}^\top\mathbf{S}_T^\phi\mathbf{w}}{\mathbf{w}^\top\mathbf{w}} \le \lambda_{max}(\mathbf{S}_T^\phi)\end{aligned} \tag{14}$$

where $\lambda_{max}(\mathbf{S}_B^\phi)$ and $\lambda_{max}(\mathbf{S}_T^\phi)$ denote the maximal eigenvalue of $\mathbf{S}_B^\phi$ and $\mathbf{S}_T^\phi$, respectively. Thus, the objective function of KFDA can be expressed as

$$\begin{aligned}\mathcal{J}(\mathbf{w}) &= \frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}}{\mathbf{w}^\top\mathbf{S}_T^\phi\mathbf{w}} = \frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}/\mathbf{w}^\top\mathbf{w}}{\mathbf{w}^\top\mathbf{S}_T^\phi\mathbf{w}/\mathbf{w}^\top\mathbf{w}} \\ &\ge \frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}/\mathbf{w}^\top\mathbf{w}}{\lambda_{max}(\mathbf{S}_T^\phi)}\end{aligned} \tag{15}$$

Hence,

$$\max_{\mathbf{w}\in\mathcal{K}}\left[\mathcal{J}(\mathbf{w})\right] \geq \max_{\mathbf{w}\in\mathcal{K}}\left(\frac{\mathbf{w}^\top\mathbf{S}_B^\phi\mathbf{w}/\mathbf{w}^\top\mathbf{w}}{\lambda_{max}(\mathbf{S}_T^\phi)}\right)$$

$$= \frac{\lambda_{max}(\mathbf{S}_B^\phi)}{\lambda_{max}(\mathbf{S}_T^\phi)} \geq \frac{\mathrm{tr}(\mathbf{S}_B^\phi)}{\mathrm{tr}(\mathbf{S}_T^\phi)} \quad (16)$$

The last inequality is based on the following two facts: (1) In a binary classification, $\mathrm{rank}(\mathbf{S}_B^\phi) = 1$ and $\mathbf{S}_B^\phi$ has one and only one non-zero eigenvalue. Thus, it can be obtained that $\lambda_{max}(\mathbf{S}_B^\phi) = \mathrm{tr}(\mathbf{S}_B^\phi)$; (2) It is known that $\sum_{i=1}^{\dim(\mathcal{K})}\lambda_i(\mathbf{S}_T^\phi) = \mathrm{tr}(\mathbf{S}_T^\phi)$ and that $\lambda_i(\mathbf{S}_T^\phi) \geq 0$ since $\mathbf{S}_T^\phi$ is PSD. Thus, it can be shown that $0 \leq \lambda_{max}(\mathbf{S}_T^\phi) \leq \mathrm{tr}(\mathbf{S}_T^\phi)$.

APPENDIX IV

THE CONVEXITY ANALYSIS OF $\mathcal{J}_{reg}^\phi(\boldsymbol{\eta})$

Correction: Equation (12) in the main text of this paper should be corrected as

$$\mathcal{J}_{reg}^\phi(\boldsymbol{\eta}) = (1-\lambda)\left(-\mathcal{J}^\phi(\boldsymbol{\eta})\right) + \lambda\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2 \quad (17)$$

and $\mathcal{J}_{reg}^\phi(\boldsymbol{\eta})$ is to be *minimized*. The following analysis is revised accordingly based on the corrected equation (12).

Since $f(x) = \exp(-x)$ is convex, the kernel, $k(\mathbf{x},\mathbf{y}) = \exp\left[-\sum_{i=1}^d \eta_i(x_i-y_i)^2\right]$, will be a convex function for η_i. Instantly, it can be obtained that $\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_i,\mathcal{D}_j}) = \sum_{\mathbf{x}_p\in\mathcal{D}_i}\sum_{\mathbf{y}_q\in\mathcal{D}_j} k(\mathbf{x}_p,\mathbf{y}_q)$ is also convex because a nonnegative weighted sum of convex functions is still convex. In addition, please note that $f(\boldsymbol{\eta}) = \|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2$ is also convex and that $0 \leq \lambda < 1$. Thus, $\mathcal{J}_{reg}^\phi(\boldsymbol{\eta})$ can be written a difference of two convex functions as follows.

$$\begin{aligned}
\mathcal{J}^{\phi}_{reg}(\boldsymbol{\eta}) &= (1-\lambda)\left(-\mathcal{J}^{\phi}(\boldsymbol{\eta})\right) + \lambda\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2 \\
&= (1-\lambda)\left(-\mathrm{tr}(\mathbf{S}^{\phi}_{B})\right) + \lambda\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2 \\
&= (1-\lambda)\left(\tfrac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n} - \textstyle\sum_{i=1}^{c}\tfrac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_i,\mathcal{D}_i})}{n_i}\right) + \lambda\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2 \\
&= \left[(1-\lambda)\tfrac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n} + \lambda\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2\right] - \left[(1-\lambda)\textstyle\sum_{i=1}^{c}\tfrac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_i,\mathcal{D}_i})}{n_i}\right] \\
&\triangleq g(\boldsymbol{\eta}) - h(\boldsymbol{\eta})
\end{aligned} \tag{18}$$

where $g(\boldsymbol{\eta}) \triangleq \left[(1-\lambda)\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D},\mathcal{D}})}{n} + \lambda\|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|^2\right]$ and $h(\boldsymbol{\eta}) \triangleq \left[(1-\lambda)\sum_{i=1}^{c}\frac{\mathrm{Sum}(\mathbf{K}_{\mathcal{D}_i,\mathcal{D}_i})}{n_i}\right]$. Both are convex for $\boldsymbol{\eta}$. This also indicates that $\mathcal{J}^{\phi}_{reg}(\boldsymbol{\eta})$ is not a convex function for $\boldsymbol{\eta}$.