BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–2

Supplement to: Kernel-based data fusion for gene prioritization

Tijl De Bie [a,b], Léon-Charles Tranchevent [c], Liesbeth M. M. van Oeffelen [c], Yves Moreau [c]

[a]Dept. of Engineering Mathematics, University of Bristol, University Walk, BS8 1TR, Bristol, UK
[b]OKP Research Group, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium
[c]ESAT-SCD, Kathlieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

PROOF OF THEOREM 1

Generalizing the problem

It is convenient to consider here a slightly more general algorithm. In particular, we consider the optimization problem

$$\max_{\mathbf{K}} \quad \max_{M,\mathbf{w},\boldsymbol{\xi}} \quad p(M,\boldsymbol{\xi}) = M - \frac{1}{n\nu}\mathbf{1}'\boldsymbol{\xi}, \tag{1}$$

$$\text{s.t.} \qquad \mathbf{w}'\mathbf{w} \leq 1, \;\; \mathbf{x}_i'\mathbf{w} \geq M - \xi_i \;(\forall i), \;\; \xi_i \geq 0 \;(\forall i), \;\; \mathbf{K} \in \left\{ \sum_j \mu_j(\mathbf{K}_j/\beta_j) : \boldsymbol{\mu}'\mathbf{1} = 1, \boldsymbol{\mu} \geq \mathbf{0} \right\}.$$

The difference with the problem introduced in the paper is that slack variables ξ_i are used here, which allow small mistakes for individual data objects. These mistakes are penalized stronger for small values of ν, and for $\nu \to 0$ the more simple optimization problem explained in the main part of the paper is recovered. Using duality theory, this problem can be shown to be equivalent to

$$\min_{t,\boldsymbol{\alpha}} t \qquad \text{s.t.} \qquad \frac{1}{n\nu} \geq \alpha_i \geq 0 \;(\forall i), \;\; \mathbf{1}'\boldsymbol{\alpha} = 1, \;\; t \geq \boldsymbol{\alpha}'(\mathbf{K}_j/\beta_j)\boldsymbol{\alpha} \;(\forall j).$$

For any value of the margin M, optimization problem (1) minimizes $\mathbf{1}'\boldsymbol{\xi} = \sum_{i=1}^n \max(M - f(x), 0) = \gamma \hat{E}_X\left(g_{M,\gamma}(\mathbf{x}'\mathbf{w})\right) = \gamma \hat{E}_X\left(g_{M,\gamma}(f(x))\right)$ for f belonging to the function class $\mathcal{F}_\mathcal{K}$ defined as:

$$\mathcal{F}_\mathcal{K} = \left\{ f : x \to \sum_{i=1}^n \alpha_i k(x, x_i) / \sqrt{\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}} \text{ and } k \in \mathcal{K} \right\}.$$

A more general Theorem

We prove a more general Theorem, from which Theorem 1 follows immediately. First we give a few definitions. Given values M and γ, define the function $\phi_{M,\gamma}$ as:

$$\phi_{M,\gamma}(a) = \min\left(\max\left(\tfrac{M-a}{\gamma}, 0\right), 1\right) = \begin{cases} 1 & a \leq M - \gamma, \\ \frac{M-a}{\gamma} & \text{if} \quad M - \gamma < a \leq M, \\ 0 & M \leq a. \end{cases}$$

Furthermore, define $g_{M,\gamma}(a) = \max\left(\frac{M-a}{\gamma}, 0\right)$. Then, with I the indicator function:

$$g_{M,\gamma}(a) \geq \phi_{M,\gamma}(a) \geq I(a \leq M - \gamma).$$

THEOREM 2. *Given a set X of n objects (genes) x_i sampled* iid *from an unknown distribution $\mathcal{D}$. Let $\lambda(\mathbf{K}_j)$ denote the largest eigenvalue of $\mathbf{K}_j$. Then, for any $M, \gamma \in \Re_+$ and for any $\delta \in (0,1)$, with probability of at least $1 - \delta$ the following holds for all functions $f \in \mathcal{F}$:*

$$P_\mathcal{D}\left(f(x) \leq M - \gamma\right) \leq \hat{E}_X\left(\phi_{M,\gamma}\left(f(x)\right)\right) + \frac{4}{n\gamma}\sqrt{\min\left(n \max_j \frac{\lambda(\mathbf{K}_j)}{\beta_j}, \sum_{j=1}^m \frac{\text{trace}\left(\mathbf{K}_j\right)}{\beta_j}\right)} + \sqrt{\frac{2}{n}\ln\frac{2}{\delta}}.$$

© Oxford University Press 2007.

Proof of Theorem 2

The proof has the same structure as the Rademacher complexity proofs in Bartlett and Mendelson (2002); Shawe-Taylor and Cristianini (2004); Lanckriet *et al.* (2004). For any M and γ:

$$\begin{aligned} P_{\mathcal{D}}\left(f(x) \leq M-\gamma\right) &= E_{\mathcal{D}}\left(I\left(f(x) \leq M-\gamma\right)\right) \leq E_{\mathcal{D}}\left(\phi_{M,\gamma}(f(x))\right) \\ &\leq \hat{E}_X\left(\phi_{M,\gamma}(f(x))\right) + \sup_{f\in\mathcal{F}}\left(E_{\mathcal{D}}\left(\phi_{M,\gamma}(f(x))\right) - \hat{E}_{\mathbf{X}}\left(\phi_{M,\gamma}(f(x))\right)\right) \end{aligned}$$

We now make use of the fact that $\hat{E}_X\left(\phi_{M,\gamma}(f(x))\right)$ is close to its expectation, as shown by McDiarmid's inequality for functions with bounded differences. With probability $1-\frac{\delta}{2}$ over X:

$$\begin{aligned} &\sup_{f\in\mathcal{F}}\left(E_{\mathcal{D}}\left(\phi_{M,\gamma}(f(x))\right) - \hat{E}_X\left(\phi_{M,\gamma}(f(x))\right)\right) \\ \leq\ & E_{X\sim\mathcal{D}^n}\sup_{f\in\mathcal{F}}\left(E_{\mathcal{D}}\left(\phi_{M,\gamma}(f(x))\right) - \hat{E}_X\left(\phi_{M,\gamma}(f(x))\right)\right) + \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}. \end{aligned}$$

Now consider a new sample Z of n data objects, sampled *iid* from $\mathcal{D}$. From linearity, $E_{\mathcal{D}}\left(\phi_{M,\gamma}(f(x))\right) = E_{Z\sim\mathcal{D}^n}\left(\hat{E}_Z\left(\phi_{M,\gamma}(f(z))\right)\right)$. Furthermore, the supremum of an expectation is smaller than or equal to the expectation of a supremum, such that:

$$\begin{aligned} &E_{X\sim\mathcal{D}^n}\sup_{f\in\mathcal{F}}\left(E_{\mathcal{D}}\left(\phi_{M,\gamma}(f(x))\right) - \hat{E}_X\left(\phi_{M,\gamma}(f(x))\right)\right) \\ \leq\ & E_{X,Z\sim\mathcal{D}^n}\sup_{f\in\mathcal{F}}\left(\hat{E}_Z\left(\phi_{M,\gamma}(f(z))\right) - \hat{E}_X\left(\phi_{M,\gamma}(f(x))\right)\right) \\ =\ & \frac{1}{n}E_{X,Z\sim\mathcal{D}^n,\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\sigma_i\left(\phi_{M,\gamma}(f(z_i))\right) - \phi_{M,\gamma}(f(x_i))\right) \\ \leq\ & \frac{2}{n}E_{X\sim\mathcal{D}^n,\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\phi_{M,\gamma}(f(x_i))\right| \end{aligned}$$

where we take $\boldsymbol{\sigma} \in \{-1,1\}^n$ a so-called Rademacher random variable with a uniform distribution. Finally, we can use McDiarmid's inequality again to show that, with probability $1-\frac{\delta}{2}$ over X, this is upper bounded by $\frac{2}{n}E_{\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i\phi_{M,\gamma}(f(x_i))\right| + \sqrt{\frac{1}{2n}\ln\frac{2}{\delta}}$. The first term in this expression is twice the so-called empirical Rademacher complexity of the function class $\mathcal{H}_{M,\gamma} = \{h = \phi_{M,\gamma}\circ f \text{ with } f\in\mathcal{F}\}$:

$$\hat{\mathcal{R}}_X(\mathcal{H}_{M,\gamma}) = \frac{1}{n}E_{\boldsymbol{\sigma}}\sup_{h\in\mathcal{H}_{M,\gamma}}\left|\sum_{i=1}^{n}\sigma_i h(x_i)\right|.$$

Since $\exists a : \phi_{M,\gamma}(a) = 0$ and $\phi_{M,\gamma}$ is an L-Lipschitz-function with $L=\frac{1}{\gamma}$, we can invoke Lemma 3 from Bartlett *et al.* (2002) to obtain that: $\hat{\mathcal{R}}_X(\mathcal{H}_{M,\gamma}) \leq \frac{2}{\gamma}\hat{\mathcal{R}}_X(\mathcal{F})$. Hence it suffices to bound $\hat{\mathcal{R}}_X(\mathcal{F})$:

$$\begin{aligned} n\hat{\mathcal{R}}_X(\mathcal{F}) &= E_{\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right| = E_{\boldsymbol{\sigma}}\sup_{k\in\mathcal{K}}\sup_{\boldsymbol{\alpha}}\left|\frac{\boldsymbol{\sigma}'\mathbf{K}\boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}}}\right| \\ &= E_{\boldsymbol{\sigma}}\sup_{k\in\mathcal{K}}\sup_{\boldsymbol{\alpha}}\left|\left(\sqrt{\mathbf{K}}\boldsymbol{\sigma}\right)'\left(\frac{\sqrt{\mathbf{K}}\boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}}}\right)\right| \\ &\leq E_{\boldsymbol{\sigma}}\sup_{k\in\mathcal{K}}\sqrt{\boldsymbol{\sigma}'\mathbf{K}\boldsymbol{\sigma}} \leq \sqrt{E_{\boldsymbol{\sigma}}\sup_{k\in\mathcal{K}}\boldsymbol{\sigma}'\mathbf{K}\boldsymbol{\sigma}} \end{aligned}$$

The value of $E_{\boldsymbol{\sigma}}\sup_{k\in\mathcal{K}}\boldsymbol{\sigma}'\mathbf{K}\boldsymbol{\sigma}$ can be upper bounded in two ways. From $\boldsymbol{\sigma}'\mathbf{K}_j\boldsymbol{\sigma} \leq \boldsymbol{\sigma}'\boldsymbol{\sigma}\lambda(\mathbf{K}_j) = n\lambda(\mathbf{K}_j)$, it is clear that $n\max_j\lambda(\mathbf{K}_j)/\beta_j$ is an upper bound. Alternatively, observe that $E_{\boldsymbol{\sigma}}\sup_{k\in\mathcal{K}}\sigma'\mathbf{K}\sigma \leq E_{\boldsymbol{\sigma}\boldsymbol{\sigma}'}\left(\sum_{j=1}^{m}\mathbf{K}_j/\beta_j\right)\boldsymbol{\sigma} = \sum_{j=1}^{m}\text{trace}\left(\mathbf{K}_j\right)/\beta_j$. Putting all pieces together completes the proof. □

REFERENCES

Bartlett, P., Bousquet, O., and Mendelson, S. (2002). Localized rademacher complexities. In *Proceedings of the 15th annual conference on Computational Learning Theory (COLT02)*, pages 44–58.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, **3**, 463–482.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, **5**, 27–72.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K.