Lecture-18 Supplementary

29-Mar-10

A simple formulation where the ~~feature~~ all tasks are assumed to ~~be~~ ~~low-dimensional~~ share the same (low number of) features:

$$\min_{w,b,\xi} \; \frac{1}{2}\left(\sum_{j=1}^{d} \|a_j\|_2\right)^2 + C\sum_{t,i}\xi_{ti}$$

$$\begin{array}{c|cccccc}
 & v_1 & v_2 & & \cdots & & v_T \\
 & \downarrow & \downarrow & & & & \downarrow \\
 & 1 & 2 & 3 & \cdots & & T \\
\hline
a_1 \longrightarrow 1 & w_{11} & w_{12} & & \cdots & & w_{1T} \\
a_2 \longrightarrow 2 & w_{21} & & & \cdots & & w_{2T} \\
\vdots & & & & & & \\
a_d \longrightarrow d & w_{d1} & & & \cdots & & w_{dT}
\end{array}$$

(I) $\underline{\text{s.t.}}$ $y_{ti}\left(v_t^\top x_{ti}\right) \geq 1-\xi_{ti},\ \xi_{ti}\geq 0$

→ Promotes structured sparsity: if a feature is useful it is employed in all tasks else it is not employed in any task.

Interestingly this formulation can be written as an MKL formulation:

Using "lambda trick" (I) is same as:

$$\min_{\lambda\in\Delta}\;\min_{w,v,\xi}\; \frac{1}{2}\sum_{j=1}^{d} \frac{\sum_{t=1}^{T} w_{jt}^2}{\lambda_j} + C\sum_{t,i}\xi_{ti} \qquad \text{(II)}$$

$\underline{\text{s.t.}}$ $y_{ti}\left(\sum_{j=1}^{d} w_{jt}\, x_{tij}\right) \geq 1-\xi_{ti},\ \xi_{ti}\geq 0$

→ j^{th} feature of i^{th} eg. of t^{th} task.

Let $\bar{w}_{jt} = \frac{w_{jt}}{\sqrt{\lambda_j}}$, Let $M_t = [\underbrace{0}_{d\times d \text{ zero matrix}} \cdots 0\, \underbrace{I}_{d\times d \text{ identity matrix}}\, 0 \cdots 0]_{d\times Td}$; Let $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$ → $\mathrm{diag}(\lambda_1,\ldots,\lambda_d)$

~~Note that~~ Let $w = \begin{bmatrix} w_{11} \\ \vdots \\ w_{d1} \\ w_{12} \\ \vdots \\ w_{d2} \\ \vdots \\ w_{1T} \\ \vdots \\ w_{dT} \end{bmatrix}_{Td\times 1}$ Now $v_t = M_t w$

With this notation it is easy to see the ~~form~~ (II) is:

$$\min_{\lambda \in \Delta} \; \min_{w, b, \xi} \; \frac{1}{2}\|w\|_2^2 + C \sum_{t,i} \xi_{ti}$$

$$\underline{\text{s.t.}} \quad y_{ti}\left(w^T M_t^T \Lambda x_{ti} \; \cancel{}\right) \geq 1 - \xi_{ti}, \; \xi_{ti} \geq 0$$

standard dual (observe there is no 'b' term however)

$$\min_{\lambda \in \Delta} \; \max_{\alpha} \; 1^T \alpha - \frac{1}{2} \alpha^T Y K_\lambda Y \alpha$$

$$\underline{\underline{\text{s.t.}}} \quad 0 \leq \alpha \leq C,$$

no term $y^T \alpha = 0$

III

K_λ is a $Td \times Td$ matrix which is block diagonal. All entries across tasks are zero! Within a task 't' it is a simple conic combination of ~~bases~~ base kernels as linear kernel built ~~are~~ on each feature. Mathematically:

$$K_\lambda(x, y) = \begin{cases} 0 & \text{if } x \;\&\; y \text{ are not from same task} \\ \lambda_1 K_1 + \dots + \lambda_d K_d & \text{if } x \;\&\; y \text{ belong to a task} \end{cases}$$

$$\underbrace{K_1(x,y) = x_1 y_1 \;\cdots\; K_d(x,y) = x_d y_d}_{\text{simple per feature linear kernels}}.$$

[With this $\alpha^T Y K_\lambda Y \alpha$ decomposes into 'T' quadratics one per each task: $\sum_{t=1}^{T} \alpha_t^T Y_t \left(\sum_{i=1}^{d} \lambda_i K_i^t\right) Y_t \alpha_t$]

Now it is easy to see that (III) can be solved using MD or simple MKL etc.