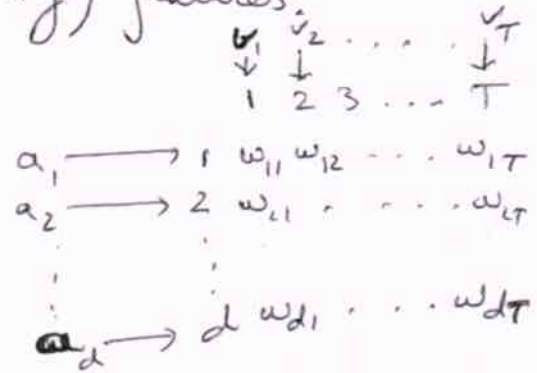


Lecture-18 Supplementary

29-Mar-10

A simple formulation where the ~~feature~~^{all} tasks are assumed to be ~~low-dimensional~~ where the name (low) number of) features:

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



(I) s.t. $y_{ti} (v_t^T x_{ti}) \geq 1 - \xi_{ti}, \xi_{ti} \geq 0$

→ 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 "kernel trick" (I) is same as:

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

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

Let $\bar{w}_{jt} = \frac{w_{jt}}{\sqrt{\lambda_j}}$, Let $M_t = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & & & \end{bmatrix}_{d \times d}$; let $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \dots & \lambda_d \end{bmatrix}$

$d \times d$ zero matrix $d \times d$ identity matrix $\text{diag}(\lambda_1, \dots, \lambda_d)$

~~Note that~~ Let $w = \begin{bmatrix} w_{11} \\ w_{21} \\ w_{31} \\ \vdots \\ w_{d1} \\ \vdots \\ w_{1T} \\ \vdots \\ w_{dT} \end{bmatrix}_{Td \times 1}$

Now $v_t = M_t w$

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

min
 $\lambda \in \mathbb{A}$

min
 w, b, ξ

$$\frac{1}{2} \|w\|_2^2 + C \sum_{t,i} \xi_{t,i}$$

s.t. $y_{t,i} (w^T M_{t,i}^T x_{t,i} - b) \geq 1 - \xi_{t,i}, \xi_{t,i} \geq 0$

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

min
 $\lambda \in \mathbb{A}$

max α
 α $1 - \frac{1}{2} \alpha^T Y K_\lambda Y \alpha$

s.t. $0 \leq \alpha \leq C$

return
 $y^T \alpha = 0$

III

K_λ is a $Td \times Td$ matrix which is block diagonal, all entries across tasks are zero! Within a task it is a simple conic combination of ~~linear~~ base kernels as linear kernel built ~~are~~ on each feature. Mathematically:

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

$K_1(x, y) = x_1 y_1 \dots K_d(x, y) = x_d y_d$

simple per feature linear kernels.

* [With this $\alpha^T Y K_\lambda Y \alpha$ decomposes into T quadratics one per each task: $\sum_{t=1}^T \alpha_t^T Y_t (\sum_{i=1}^d \lambda_i K_i^+) Y_t \alpha_t$]*

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