Learning Using Partially Available Privileged Information and Label Uncertainty: SVM+ with l2 regularizer

Yufeng Zhang

1 Description

- **Problems** Existing LULUPAI has to do multiple direction searches and the number of parameters to be updated are very huge (too many constraints)[1]

- **Solutions** To make it faster, l-2 regularizer is used to avoid too many constraints and the dual forms for LUPI and LULUPAPI are the same as the dual form for one-class SVM, therefore the embedded SVM function in MATLAB can be used to solve optimization problem so as to reduce the computation time.[2]

- Here the SVM we used is ρ-SVM. And also we absorb the bias term into weight vector.

2 One-class SVM

2.1 Definition

To identify outliers amongst the positive examples and use them as negative examples.

2.2 Primal Problem

$$R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i \quad s.t. \quad \|\phi(x_i) - b\|^2 \leq R^2 + \xi_i \quad and \quad \xi_i \geq 0$$

2.3 Lagrange

- $L = R^2 + \frac{1}{\upsilon n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i \left(\left(\phi\left(x_i\right) - b\right)^\top \left(\phi\left(x_i - b\right) - R^2 - \xi_i^2\right) - \sum_{i=1}^{n}\mu_i\xi_i\right.$

- Take derivative with respect to R, b and ξ_i

$$\frac{\partial L}{\partial R} = 0 \Rightarrow \quad \Sigma\alpha_i = 1$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \quad b = \Sigma\alpha_i\phi\left(x_i\right)$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \quad \alpha_i + \mu_i = \frac{1}{\nu n}$$

2.4 Dual Representation

$$-\sum\sum \alpha i\alpha_j k\left(x_i, x_j\right) = \alpha^\top K\alpha \quad s.t. \quad \Sigma\alpha_i = 1 \quad and \quad 0 \leq \alpha_i \leqslant \frac{1}{\nu n}$$

3 Derivation for Learning using Privileged Information and Label Uncertainty

3.1 Primal Problem

- $\min \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{n} z_i\xi^2 - \rho \qquad s.t. \quad y_i(w^T\phi(x_i)) \geq \rho - \xi^2$
- If we use privileged information to construct slack variable then:

$$\min \frac{1}{2}\left(\|w\|^2 + \gamma\|\tilde{w}\|^2\right) + \frac{1}{2}C\sum_{i=1}^{n} z_i\left(\tilde{w}^\top\phi\left(\widetilde{x}_i\right)\right)^2 - \rho \qquad s.t. \quad y_i(w^T\phi(x_i)) \geq \rho - \xi^2$$

3.2 Lagrange

- Construct Lagrange multiplier

$$L = \min \frac{1}{2}\left(\|w\|^2 + \gamma\|\tilde{w}\|^2\right) + \frac{1}{2}C\sum_{i=1}^{n} z_i\left(\tilde{w}^\top\phi\left(\widetilde{x}_i\right)\right)^2 - \rho - \sum_{i=1}^{n}\alpha_i\left(y_i\left(w^\top\phi\left(x_i\right)\right) - \rho + \widetilde{\omega}^\top\phi\left(\tilde{x}_i\right)\right)$$

- Take derivation with respect to $w, \tilde{w}, \rho$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n}\alpha_i y_i\phi\left(x_i\right)$$

$$\frac{\partial L}{\alpha\widetilde{w}} = 0 \Rightarrow \tilde{w} = \sum_{i=1}^{n}\alpha_i\left(\gamma I + C\left(Z\circ P\right)P^\top\right)^{-1}\phi\left(\widetilde{x}_i\right)$$

$$\frac{\partial L}{\partial \rho} = 0 \Rightarrow \quad \sum_{i=1}^{n}\alpha_i = 1 \quad 1'\alpha = 1$$

3.3 Dual representation

- Substitute three equations in (2), (3), (4) back into Lagrangian (1) and we could get the dual representation:

$$\begin{aligned} L = &\frac{1}{2}\alpha^\top\left(K\circ yy^\top\right)\alpha + \frac{\gamma}{2}\alpha^\top\widetilde{K}\left(\gamma I + C(Z\circ P)P^\top\right)^{-2}\alpha \\ &+ \frac{C}{2}\sum_{i=1}^{n} z_i\phi^2\left(\tilde{x}_i\right)\alpha^\top\widetilde{K}\left(\gamma I + C\left(Z\circ P\right)P^\top\right)^{-2}\alpha \\ &- \alpha^\top\left(K\circ yy^\top\right)\alpha - \alpha^\top\widetilde{K}\left(\gamma I + C(Z\circ P)P^\top\right)^{-2}\alpha \end{aligned}$$

where $P = [\phi\left(\widetilde{x}_1\right., ...\phi\left(\widetilde{x}_n\right)]$ And $Z = [z_1, ..., z_n]$

And After simplifying this long expression, we could arrive the simple dual form :

$$L = -\frac{1}{2}\alpha^\top\left(K\circ yy^\top\right)\alpha - \frac{1}{2}\alpha^\top\frac{\widetilde{K}}{(\gamma I + C(Z\circ P)P^\top)^{-1}}\alpha$$

4 Derivation for Learning using Partially Privileged Information and Label Uncertainty

4.1 Primal Problem

$$\frac{1}{2}\|w\|^2+\frac{\gamma}{2}\|\widetilde{w}\|^2+\frac{1}{2}C\sum_{i=m+1}^{n}z_i\xi^2++\frac{1}{2}C^*\sum_{i=1}^{m}z_i\left(\widetilde{w}\phi\left(\widetilde{x}_i\right)\right)^2-\rho$$
$$s.t.\quad 1\leqslant i\leqslant m\quad y_i\left(w^\top\phi\left(x_i\right)\right)\geqslant\rho-\left(\widetilde{w}\phi(\tilde{x}_i)\right)$$
$$m+1\leqslant i\leqslant n\quad y_i\left(w^\top\phi\left(x_i\right)\right)\geqslant\rho-\xi_i$$

4.2 Lagrange

- Construct Lagrange Multiplier:

$$L=\frac{1}{2}\|w\|^2+\frac{\gamma}{2}\|\widetilde{w}\|^2+\frac{1}{2}C\sum_{i=m+1}^{n}z_i\xi_i^2+\frac{1}{2}c^*\sum_{i=1}^{m}z_i\left(\tilde{w}\phi\left(\tilde{x}_i\right)\right)^2-\rho$$
$$-\sum_{i=1}^{m}\alpha_i\left(y_i(w^\top-\phi\left(x_i\right))-\rho+\widetilde{w}\phi(\tilde{x}_i)\right)$$
$$-\sum_{i=m+1}^{n}\beta_i\left(y_i\left(w^\top\phi\left(x_i\right)\right)-\rho+\xi_i\right)$$

- Take derivatives with respect to w, $\widetilde{w}$, ξ_i, ρ:

$$\frac{\alpha L}{\alpha w}=0\Rightarrow w=\sum_{i=1}^{m}\alpha_iy_i\phi\left(x_i\right)+\sum_{i=m+1}^{n}\beta_iy_i\phi\left(x_i\right)$$
$$\frac{\partial L}{\alpha\widetilde{w}}=0\Rightarrow\widetilde{w}=\sum_{i=1}^{m}\alpha_i\left(\gamma I+C^*PP'\right)^{-1}\phi\left(\tilde{x}_i\right)$$
$$\frac{\partial L}{\partial\xi i}=0\Rightarrow\xi_i=\frac{\beta_i}{cz_i}$$
$$\frac{\partial L}{\partial\rho}=0\Rightarrow-1+\sum_{7=1}^{m}\alpha_i+\sum_{i=m+1}^{n}\beta_i=0\Rightarrow\sum_{i=1}^{m}\alpha_i+\sum_{i=m+1}^{n}\beta i=1$$

4.3 Dual representation

Here we should unify α_i and β_i into one variable, because they are all Lagrange multipliers and they can be exchangeable. So in the following parts I use α_i to replace β_i. Thus the dual representation arrives:

$$L=-\frac{1}{2}\alpha^\top\left(K\circ yy^\top\right)\alpha-\frac{1}{2}\alpha^\top\frac{\widetilde{K}}{(\gamma I+C^*(Z\circ P)P^\top)^{-1}}\alpha-\frac{1}{2}\alpha^\top diag(1/(Cz_i))\alpha$$

But here we need to pay attention: the kernel in the first term should be calculated using all the data points; But the second term is only for those data points having privileged information and the third term we only need whose data points without privileged information. Therefore, a matrix consists two blocks are constructed $\begin{bmatrix} G\,0 \\ 0\,F \end{bmatrix}$

(The kernel part of the second term as G and the third term as F).

References

[1] Elyas Sabeti et al. "Learning Using Partially Available Privileged Information and Label Uncertainty: Application in Detection of Acute Respiratory Distress Syndrome". In: *IEEE Journal of Biomedical and Health Informatics* (2020).

[2] Wen Li et al. "Fast algorithms for linear and kernel svm+". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2258–2266.