svr_bennett2006

This bilevel program was first introduced by Bennett et. al. in 2006 [1, Section IV, Equation 1] though it has been presented in many forms since then.

The classic Support Vector Regression (SVR) approach has two hyperparameters, the regularisation constant C and the tube width ϵ. A third hyperparameter λ controls the multitask learning learning [2]. Then $\bar{\mathbf{w}}$ and $\underline{\mathbf{w}}$ are lower and upper bounds, respectively, on the model weights providing feature selection.

The data consists of feature vectors $x_i \in \mathbb{R}^d$ and labels $y_i \in \mathbb{R}$ for $i \in \Omega$. The data indices are split into T distinct partitions Ω_t for $t = 1, \ldots, T$. T-fold cross-validation methodology is used such that Ω_t is used to evaluate the validation loss in the upper-level, while its complement $\bar{\Omega}_t = \Omega \setminus \Omega_t$ is used to evaluate training loss in the lower-level.

The upper-level program seeks to choose optimal hyperparameters $C, \epsilon, \lambda, \underline{\mathbf{w}}, \bar{\mathbf{w}}$ such that the optimal regression weights $\mathbf{w}^t$ chosen by the lower-level (SVR) program result in the minimum average validation loss.

$$\begin{aligned} \underset{C,\epsilon,\lambda,\bar{\mathbf{w}},\underline{\mathbf{w}},\mathbf{w}^t}{\text{minimise}} \quad & \frac{1}{T}\sum_{t=1}^{T}\frac{1}{|\Omega_t|}\sum_{i\in\Omega_t}|x_i^\top \mathbf{w}^t - y_i| \\ \text{subject to} \quad & C, \epsilon, \lambda \geq 0, \\ & \bar{\mathbf{w}} \leq \underline{\mathbf{w}}, \\ & \mathbf{w}^t \text{ solve (SVR)}. \end{aligned}$$

The lower-level program seeks to minimise the sum of ϵ-insensitive residuals over training data $\max\{||x_j^\top \mathbf{w} - y_i - \epsilon, 0\}$. The slack variable $\{e_j\}_{j\in\bar{\Omega}_t}$ are introduced to remove the max operator. Two extra terms are added for regularisation.

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & C\sum_{j\in\bar{\Omega}_t} e_j + \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_0\|_2^2 \\ \text{subject to} \quad & \bar{\mathbf{w}} \leq \mathbf{w} \leq \underline{\mathbf{w}} \\ & \left\{ \begin{array}{l} e_j \geq x_j^\top \mathbf{w} - y_j - \epsilon \\ e_j \geq -x_j^\top \mathbf{w} + y_j - \epsilon \\ e_j \geq 0 \end{array} \right\} \quad \text{for } j \in \bar{\Omega}_t \end{aligned} \tag{SVR}$$

Since the lower-level program is convex and has a Slater's point, Bennett et. al. suggest solving this with a KKT reformulation.

References

[1] Kristin P Bennett, Jing Hu, Xiaoyun Ji, Gautam Kunapuli, and Jong-Shi Pang. Model selection via bilevel optimization. In *The 2006 IEEE Interna-*

tional Joint Conference on Neural Network Proceedings, pages 1922–1929. IEEE, 2006.

[2] Rich Caruana. Multitask learning. *Machine Learning*, 28, 07 1997.