02471 Machine Learning for Signal Processing

Solution

Exercise 13: Support vector regression

13.1 Support vector regression (SVR)

Exercise 13.1.1

We have the model

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x} + \theta_0$$

$$\mathcal{L}(y, f(\boldsymbol{x})) = \begin{cases} |y - f(\boldsymbol{x})| - \epsilon, & \text{if} \quad |y - f(\boldsymbol{x})| > \epsilon \\ 0, & \text{if} \quad |y - f(\boldsymbol{x})| \leq \epsilon \end{cases}$$

Two cases: if $y_n - f(\boldsymbol{x}_n) \geq \epsilon$ then

$$\begin{aligned} y_n - f(\boldsymbol{x}_n) \geq \epsilon \quad &\Leftrightarrow \\ y_n - (\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0) \geq \epsilon \quad &\Leftrightarrow \\ y_n - (\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0) \leq \epsilon + \tilde{\xi}_n & \end{aligned}$$

where $\tilde{\xi}_n$ is chosen big enough for the inequality to be true. The bound is then

$$\tilde{\xi}_n \geq y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0 - \epsilon \geq 0$$

From this we see that the smallest $\tilde{\xi}_n$ we can choose is $\tilde{\xi}_n = 0$, and if this choice is made when $y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0 \leq \epsilon$. In this case, the loss for that particular point would be 0 (because we have the case $|y - f(\boldsymbol{x})| \leq \epsilon$), so ideally, our optimization would select $\boldsymbol{\theta}$ and θ_0 so that $\tilde{\xi}_n = 0$.

The other case is $y_n - f(\boldsymbol{x}_n) \leq -\epsilon$ then

$$\begin{aligned} y_n - (\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0) \leq -\epsilon \quad &\Leftrightarrow \\ \boldsymbol{\theta}^T \boldsymbol{x}_n - y_n - \theta_0 \geq \epsilon & \\ \boldsymbol{\theta}^T \boldsymbol{x}_n - y_n - \theta_0 \leq \epsilon + \xi_n & \end{aligned}$$

Following similar arguments, $\xi_n \geq 0$, and we ideally want the optimization to end up with $\xi_n = 0$.

That means we can restate the optimization problem (with an added regularization) as

$$\begin{aligned} \underset{\boldsymbol{\theta}, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}},}{\arg\min} \quad & J(\boldsymbol{\theta}, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}) := \frac{1}{2} \|\boldsymbol{\theta}\|^2 + C \left(\sum_{n=1}^{N} \xi_n + \sum_{n=1}^{N} \tilde{\xi}_n \right) \\ \text{s.t.} \quad & y_n - f(\boldsymbol{x}_n) \leq \epsilon + \tilde{\xi}_n \\ & -(y_n - f(\boldsymbol{x}_n)) \leq \epsilon + \xi_n \\ & \tilde{\xi}_n \geq 0 \\ & \xi_n \geq 0 \end{aligned}$$

Exercise 13.1.2

We have 4 constraints that can all be written as

$$\begin{aligned}
\text{s.t.} \quad & y_n - f(\boldsymbol{x}_n) - (\epsilon + \tilde{\xi}_n) \leq 0 \quad \Leftrightarrow \quad y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n - \theta_0 - \epsilon - \tilde{\xi}_n \leq 0 \\
& -(y_n - f(\boldsymbol{x}_n)) - (\epsilon + \xi_n) \leq 0 \quad \Leftrightarrow \quad \boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n \leq 0 \\
& \tilde{\xi}_n \geq 0 \\
& \xi_n \geq 0
\end{aligned}$$

Then by using the results from C.2 as stated in the exercise, we introduce the Lagrange multipliers $\tilde{\lambda}_n$, λ_n, $\tilde{\mu_n}$, $\mu_n \geq 0$ to obtain

$$\begin{aligned}
& \tilde{\lambda}_n(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n - \theta_0 - \epsilon - \tilde{\xi}_n) = 0 \\
& \lambda_n(\boldsymbol{\theta}^T \boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n) = 0 \\
& \tilde{\mu}_n \tilde{\xi}_n = 0 \\
& \mu_n \xi_n = 0
\end{aligned}$$

Using the result:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J(\boldsymbol{\theta}) - \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{\theta})$$

will readily give the problem stated in the exercise.

Exercise 13.1.3

We need the rules

$$\begin{aligned}
\frac{\partial a^T \boldsymbol{x}}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{x}^T a}{\partial \boldsymbol{x}} = a \\
\frac{\partial \boldsymbol{x}^T A \boldsymbol{x}}{\partial \boldsymbol{x}} &= (A + A^T)\boldsymbol{x}
\end{aligned}$$

Then we get (only including the terms with $\boldsymbol{\theta}$)

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L} &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta} - \sum_{n=1}^{N} \tilde{\lambda}_n \boldsymbol{\theta}^T \boldsymbol{x}_n + \sum_{n=1}^{N} \lambda_n \boldsymbol{\theta}^T \boldsymbol{x}_n \right) \\
&= \boldsymbol{\theta} - \sum_{n=1}^{N} \tilde{\lambda}_n \boldsymbol{x}_n + \sum_{n=1}^{N} \lambda_n \boldsymbol{x}_n
\end{aligned}$$

Setting this derivate to zero gives

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \sum_{n=1}^{N} \tilde{\lambda}_n \boldsymbol{x}_n - \sum_{n=1}^{N} \lambda_n \boldsymbol{x}_n \\
&= \sum_{n=1}^{N} (\tilde{\lambda}_n - \lambda_n) \boldsymbol{x}_n
\end{aligned}$$

If we have $f(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x} + \theta_0$ we get the following prediction

$$\hat{y} = \hat{\boldsymbol{\theta}}^T \boldsymbol{x} + \hat{\theta}_0 = \sum_{n=1}^{N} (\tilde{\lambda}_n - \lambda_n) \boldsymbol{x}_n^T \boldsymbol{x} + \hat{\theta}_0$$

For the next derivate we get (only including the terms for ξ_n)

$$\begin{aligned}\frac{\partial}{\partial \xi_n}\mathcal{L} &= \frac{\partial}{\partial \xi_n}\left(C\sum_{n=1}^{N}\xi_n + \sum_{n=1}^{N}\lambda_n(-\xi_n) - \sum_{n=1}^{N}\mu_n\xi_n\right)\\ &= C - \lambda_n - \mu_n\end{aligned}$$

Setting this derivate to zero gives

$$\lambda_n + \mu_n = C$$

Since both $\mu_n \geq 0$ and $\lambda_n \geq 0$, we can deduce the constraint

$$0 \leq \lambda_n \leq \ C$$

And finally for θ_0 we get

$$\frac{\partial}{\partial \theta_0}\mathcal{L} = -\sum_{n=1}^{N}\tilde{\lambda}_n + \sum_{n=1}^{N}\lambda_n$$

Setting the derivative to zero gives

$$\sum_{n=1}^{N}\lambda_n = \sum_{n=1}^{N}\tilde{\lambda}_n$$

Exercise 13.1.4

For $\xi_n > 0$, the reason why this implies $\mu_n = 0$ is because we have a constraint $\xi_n\mu_n = 0$. The second implication is then due to the solution from previous exercise: $C = \lambda_n + \mu_n$. If $\mu_n = 0$, then $C = \lambda_n$.

For $\xi_n = 0$, we get

$$\begin{aligned}\lambda_n(\boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n) = 0 \quad &\Rightarrow\\ \lambda_n(\boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 - y_n - \epsilon) = 0\end{aligned}$$

If the prediction is exactly $-\epsilon$ off, we have $y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 = -\epsilon$, which leads to

$$\begin{aligned}\lambda_n(\boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 - y_n - \epsilon - \xi_n) = 0 \quad &\Rightarrow\\ \lambda_n(\epsilon - \epsilon) = 0\end{aligned}$$

This implies that λ_n can be any value. By similar arguments, we see that if $y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0 < \epsilon$, the same constraint enforces $\lambda_n = 0$.