Outline

t_2 Loss Function

- We start with a linear model for regression as

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$

and we have used the squared loss in ordinary linear regression

$$E_2^t(r^t, f(\mathbf{x}^t)) = |r^t - f(\mathbf{x}^t)|^2$$

- Total loss:

$$E_2 = \sum_t E_2^t(r^t, f(\mathbf{x}^t)) = \sum_t |r^t - f(\mathbf{x}^t)|^2$$

- Squared regression (or least squares regression):

$$\underset{\mathbf{w},\, w_0}{\text{minimize}} \quad \frac{1}{N} \sum_{t=1}^{N} |r^t - f(\mathbf{x}^t)|^2$$

ϵ-Insensitive Loss Function – I

- In order for the sparseness property of support vectors in SVM for classification to carry over to regression, we do not use the squared loss but the ϵ-insensitive loss function:

$$E_\epsilon^t(r^t, f(\mathbf{x}^t)) = (|r^t - f(\mathbf{x}^t)| - \epsilon)_+ = \begin{cases} 0 & \text{if } |r^t - f(\mathbf{x}^t)| \le \epsilon \\ |r^t - f(\mathbf{x}^t)| - \epsilon & \text{otherwise} \end{cases}$$

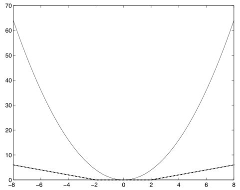
- Two characteristics:
  - Errors are tolerated up to a threshold of ϵ, i.e., no loss for point lying inside an ϵ-tube around the prediction.
  - Errors beyond ϵ have a linear (rather than quadratic) effect so that the model is more more tolerant to noise and robust against noise.
- Total loss:

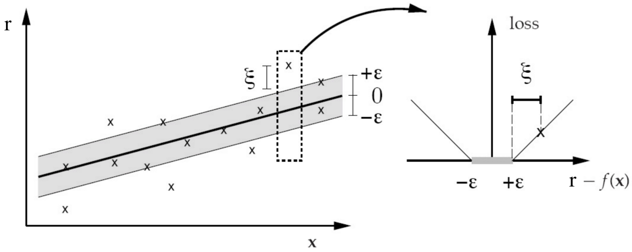
$$E_\epsilon = \sum_t E_\epsilon^t(r^t, f(\mathbf{x}^t)) = \sum_t (|r^t - f(\mathbf{x}^t)| - \epsilon)_+$$

- Tube regression:

$$\underset{\mathbf{w},\, w_0}{\text{minimize}} \quad \frac{1}{N} \sum_{t=1}^{N} (|r^t - f(\mathbf{x}^t)| - \epsilon)_+$$

ϵ-Insensitive Loss Function – II





Support Vector Regression

- Support vector (machine) regression (SVR) is given as

$$\underset{\mathbf{w},\, w_0}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t (|r^t - f(\mathbf{x}^t)| - \epsilon)_+$$

  where C trades off the model complexity (i.e., the flatness of the model) and data misfit.
- The value of ϵ determines the width of the tube (a smaller value indicates a lower tolerance for error) and also affects the number of support vectors and, consequently, the solution sparsity.
  - If ϵ is decreased, the boundary of the tube is shifted inward. Therefore, more datapoints are around the boundary indicating more support vectors.
  - Similarly, increasing ϵ will result in fewer points around the boundary.
- A convex problem, but not a standard QP.
- We will rewrite it to a form similar to SVM which can be QP-solvable.

Primal Optimization Problem

- We introduce slack variables ξ_t^+ and ξ_t^- to account for deviations out of the ϵ-zone.
- Primal optimization problem:

$$\begin{aligned}
\underset{\mathbf{w},\, w_0,\, \{\xi_t^+\},\, \{\xi_t^-\}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t(\xi_t^+ + \xi_t^-) \\
\text{subject to} \quad & r^t - (\mathbf{w}^T\mathbf{x}^t + w_0) \leq \epsilon + \xi_t^+, \quad \forall t \\
& (\mathbf{w}^T\mathbf{x}^t + w_0) - r^t \leq \epsilon + \xi_t^-, \quad \forall t \\
& \xi_t^+, \xi_t^- \geq 0, \quad \forall t
\end{aligned}$$

  which is a standard QP.
- Two types of slack variables:
  - ξ_t^+: for positive deviation such that $r^t - (\mathbf{w}^T\mathbf{x}^t + w_0) > \epsilon$.
  - ξ_t^-: for negative deviation such that $(\mathbf{w}^T\mathbf{x}^t + w_0) - r^t > \epsilon$.
- If $r^t - (\mathbf{w}^T\mathbf{x}^t + w_0) \leq \epsilon$ and $(\mathbf{w}^T\mathbf{x}^t + w_0) - r^t \leq \epsilon$, then $\xi_t^+ = \xi_t^- = 0$, contributing no cost to the objective function.

Lagrangian

- Similar to SVM for classification, the optimization problem for SVR can also be rewritten in the dual form.
- Lagrangian:

$$
\begin{aligned}
&\mathcal{L}(\mathbf{w}, w_0, \{\xi_t^+\}, \{\xi_t^-\}, \{\alpha_t^+\}, \{\alpha_t^-\}, \{\mu_t^+\}, \{\mu_t^-\}) \\
=&\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t(\xi_t^+ + \xi_t^-) \\
&- \sum_t \alpha_t^+\left[\epsilon + \xi_t^+ - r^t + (\mathbf{w}^T\mathbf{x}^t + w_0)\right] - \sum_t \alpha_t^-\left[\epsilon + \xi_t^- + r^t - (\mathbf{w}^T\mathbf{x}^t + w_0)\right] \\
&- \sum_t(\mu_t^+\xi_t^+ + \mu_t^-\xi_t^-)
\end{aligned}
$$

where $\alpha_t^+, \alpha_t^-, \mu_t^+, \mu_t^- > 0$.

Eliminating Primal Variables

- Setting the gradients of $\mathcal{L}$ w.r.t. $\mathbf{w}$, w_0, $\{\xi_t^+\}$, and $\{\xi_t^-\}$ to 0:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{w} = \sum_t (\alpha_t^+ - \alpha_t^-)\mathbf{x}^t \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \quad \Rightarrow \quad \sum_t (\alpha_t^+ - \alpha_t^-) = 0 \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_t^+} = 0 \quad \Rightarrow \quad \mu_t^+ = C - \alpha_t^+, \quad \forall t \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_t^-} = 0 \quad \Rightarrow \quad \mu_t^- = C - \alpha_t^-, \quad \forall t \tag{13}$$

- Plugging (9), (10), (11), and (12) into $\mathcal{L}$ gives the objective function G for the dual problem:

$$G(\{\alpha_t^+\}, \{\alpha_t^-\}) = -\frac{1}{2}\sum_t \sum_{t'} (\alpha_t^+ - \alpha_t^-)(\alpha_{t'}^+ - \alpha_{t'}^-)(\mathbf{x}^t)^T \mathbf{x}^{(t')}$$
$$- \epsilon \sum_t (\alpha_t^+ + \alpha_t^-) + \sum_t r^t (\alpha_t^+ - \alpha_t^-)$$

Dual Optimization Problem – I

- Dual optimization problem:

$$\begin{aligned}
\underset{\{\alpha_t^+\}, \{\alpha_t^-\}}{\text{maximize}} \quad & -\frac{1}{2}\sum_t \sum_{t'} (\alpha_t^+ - \alpha_t^-)(\alpha_{t'}^+ - \alpha_{t'}^-)(\mathbf{x}^t)^T \mathbf{x}^{(t')} \\
& - \epsilon \sum_t (\alpha_t^+ + \alpha_t^-) + \sum_t r^t (\alpha_t^+ - \alpha_t^-) \\
\text{subject to} \quad & \sum_t (\alpha_t^+ - \alpha_t^-) = 0 \\
& 0 \le \alpha_t^+ \le C, \ \forall t \\
& 0 \le \alpha_t^- \le C, \ \forall t
\end{aligned}$$

- Instances in the ϵ-tube ($\alpha_t^+ = \alpha_t^- = 0$) are instances fitted with enough precision.
- The support vectors satisfy either $\alpha_t^+ > 0$ or $\alpha_t^- > 0$ and are of two types.
  - instances on the boundary of the ϵ-tube (either $0 < \alpha_t^+ < C$ or $0 < \alpha_t^- < C$), and we use these to calculate w_0
  - instances outside the ϵ-tube are instances for which we do not have a good fit (either $\alpha_t^+ = C$ or $\alpha_t^- = C$)

Dual Optimization Problem – II

- We have the fitted line as a weighted sum of the support vectors:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 = \sum_{\mathbf{x}^t \in \mathcal{SV}} (\alpha_t^+ - \alpha_t^-)(\mathbf{x}^t)^T \mathbf{x} + w_0$$

- Due to the sparseness property of the ϵ-insensitive loss function, only a small fraction of the training instances are support vectors which are used in defining the regression function (like the discriminant function for classification).
- Nonlinear (kernel) extension is possible by introducing appropriate kernel functions.

SVR