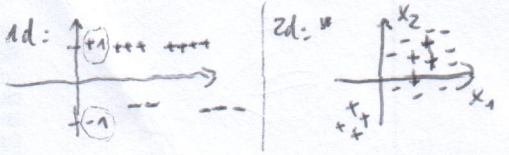
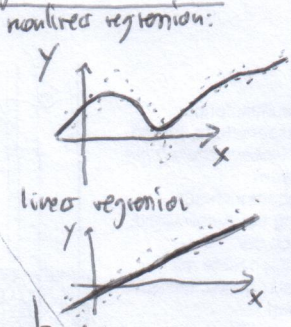
Support Vector Regression

reminder: classification problems in contrast have labels with only finitely many possible values, e.g., $y_T^{(\alpha)} \in \{-1, +1\}$ for binary classification





Data: $\{(\underline{x}^{(\alpha)}, y_T^{(\alpha)})\}$

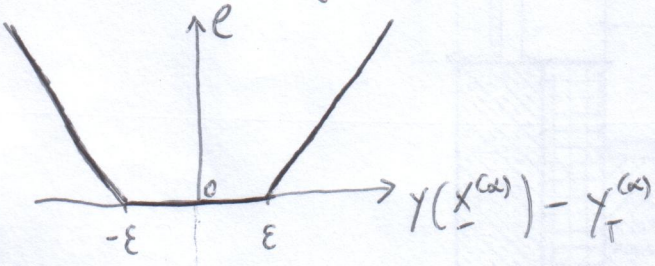
$\underline{x}^{(\alpha)} \in \mathbb{R}^N$, $y_T^{(\alpha)} \in \mathbb{R}$

Model: linear mapping $y(\underline{x}) = \underline{w}^T \underline{x} + b$

$\underline{w}$: normal vector of hyperplane; b: offset of hyperplane

a) **ε-sensitive cost function**

$$e(\underline{x}^{(\alpha)}, y_T^{(\alpha)}) = \max\left(0, |y(\underline{x}^{(\alpha)}) - y_T^{(\alpha)}| - \varepsilon\right)$$



$y_T^{(\alpha)} - y(\underline{x}^{(\alpha)}) > \varepsilon$ $\quad$ $y(\underline{x}^{(\alpha)}) - y_T^{(\alpha)} > \varepsilon$

$0 < y_T^{(\alpha)} - y(\underline{x}^{(\alpha)}) \leq \varepsilon$ $\quad$ $0 \leq y(\underline{x}^{(\alpha)}) - y_T^{(\alpha)} \leq \varepsilon$

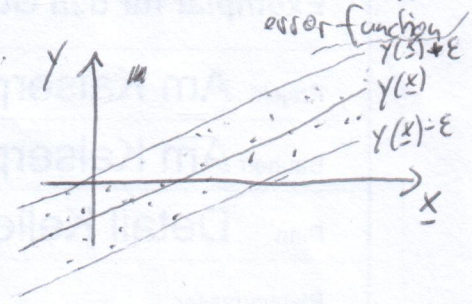
$\Rightarrow$ cost function parametrized by $\varepsilon \geq 0$

$\Rightarrow$ no error for model output $y(\underline{x}^{(\alpha)}) \in [y_T^{(\alpha)} - \varepsilon, y_T^{(\alpha)} + \varepsilon]$

$\Rightarrow$ linear error outside this interval

b) **Derivation of the primal optimization problem for the "ε-SVR":**

i) simplifying assumptions (later relaxed): (1) perfect regression solution (0 error) and (2) linear regression problem (w.r.t. ε-sensitive error function)



$$\min_{\underline{w}, b} \ \frac{1}{2}\|\underline{w}\|^2$$

s.t. $y(\underline{x}^{(\alpha)}) - y_T^{(\alpha)} \leq \varepsilon$ ← parameter defining the zero-error intervall

$y_T^{(\alpha)} - y(\underline{x}^{(\alpha)}) \leq \varepsilon$ $\quad$ for $\alpha = 1, \dots, P$

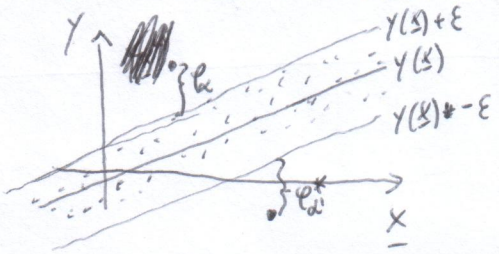
$\Leftrightarrow$

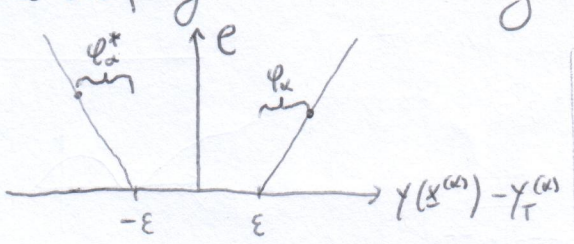
$$\min_{\underline{w}, b} \ \frac{1}{2}\|\underline{w}\|^2$$

s.t. $\underline{w}^T \underline{x}^{(\alpha)} + b - y_T^{(\alpha)} \leq \varepsilon$

$y_T^{(\alpha)} - \underline{w}^T \underline{x}^{(\alpha)} - b \leq \varepsilon$

b) ii) extension: allow some error to reduce overfitting to noise in training data (cf. C-SVM)





φ_α: "positive" slack variable

φ_α^*: "negative" slack variable

primal optimization problem ("ε-SVR"):

$$\min_{\underline{w}, b, \{\varphi_\alpha\}, \{\varphi_\alpha^*\}} \quad \frac{1}{2}\|\underline{w}\|^2 + C\,\frac{1}{P}\sum_{\alpha=1}^{P}\left(\varphi_\alpha + \varphi_\alpha^*\right)$$

($\varphi_\alpha, \varphi_\alpha^*$: new variables (slacks); C: parameter weighting model error (to model complexity))

ε-SVR

$$\text{s.t.} \quad \underline{w}^T \underline{x}^{(\alpha)} + b - y_T^{(\alpha)} \le \varepsilon + \varphi_\alpha$$

$$y_T^{(\alpha)} - \underline{w}^T \underline{x}^{(\alpha)} - b \le \varepsilon + \varphi_\alpha^*$$

$$\varphi_\alpha \ge 0 \,, \quad \varphi_\alpha^* \ge 0 \qquad \text{for } \alpha = 1, \dots, P$$

(← new constraints)

Derivation of the primal problem for the "ν-SVR"

c) • choosing ε properly can be difficult as it depends on the unknown data noise
• extension: make ε a variable with respect to which we optimize:

$$\min_{\underline{w}, b, \{\varphi_\alpha\}, \{\varphi_\alpha^*\}, \varepsilon} \quad \frac{1}{2}\|\underline{w}\|^2 + C\left[\nu\,\varepsilon + \frac{1}{P}\sum_{\alpha=1}^{P}\left(\varphi_\alpha + \varphi_\alpha^*\right)\right]$$

(ε: new variable; ν: new parameter)

primal optimization problem of the ν-SVR

$$\text{s.t.} \quad \underline{w}^T \underline{x}^{(\alpha)} + b - y_T^{(\alpha)} \le \varepsilon + \varphi_\alpha$$

$$y_T^{(\alpha)} - \underline{w}^T \underline{x}^{(\alpha)} - b \le \varepsilon + \varphi_\alpha^* \qquad \text{for } \alpha = 1, \dots, P$$

$$\varphi_\alpha \ge 0 \,, \quad \varphi_\alpha^* \ge 0 \,, \quad \varepsilon \ge 0$$

(← new constraint)

• it can be shown that $\nu \le \frac{\#SV}{P}$, i.e., ν lower bounds the fraction of support vectors (on $\pm\varepsilon$ tube or in zero error region)

• ν-SVR is standard SV-based regression method when combined with a suitable kernel (e.g. RBF)

• ν has still to be chosen (e.g. cross-validation based hyperparam. selection)

d) Lagrangian:

$$\mathcal{L}\Big(\underbrace{\underline{w}, b, \{\varphi_\alpha\}, \{\varphi_\alpha^*\}, \varepsilon}_{\text{primal variables}}, \underbrace{\{\lambda_\alpha\}, \{\lambda_\alpha^*\}, \{\eta_\alpha\}, \{\eta_\alpha^*\}, \delta}_{\text{dual variables (Lagrange multipliers)}}\Big)$$

$$= \frac{1}{2}\|\underline{w}\|^2 + C\left[\nu\varepsilon + \frac{1}{p}\sum_{\alpha=1}^{p}\left(\varphi_\alpha + \varphi_\alpha^*\right)\right]$$

$$- \sum_{\alpha=1}^{p}\lambda_\alpha\left(\varphi_\alpha + \varepsilon + y_T^{(\alpha)} - \underline{w}^T\underline{x}^{(\alpha)} - b\right)$$

$$- \sum_{\alpha=1}^{p}\lambda_\alpha^*\left(\varphi_\alpha^* + \varepsilon + \underline{w}^T\underline{x}^{(\alpha)} + b - y_T^{(\alpha)}\right)$$

$$- \sum_{\alpha=1}^{p}\eta_\alpha\varphi_\alpha - \sum_{\alpha=1}^{p}\eta_\alpha^*\varphi_\alpha^* - \delta\varepsilon$$

s.t. $\lambda_\alpha \geq 0$, $\lambda_\alpha^* \geq 0$, $\eta_\alpha \geq 0$, $\eta_\alpha^* \geq 0$, $\delta \geq 0$ $\forall\alpha$

---

in the homework you will see/show that both i) the dual optimization problem as well as ii) the model evaluation for new data $\underline{x}$, i.e. $y(\underline{x})$, depends on the data exclusively via scalar products $\underline{x}^{(\alpha)T}\underline{x}^{(\beta)}$ [and $\underline{x}^{(\alpha)T}\underline{x}$ respectively]

e.g. final hyper-para opt: $C, \sigma(\gamma)$, ν

$\Longrightarrow$ kernel usage: replace these scalar products by kernel evaluation $k(\underline{x}^{(\alpha)}, \underline{x}^{(\beta)})$ which then implicitly does a linear regression on the ~~nonlinearly expanded~~ data set in feature space, $\{(\underline{\Phi}(\underline{x}^{(\alpha)}), y_T^{(\alpha)})\}_\alpha$, i.e. $\underline{x}$ expanded by nonlinear transformation $\Phi(\underline{x})$ with $\Phi(\underline{x}^{(\alpha)})^T\Phi(\underline{x}^{(\beta)}) = k(\underline{x}^{(\alpha)}, \underline{x}^{(\beta)})$

$\Uparrow$ yields nonlinear regression in data space $\Uparrow$