Interval-valued regression and classification models in the framework of machine learning

Lev V. Utkin and Frank P.A. Coolen

St.Petersburg Forest Technical University, Durham University

lev.utkin@mail.ru, frank.coolen@durham.ac.uk

1. Regression and classification in the machine learning framework

Given: a training set $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, $\mathbf{x} \in \mathbb{R}^m$ is a multivariate input of features and a scalar output: regression: $y \in \mathbb{R}$, classification: binary $y \in \{-1, 1\}$ or multi-class $y \in \{1, 2, ..., l\}$.
The learning problem: to select a function $f(\mathbf{x}, w_{\text{opt}})$ from a set of functions $f(\mathbf{x}, w)$ parameterized by a set of parameters $w \in \Lambda$, which
regression: best approximates the system response y
classification: separates examples of different classes y.
A general problem solution: To minimize the risk measure $R(w)$ over $w \in \Lambda$:
regression

$$R(w) = \int_{\mathbb{R}} L(z, w) \mathrm{d}F(z), \ z = y - f(\mathbf{x}, w).$$

classification

$$R(w) = \int_{\mathbb{R}^m \times \{-1,1\}} L(y, f) \mathrm{d}F_0(\mathbf{x}, y).$$

Loss functions in regression: the quadratic loss $L(z) = z^2$, the linear loss $L(z) = |z|$, the so-called 'ε-insensitive' and 'pinball' loss functions.
In classification: the indicator loss function $L(\mathbf{x}, y) = \mathbf{1}\{\text{sgn}(f(\mathbf{x}, w)) \neq y\}$, the logistic loss, the hinge loss $L(y, f(\mathbf{x}, w)) = \max(0, 1 - yf)$, the squared hinge loss.

2. Regression with a set of distributions

We do not know the precise CDF $F(z)$, but we know that it belongs to a set $\mathcal{F}(w)$ bounded by lower CDF $\underline{F}(z \mid w)$ and upper CDF $\overline{F}(z \mid w)$ inferrred from $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$.
The minimax strategy: It can be interpreted as an insurance against the worst case. The upper risk measure is

$$\overline{R}(w) = \max_{F(z \mid w) \in \mathcal{F}(w)} \int L(z \mid w) \mathrm{d}F(z \mid w).$$

Loss functions in regression models have one minimum at point 0. The optimal CDF is

$$F_U(z) = \begin{cases} \overline{F}(z), & z \leq \overline{F}^{-1}(\tau), \\ \tau, & \overline{F}^{-1}(\tau) < z < \underline{F}^{-1}(\tau), \\ \underline{F}(z), & z \geq \underline{F}^{-1}(\tau), \end{cases}$$

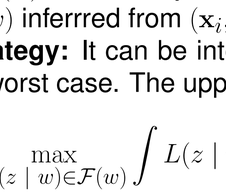
where τ is one of the roots of the equation $L\left(\overline{F}^{-1}(\tau)\right) = L\left(\underline{F}^{-1}(\tau)\right)$.
The minimin strategy: It can be interpreted as an 'optimistic' decision. The lower risk measure is

$$\underline{R}(w) = \min_{F(z \mid w) \in \mathcal{F}(w)} \int L(z \mid w) \mathrm{d}F(z \mid w).$$

The optimal CDF is

$$F_L(z) = \begin{cases} \underline{F}(z), \ z \leq 0, \\ \overline{F}(z), \ z > 0. \end{cases}$$

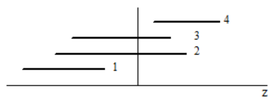
3. Regression with interval-valued observations



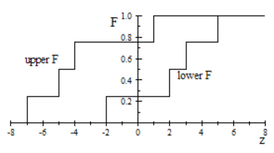
The training set: $(\mathbf{x}_i, \mathcal{Y}_i)$, $i = 1, ..., n$, with intervals $\mathcal{Y}_i = [\underline{y}_i, \overline{y}_i]$. The random noise $Z \in \mathcal{Z}_i(w) = \mathcal{Y}_i - f(\mathbf{x}_i, w)$.
A p-box is constructed in the framework of Dempster-Shafer theory

$$\underline{F}(z \mid w) = \text{Bel}((-\infty, z]) = n^{-1} \sum\nolimits_{i: \overline{\mathcal{Z}}_i(w) \leq z} 1,$$
$$\overline{F}(z \mid w) = \text{Pl}((-\infty, z]) = n^{-1} \sum\nolimits_{i: \underline{\mathcal{Z}}_i(w) \leq z} 1.$$

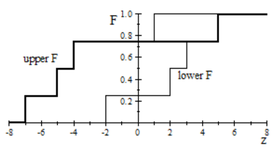
"Optimal distribution functions" (regression) (Utkin & Destercke 2009)



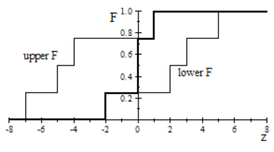
Interval-valued estimates $\left[\underline{\mathcal{Z}}_i(w), \overline{\mathcal{Z}}_i(w)\right]$



Lower and upper probability distributions produced by four intervals



The optimal probability distribution (thick) by the minimax strategy



The optimal probability distribution (thick) by the minimin strategy

3.1 The minimax strategy

The upper risk measure is

$$\overline{R}(w) = n^{-1} \sum\nolimits_{i: \underline{\mathcal{Z}}_i(w) \leq \overline{F}^{-1}(\tau)} L(\underline{\mathcal{Z}}_i(w)) + n^{-1} \sum\nolimits_{i: \overline{\mathcal{Z}}_i(w) \geq \underline{F}^{-1}(\tau)} L(\overline{\mathcal{Z}}_i(w)).$$

with τ such that $\underline{F}^{-1}(\tau) = -\overline{F}^{-1}(\tau)$. The upper risk measure uses only the boundary points $\underline{\mathcal{Z}}_i(w)$ and $\overline{\mathcal{Z}}_i(w)$ of the intervals $\mathcal{Z}_i(w)$! Hence

$$\overline{R}(w) = n^{-1} \sum\nolimits_{i=1}^{n} \max\left\{L(\underline{\mathcal{Z}}_i(w)), L(\overline{\mathcal{Z}}_i(w))\right\}.$$

Linear and pinball loss function:
The pinball loss function with parameter $\tau \in [0, 1]$:

$$L_\tau(z) = \begin{cases} \tau z, & z > 0, \\ (\tau - 1)z, & z \leq 0. \end{cases}$$

The optimisation problem is

$$\min_{w, G_i} \sum\nolimits_{i=1}^{n} G_i,$$

subject to

$$G_i \geq \tau \cdot \underline{\mathcal{Z}}_i(w), \ G_i \geq \tau \cdot \overline{\mathcal{Z}}_i(w),$$
$$G_i \geq (\tau - 1) \cdot \underline{\mathcal{Z}}_i(w), \ G_i \geq (\tau - 1) \cdot \overline{\mathcal{Z}}_i(w), \ i = 1, ..., n.$$

3.2 SVM

The ε-insensitive loss function is applied in the corresponding regression models. If all observations are point-valued, so $\underline{y}_i = \overline{y}_i = y_i$, then parameters w are determined by the quadratic programming problem

$$\min_{w} \left(\frac{1}{2} \langle w, w \rangle + C \sum\nolimits_{i=1}^{n} (\xi_i + \xi_i^*)\right)$$

subject to

$$\begin{aligned} &\xi_i \geq 0, \ \ \xi_i + \varepsilon \geq (\langle w \mathbf{x}_i \rangle + w_0) - \underline{y}_i, \\ &\xi_i \geq 0, \ \ \xi_i + \varepsilon \geq (\langle w \mathbf{x}_i \rangle + w_0) - \overline{y}_i, \\ &\xi_i^* \geq 0, \ \ \xi_i^* + \varepsilon \geq \underline{y}_i - (\langle w \mathbf{x}_i \rangle + w_0), \\ &\xi_i^* \geq 0, \ \ \xi_i^* + \varepsilon \geq \overline{y}_i - (\langle w \mathbf{x}_i \rangle + w_0). \end{aligned}$$

Here C is a constant 'cost' parameter, ξ_i, ξ_i^*, $i = 1, ..., n$, are slack variables, and $\frac{1}{2} \langle w, w \rangle$ is the Tikhonov regularization term (the most popular penalty or smoothness term) which enforces uniqueness by penalizing functions with wild oscillation and effectively restricting the space of admissible solutions. The positive slack variables ξ_i, ξ_i^* represent the distance from y_i to the corresponding boundary values of the ε-tube.

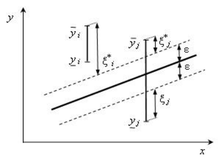


Figure 2: *The relationship between slack variables in SVM and the observed intervals by the minimax strategy*

The minimax strategy searches for the largest residuals (or 'margins' in terms of classification) from all residuals in every interval $\mathcal{Z}_i$, $i = 1, ..., n$.

3.3 The minimin strategy

The lower risk measure is

$$\underline{R}(w) = n^{-1} \sum_{i: \overline{\mathcal{Z}}_i(w) \leq 0} L(\overline{\mathcal{Z}}_i(w)) + n^{-1} \sum_{i: \underline{\mathcal{Z}}_i(w)) \geq 0} L(\underline{\mathcal{Z}}_i(w)).$$

3.4 SVM

The convex optimisation problem

$$\min_{w} \left(\frac{1}{2} \langle w, w \rangle + C \sum\nolimits_{i=1}^{n} (\xi_i + \xi_i^*)\right),$$

subject to

$$\begin{aligned} &\xi_i \geq 0, \ \xi_i + \varepsilon \geq (\langle w \mathbf{x}_i \rangle + w_0) - \overline{y}_i, \ i = 1, ..., n, \\ &\xi_i^* \geq 0, \ \xi_i^* + \varepsilon \geq \underline{y}_i - (\langle w \mathbf{x}_i \rangle + w_0), \ i = 1, ..., n. \end{aligned}$$

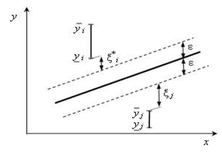
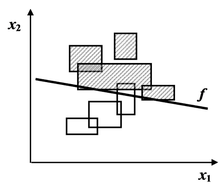


Figure 3: *The relationship between slack variables in SVM and the observed intervals by the minimin strategy*

The minimin strategy searches for the smallest residuals in each interval $\mathcal{Z}_i$, $i = 1, ..., n$.

4. Classification with interval-valued observations



Given: a training set $(\mathcal{X}_i, y_i)$, $i = 1, ..., n$. Here $\mathcal{X}_i \subset \mathbb{R}^m$ is the Cartesian product of m intervals $[\underline{x}_k^{(i)}, \overline{x}_k^{(i)}]$.
Introduce: $f_L = \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \alpha)$, $f_U = \max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \alpha)$.
Then

$$R(w) = \sum_{y=0,1} p(y) \int_{\mathbb{R}} L(f \mid y) \mathrm{d}F(f \mid y).$$

P-boxes:

$$\mathcal{F}(y) = \{F(f) \mid \forall f \in \mathbb{R}, \underline{F}(f|y) \leq F(f) \leq \overline{F}(f|y)\}.$$

4.1 The minimax strategy

The upper risk measure is

$$\overline{R}(w) = \max_{F(f|-1) \in \mathcal{F}(-1)} R_{-1}(w) + \max_{F(f|1) \in \mathcal{F}(1)} R_{+1}(w).$$

Loss functions $L(f, y)$ in classification are increasing if $y = -1$ and decreasing if $y = 1$, so

$$\overline{R}_{-1}(w) = \int_{\mathbb{R}} L(f, -1) \mathrm{d}\underline{F}(f, -1),$$
$$\overline{R}_{+1}(w) = \int_{\mathbb{R}} L(f, 1) \mathrm{d}\overline{F}(f, 1).$$

The linear problem is

$$\min_{w} \left(\frac{p(-1)}{r} \sum\nolimits_{i=1}^{r} G_i + \frac{p(1)}{n - r} \sum\nolimits_{i=r+1}^{n} G_i\right)$$

subject to

$$\begin{aligned} &G_i \geq 1 - y_i (\langle w \mathbf{x}_i \rangle + w_0), \ \forall x_k^{(i)} \in \{\underline{x}_k^{(i)}, \overline{x}_k^{(i)}\}, \\ &G_i \geq 0, \ i = 1, ..., n. \end{aligned}$$

By adding the standard Tikhonov regularization term to the objective function, we get the SVM classifier with cost parameters $C_- = \pi_-/r$ and $C_+ = \pi_+/(n - r)$.