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

Lev Utkin and Frank Coolen

Innsbruck, July 2011

A general problem statement

- **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 functional $R(w)$ over $w \in \Lambda$:

- regression

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

- classification

$$R(w) = \int_{\mathbb{R}^m \times \{-1,1\}} L(y, f)\mathrm{d}F_0(\mathbf{x}, y)$$
$$= \sum_{y=0,1} p(y) \int_{\mathbb{R}^m} L(y, f)\mathrm{d}F_0(\mathbf{x} \mid y).$$

Loss functions

- in regression models: quadratic, linear, the "pinball" function (for quantile regression), **the ε-insensitive loss function**.
- in classification models: indicator, logistic, **hinge loss.**

The main idea

1. It is assumed that the CDF $F(z) \in \mathcal{F}$ bounded by the lower $\underline{F}$ and upper $\overline{F}$ CDFs (P-boxes):

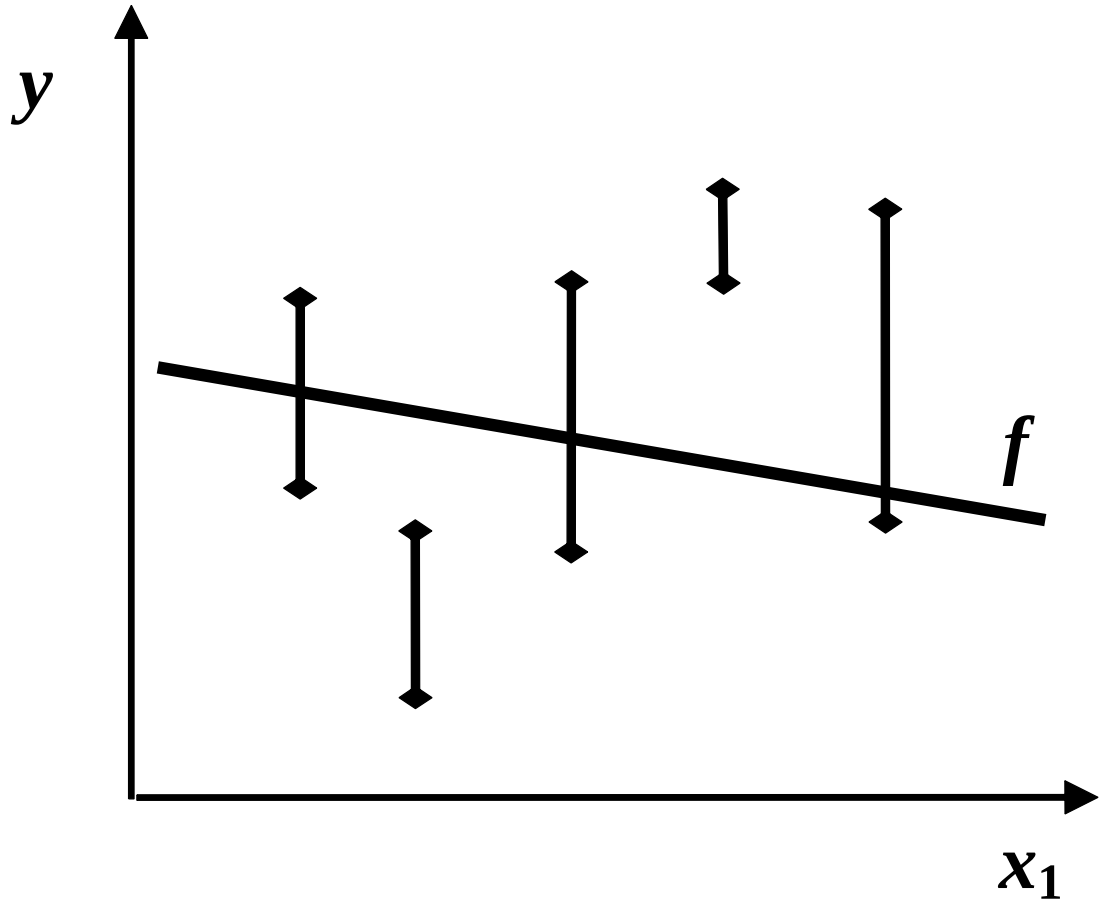
$$\mathcal{F} = \{F(z) \mid \forall z,\ \underline{F}(z \mid y) \leq F(z \mid y) \leq \overline{F}(z \mid y)\}.$$

2. P-boxes are constructed from training data and they are parametric, i.e., $\mathcal{F} \rightarrow \mathcal{F}(w)$.
3. Two CDFs maximizing and minimizing $R(w)$ are taken from $\mathcal{F}$, which determine the largest $\overline{R}$ and smallest $\underline{R}$ risk measures as functions of w.
4. w are computed by minimizing the lower and upper risk measures.

Two main tasks to be solved

1. How to construct parametric P-boxes, i.e., $\underline{F}$ and $\overline{F}$ from the training set?
2. How to find "optimal" distributions from the P-box, i.e., the distributions maximizing and minimizing the risk functional (corresponding to minimax and minimin strategies, respectively)?

Interval regression (simplest case)



Intervals and P-boxes (regression)

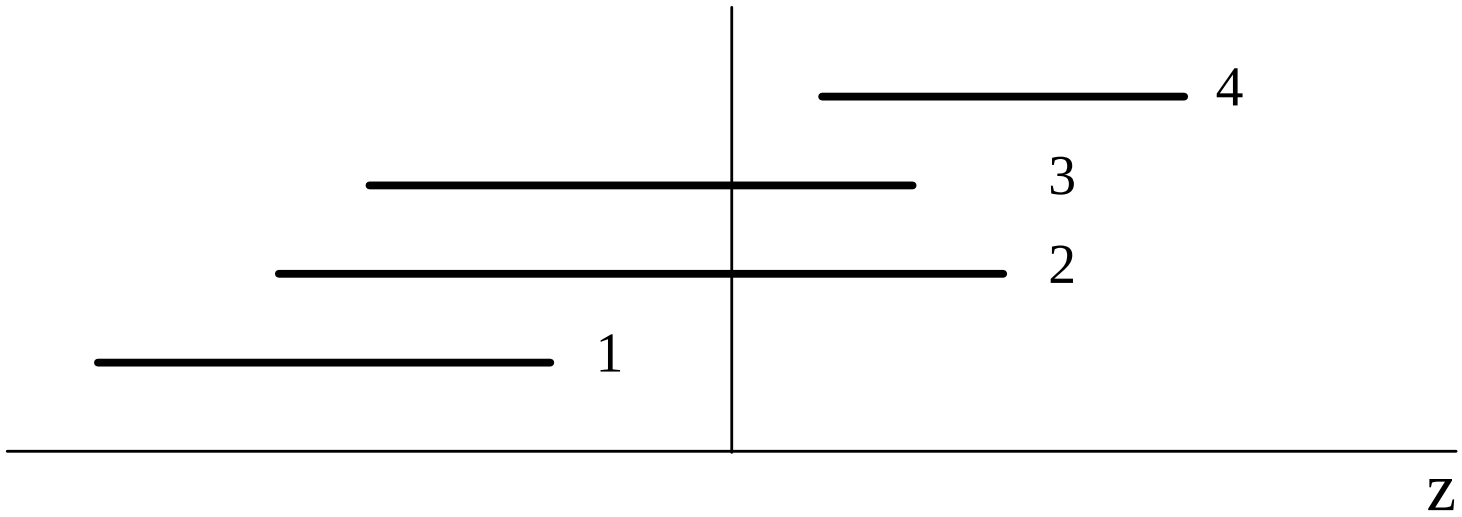
Given: a training set $(\mathbf{x}_i, \mathcal{Y}_i)$, $i = 1, ..., n$, $\mathbf{x} \in \mathbb{R}^m$, $\mathcal{Y}_i = [\underline{y}_i, \overline{y}_i]$.
P-boxes:

$$\underline{F}(z \mid w) = \mathrm{Bel}((-\infty, z]) = n^{-1} \sum_{i:\overline{\mathcal{Z}}_i(w) \le z} 1,$$

$$\overline{F}(z \mid w) = \mathrm{Pl}((-\infty, z]) = n^{-1} \sum_{i:\underline{\mathcal{Z}}_i(w) \le z} 1.$$

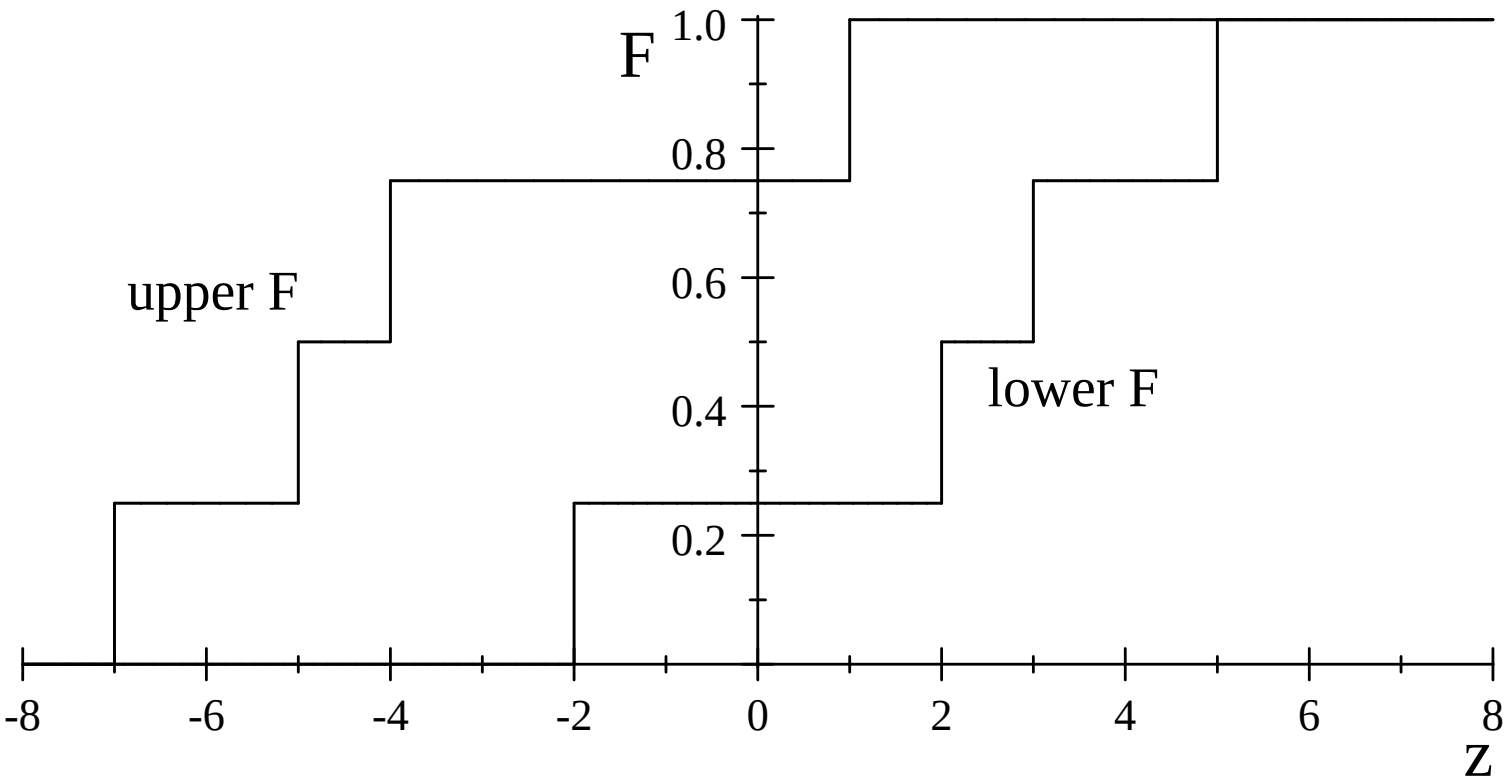
Here $\underline{\mathcal{Z}}_i(w) = \underline{y}_i - f(\mathbf{x}_i, w)$ and $\overline{\mathcal{Z}}_i(w) = \overline{y}_i - f(\mathbf{x}_i, w)$.

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



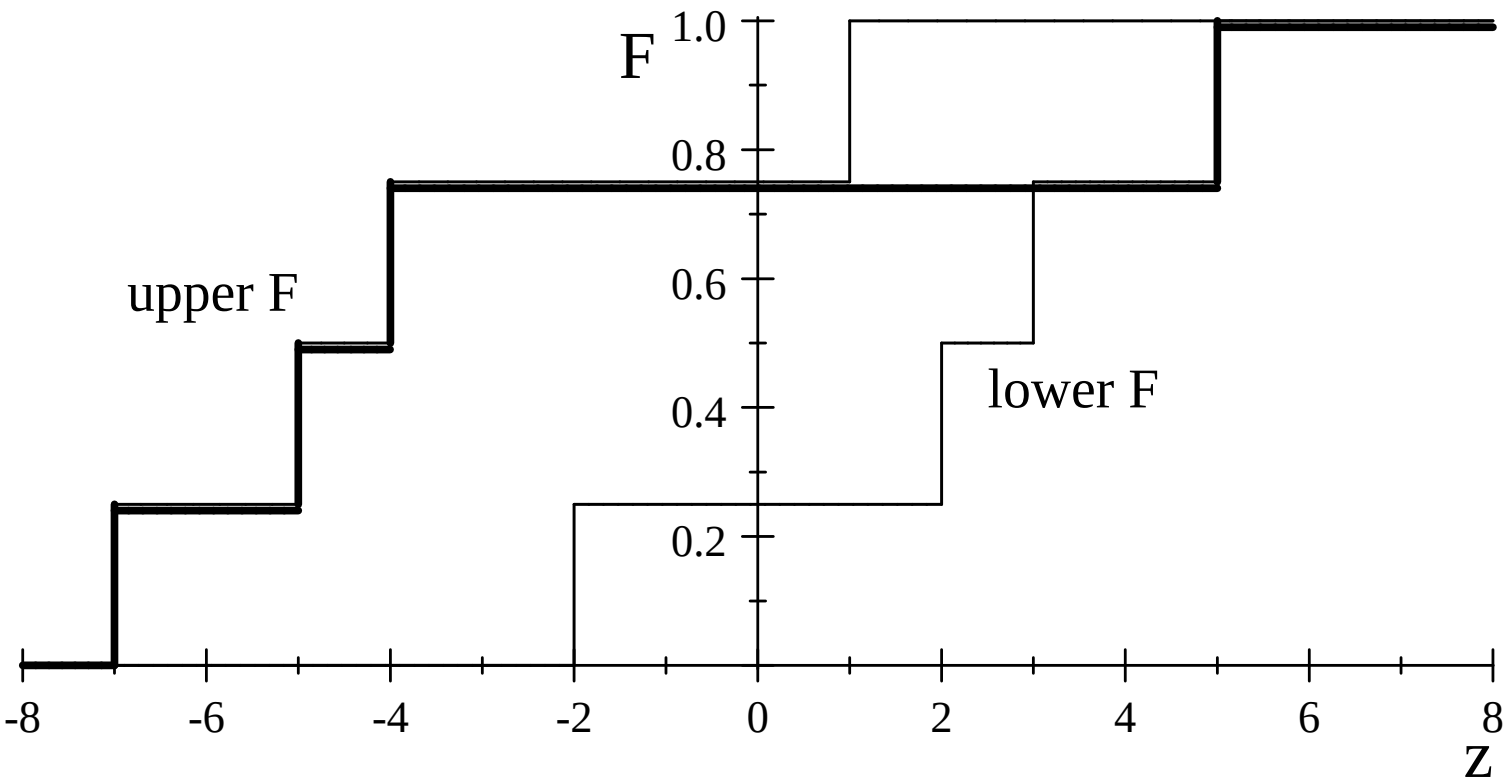
Interval-valued estimates $\left[\underline{\mathcal{Z}}_i(w), \overline{\mathcal{Z}}_i(w)\right]$

Lower and upper CDFs



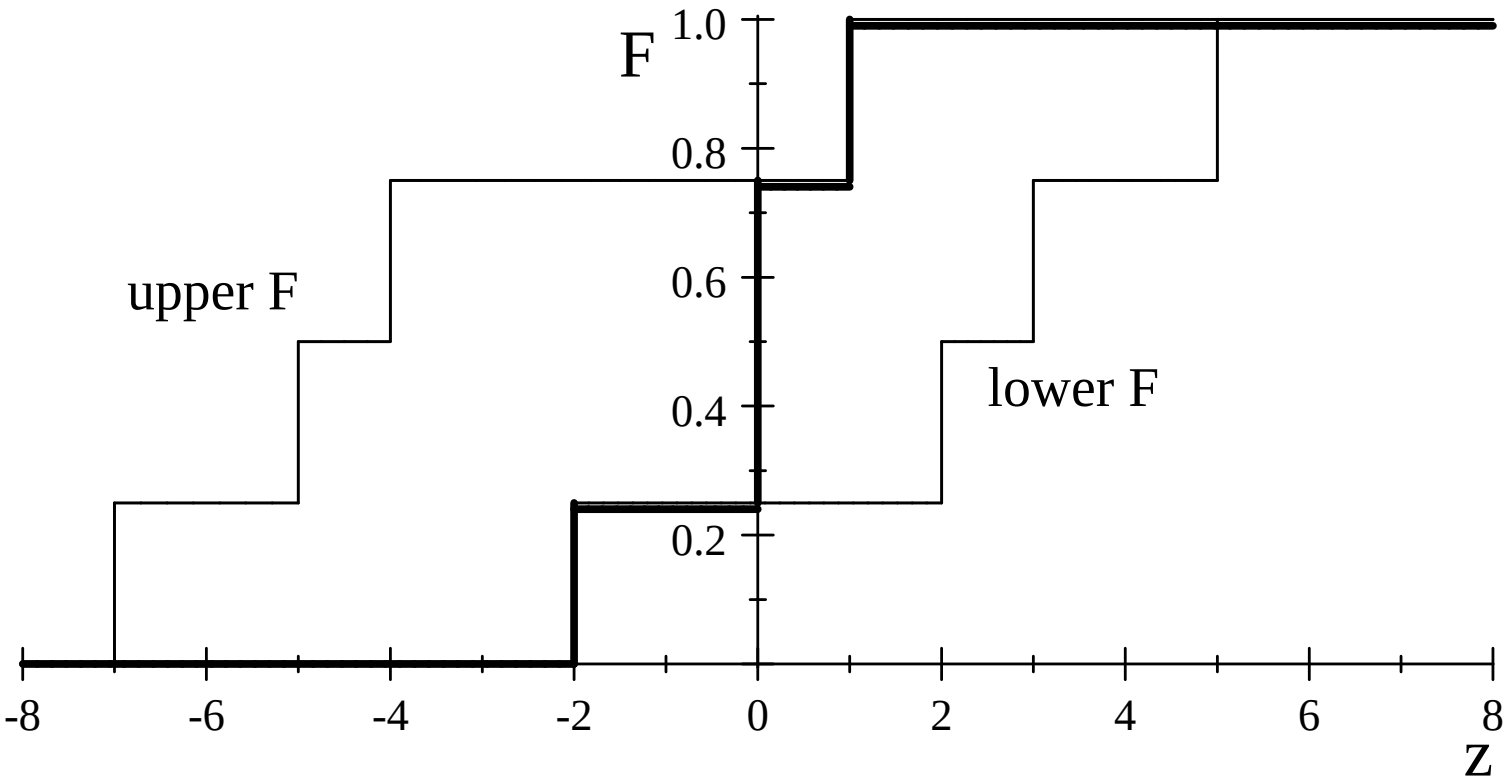
Lower and upper probability distributions produced by four intervals

The optimal CDF by the minimax strategy



The optimal probability distribution (thick) by the minimax strategy

The optimal CDF by the minimin strategy



The optimal probability distribution (thick) by the minimin strategy

Intervals and expectations in the framework of belief structures (regression)

The upper expectation of risk functional (Nguyen & Walker 1994, Strat 1990):

$$\overline{R}(w) = n^{-1} \sum_{i=1}^{n} \max_{z \in [\underline{Z}_i(w), \overline{Z}_i(w)]} L(z) \rightarrow \min_{w}.$$

The optimization problem for computing w:

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

subject to

$$G_i \geq L(\underline{Z}_i(w)), \;\; G_i \geq L(\overline{Z}_i(w)), \quad i = 1, ..., n.$$

If $L(z)$ and $f(\mathbf{x}, w)$ are linear, then we have the LP problem.

Support vector machine (SVM) and interval observations

If we take the ε-insensitive loss function L, then

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

subject to

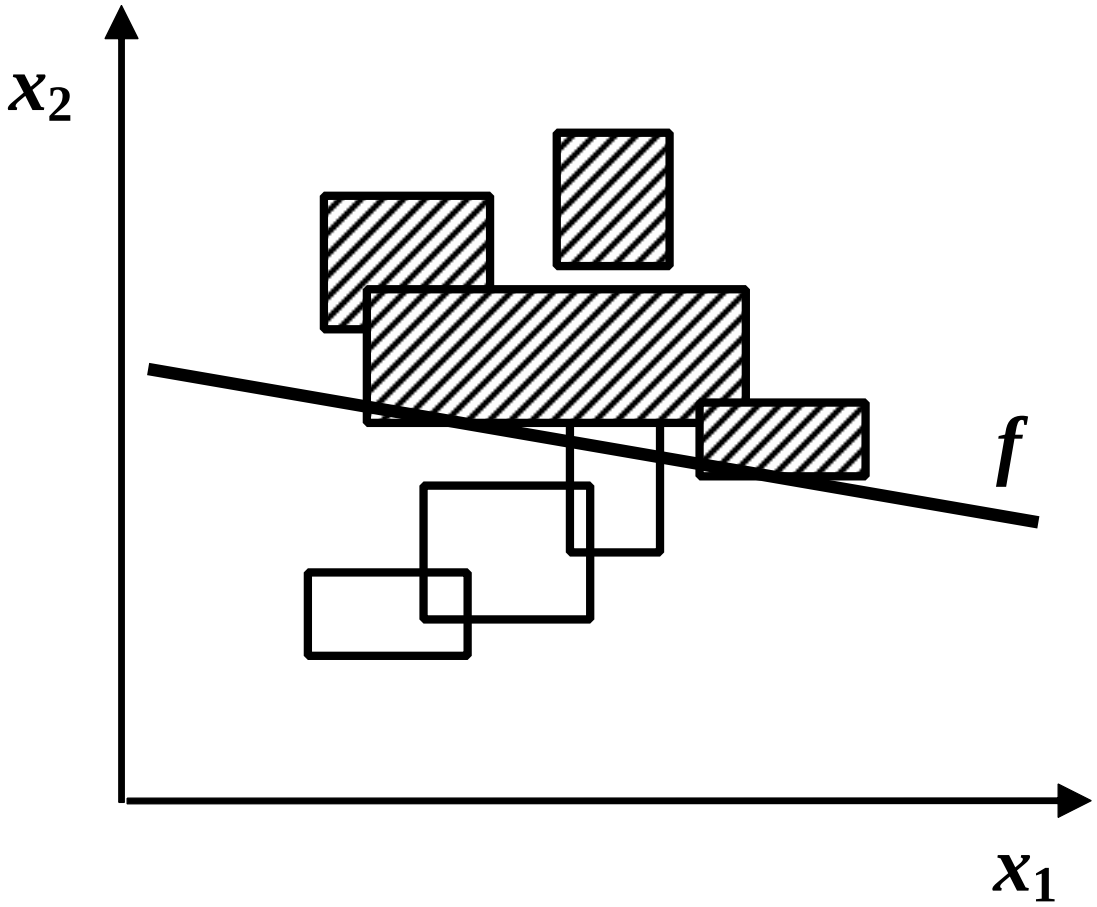
$$\begin{gathered}\xi_i \geq 0, \quad \xi_i^* \geq 0,\\ \xi_i+\varepsilon \geq \left(\langle\alpha\mathbf{x}_i\rangle+\alpha_0\right)-\underline{y}_i, \;\; \xi_i+\varepsilon \geq \left(\langle\alpha\mathbf{x}_i\rangle+\alpha_0\right)-\overline{y}_i,\\ \xi_i^*+\varepsilon \geq \underline{y}_i-\left(\langle\alpha\mathbf{x}_i\rangle+\alpha_0\right), \;\; \xi_i^*+\varepsilon \geq \overline{y}_i-\left(\langle\alpha\mathbf{x}_i\rangle+\alpha_0\right).\end{gathered}$$

$\frac{1}{2}\langle\alpha,\alpha\rangle$ is the Tikhonov regularization term (the most popular penalty or smoothness term)

Advantages of SVMs

1. SVMs are flexible in the choice of the form of the discriminant and regression functions (**non-linear functions f due to kernel methodology**);
2. SVMs provide a unique solution (due to convex objective function), there are no false local minima;
3. SVMs are simple to use;
4. SVMs have a clear geometric explanation.

Interval data in classification



What do the minimax and minimin strategies mean?

1. Regression:
   - minimax: outlying points are taken into account;
   - minimin: neighboring points are taken into account.
2. Classification:
   - minimax: points from two classes approach each other (get mixed);
   - minimin: points are separated.

Questions

?