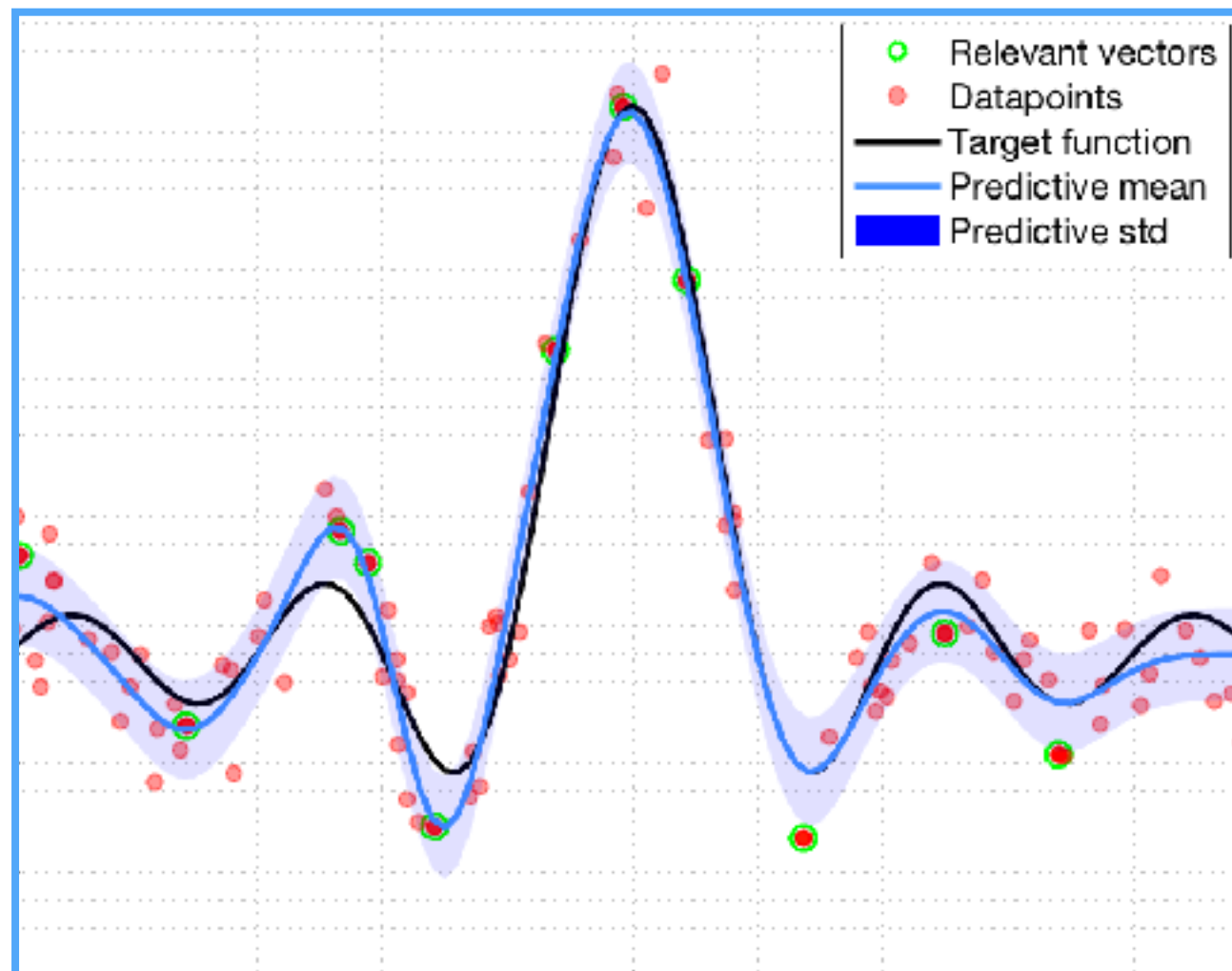


Support Vector Regression vs. Relevance Vector Regression a sparsity / performance study



L.Faury

G.Gallois-Montbrun

H.Hendrikx

26/05/2017

Outline

- ▶ **Theoretical** reminders on both methods
- ▶ Datasets presentation & test runs
- ▶ Introduction to a **sparse-regression** metric, experimental justification
- ▶ Sparse-regression metric based **cross-validation**
- ▶ Comparison outcomes

■ Regression

Learn $f : \mathbb{R}^d \rightarrow \mathbb{R}$ thanks to a dataset $\{X, t\} \in (\mathbb{R}^d)^n \times \mathbb{R}^n$

Assuming a Gaussian **conditional p.d.f** around a linear transformation of features :

$$p(t \mid x, w) = \mathcal{N}(t \mid w^T \phi(x), \beta^{-1})$$

the maximum-likelihood estimator (MLE) writes :

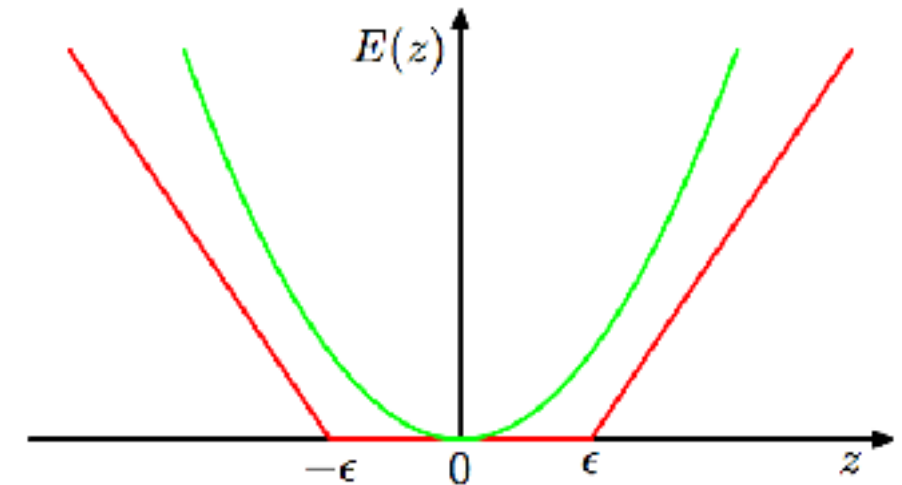
$$\begin{aligned} \hat{w} &= \operatorname{argmax}_w p(t \mid X, w) \\ &= \operatorname{argmin}_w \frac{1}{2} \sum_{i=1}^n \|w^T \phi(x_i) - t_i\|^2 \end{aligned}$$

■ Support Vector Regression

- Introduce the ε -insensitive loss-function.

Equivalent to the QP program :

$$\begin{aligned} \min_w \quad & \frac{C}{n} \sum_n (\xi_n + \hat{\xi}_n) + \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & \begin{cases} \xi, \hat{\xi} \geq 0 \\ w^T \phi(x_n) + \xi_n + \varepsilon \geq t_n \\ w^T \phi(x_n) - \hat{\xi}_n - \varepsilon \leq t_n \end{cases} \end{aligned}$$



Source : Bishop, *Pattern Recognition and Machine Learning* (2006)

- Inactive constraints leads to a sparse model. Only points outside the ε -tube are used for predictions :

$$y(x) = \sum_{n \in \mathcal{S}} (a_n - \hat{a}_n) k(x, x_n) \quad \longrightarrow \quad \text{Posterior decision}$$

■ Relevance Vector Regression

- Provide the predictor with a Gaussian prior : $w \sim \prod_i \mathcal{N}(w_i | 0, \alpha_i^{-1})$

$$y(x) = \sum_n w_n k(x, x_n)$$

- Use **type-2 likelihood** (*evidence approximation*) to determine :

$$(\alpha^*, \beta^*) = \operatorname{argmax}_{\alpha, \beta} [p(t | \alpha, \beta) = \int_w p(t | w, \beta) p(w | \alpha)]$$

- Automatic Relevance Detection : drives some α_i to $+\infty$ (sparse model). Others are called **relevant** vectors.
- Compute posterior and **predictive distribution**

■ Comparison

SVR

- ▶ Decision choices
- ▶ Held-out method for hyper-parameters (at least 3)
- ▶ Mercer kernel
- ▶ **Training** : SMO (efficient : somewhere between linear and quadratic)
- ▶ **Testing** : linear in the SV

RVR

- ▶ Posterior probability
- ▶ Hyper-parameters are determined automatically (except kernel)
- ▶ Arbitrary base functions
- ▶ **Training** : cubic in the datapoint
- ▶ **Testing** : linear in the RV