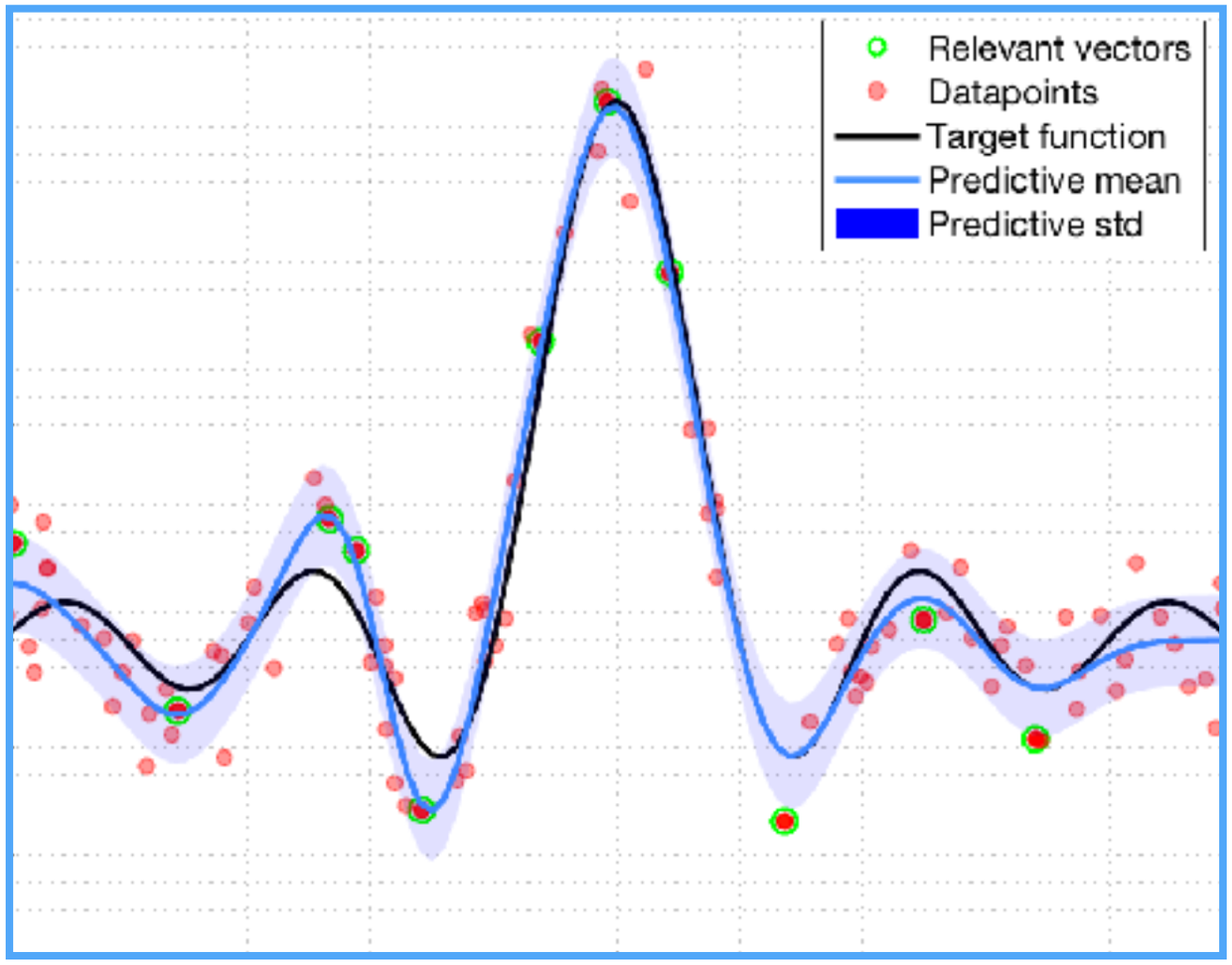
Support Vector Regression vs. Relevance Vector Regression

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 $f : \mathbb{R}^d \to \mathbb{R}$ 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 :

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

the maximum-likelihood estimator (MLE) writes :

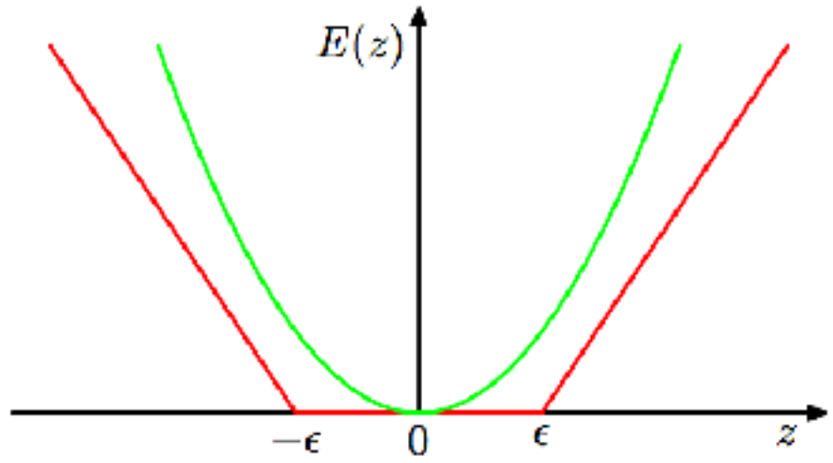
$$\begin{aligned} \hat{w} &= \text{argmax}_w \, p(t \,|\, X, w) \\ &= \text{argmin}_w \, \frac{1}{2} \sum_{i=1}^{n} \| w^t \phi(x) - t \|^2 \end{aligned}$$

Support Vector Regression

- Introduce the ε-insensitive[1] loss-function.
  Equivalent to the QP program :

$$min_w \frac{C}{n}\sum_n(\xi_n+\hat{\xi}_n)+\frac{1}{2}\|w\|^2$$

$$\text{s.t} \begin{cases} \xi,\hat{\xi}\geq 0 \\ w^T\phi(x_n)+\xi_n+\varepsilon \geq t_n \\ w^T\phi(x_n)-\hat{\xi}_n-\varepsilon \leq t_n \end{cases}$$



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 :

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

[1] *Vladimir Vapnik,* The nature of statistical learning theory, *1995*

Relevance Vector Regression[2]

- 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 :

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

- Automatic Relevance Detection : drives some α_i to $+\infty$ (sparse model).

Others are called **relevant** vectors.

- Compute posterior and **predictive distribution**

$$p(t \,|x, X, \beta^*, \alpha^*) = \int_w p(t \,|\, x, w, \beta^*) p(w \,|\, X, \alpha^*) dw$$

[2] *Tipping Michael,* Sparse Bayesian learning and the relevance vector machine, *Journal of machine learning research, 2001*

Comparison

SVR	RVR
Decision choices	Posterior probability
Held-out method for hyper-parameters (at least 3)	Hyper-parameters are determined automatically (except kernel)
Mercer kernel	Arbitrary base functions
Training : SMO[3] (somewhere between linear and quadratic)	**Training** : cubic complexity
Testing : linear in the SV	**Testing** : linear in the RV

[3] *John Platt,* Sequential minimal optimization : A fast algorithm fort raining support vector machines. *1998*

Dataset presentation