Learning with Support Vector Machines

Vrije Universiteit Brussel
Arnout Van Messem
avmessem@vub.ac.be

Given: training data $\{(\mathbf{x}_i, y_i) \mid i = 1 \dots n\} \in X \times Y$, $X \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}$, (X_i, Y_i) i.i.d. $\sim \mathbb{P}$ unknown
Goal: minimize the regularized risk

$$f_{\mathbb{P},\lambda} = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{\mathbb{P}} L(Y, f(X)) + \lambda \|f\|_{\mathcal{H}}^2$$

1. Loss function

Difference between Y and $f(X)$

$L : Y \times \mathbb{R} \to [0, \infty)$

Convex $\Rightarrow$ problem not NP-hard

2. Kernel

Difference between $\mathbf{x}_i$ and $\mathbf{x}_j$

$X \neq \emptyset$, $\mathcal{H}$ reproducing kernel Hilbert space over X

$k : X \times X \to \mathbb{R}$ is **reproducing kernel** of $\mathcal{H}$ if

- $\forall \mathbf{x} \in X$: $k(\mathbf{x}, \cdot) \in \mathcal{H}$ and
- $\forall f \in \mathcal{H}$, $\forall \mathbf{x} \in X$: $f(\mathbf{x}) = \langle f, k(\mathbf{x}, \cdot) \rangle$

Linear: $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
Gaussian RBF: $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma^{-2} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$

3. Penalizing term

Avoids overfitting

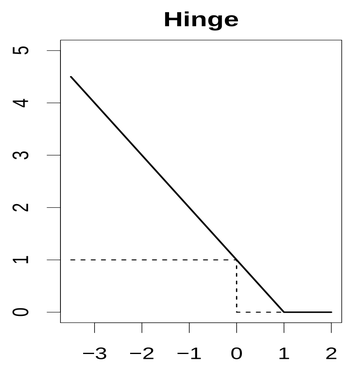
Improves generalization

$\|f\|_{\mathcal{H}}$ also possible

Classification

$Y = \{-1, 1\}$

Loss function

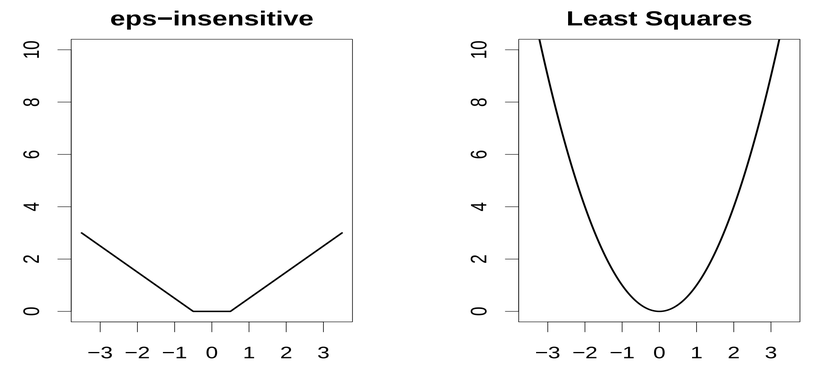


$L_{Hinge}(y, f(\mathbf{x})) = \max\{0, 1 - yf(\mathbf{x})\}$ $L_{Log}(y, f(\mathbf{x})) = \ln(1 + \exp(-yf(\mathbf{x})))$

Regression

$Y \subseteq \mathbb{R}$

Loss function



$L_\varepsilon(y, f(\mathbf{x})) = \max\{0, |y - f(\mathbf{x})| - \varepsilon\}$ $L_{LS}(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$

Remark: $L = L_{LS}$, k linear, $\lambda = 0 \Rightarrow$ standard LS-regression.

Lagrange approach gives the following dual problems

Classification

with hinge loss

$$\arg\min \tfrac{1}{2}\alpha' Q \alpha - \alpha' \mathbf{1}$$

s.t. $\Sigma_i \alpha_i y_i = 0$; $\alpha_i \in [0, C]$; $C > 0$; where $Q_{ij} = y_i y_j k(x_i, x_j)$

Regression

with ε-insensitive loss

$$\min W(\alpha, \alpha^*) = \varepsilon \Sigma_i (\alpha_i + \alpha_i^*) - \Sigma_i (\alpha_i - \alpha_i^*) y_i + \tfrac{1}{2} \Sigma_i \Sigma_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i, x_j)$$

s.t. $\Sigma_i (\alpha_i - \alpha_i^*) = 0$; $\alpha_i, \alpha_i^* \in [0, C]$; $\varepsilon, C > 0$

KKT conditions: necessary and sufficient conditions for a viable solution

Questions

Which properties should L, k, $\mathcal{H}$, and $\lambda = (\lambda_n)_{n \in \mathbb{N}}$ have such that

- $f_{\mathbb{D}_n, \lambda_n}$ can be efficiently computed, with $\mathbb{D}_n$ the empirical distribution
- $\mathbb{E}_{\mathbb{P}} L(Y, f_{\mathbb{D}_n, \lambda_n}(\mathbf{x})) \xrightarrow{\mathbb{P}} \mathcal{R}^*_{L,\mathbb{P}} = \inf_{f: X \to \mathbb{R} \text{ measurable}} \mathbb{E}_{\mathbb{P}} L(Y, f(X))$
- $f_{\mathbb{P},\lambda}$ has good robustness properties

$\longrightarrow$ Cristianini and Shawe-Taylor (2000)
Schölkopf and Smola (2002)
Christmann and Steinwart (2004, 2005)

References

[1] Burges C. (1998), *A tutorial on support vector machines for pattern recognition*, Data Mining and Knowledge Discovery, **2**, 121–167.

[2] Christmann A. and Steinwart I. (2004), *On robustness properties of convex risk minimization methods for pattern recognition*, Journal of Machine Learning Research, **5**, 1007–1034.

[3] Christmann A. and Steinwart I. (2005), *Consistency and robustness of kernel based regression*, University of Dortmund, SFB–475, TR–01/05, Submitted.

[4] Cristianini N. and Shawe-Taylor J. (2000), *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, UK.

[5] Schölkopf B. and Smola A. (2002), *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, Massachusetts.

[6] Suykens J. et al. (2002), *Least Squares Support Vector Machines*, World Scientific, Singapore.

[7] Vapnik V. (1998), *Statistical Learning Theory*, Wiley, New York.