Handling Uncertainties in SVM Classification

Emilie Niaf, Rémi Flamary, Carole Lartizien, and Stéphane Canu

emilie.niaf@creatis.insa-lyon.fr, remi.flamary@insa-rouen.fr

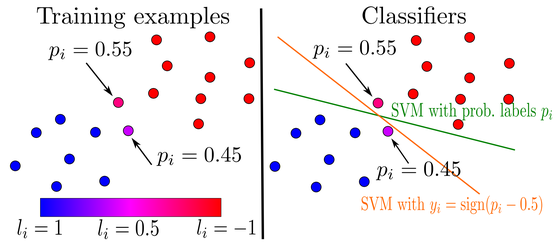
Contribution: SVM extension to probabilistic labels

We address the pattern classification problem arising when available target data include some uncertainties. Suppose that target data is either qualitative (a class label) or quantitative (a probability). We propose a SVM inspired formulation of this problem allowing to take into account class label through a hinge loss as well as probability estimates using ε-insensitive cost function together with a minimum norm (maximum margin) objective. The solution provided can be used for both decision and posterior probability estimation.

Problem formulation

Let X be a feature space and $(x_i, l_i)_{i=1\dots m}$ the learning dataset of input vectors $(x_i)_{i=1\dots m} \in X$ along with their corresponding labels $(l_i)_{i=1\dots m}$, such that

- **class labels**: $l_i{=}y_i \in \{-1, +1\}$ for $i = 1 \dots n$ (in classification),
- **real values**: $l_i{=}p_i = \mathbb{P}(Y_i = 1 \mid X_i = x_i) \in$ [0,1] for $i = n+1 \dots m$ (in regression).



Problem solution

Let k be a positive kernel satisfying Mercer's condition and H the associated Reproducing Kernel Hilbert Space. The associated P-SVM (probabilistic) pattern recognition problem is

$$\min_{f,b,\xi,\xi^-,\xi^+} \frac{1}{2}\|f\|_H^2 + C\sum_{i=1}^{n}\xi_i + \tilde{C}\sum_{i=n+1}^{m}(\xi_i^- + \xi_i^+)$$

subject to

$$\begin{cases} y_i(f(x_i)+b) \geq 1-\xi_i, & i = 1...n \\ z_i^- - \xi_i^- \leq f(x_i)+b \leq z_i^+ + \xi_i^+, & i = n+1...m \\ 0 \leq \xi_i, & i = 1...n \\ 0 \leq \xi_i^- \text{ and } 0 \leq \xi_i^+ & i = n+1...m \end{cases}$$

Following the idea of soft margin introduced in regular C-SVM, slack variables ξ_i measure the degree of misclassification of the datum x_i. C and $\tilde{C} \in \mathbb{R}^*$ control the relative weighting of classification and regression performances. Let ε be the labelling precision, δ be the confidence in the labelling and $\eta = \varepsilon + \delta$. The regression problem consists in finding optimal f such that

$$\left| \frac{1}{1+e^{-a(f(x_i)+b)}} - p_i \right| < \eta \ ,$$

Thus constraining the probability prediction for point x_i to remain around to $\frac{1}{1+e^{-a(f(x_i)+b)}}$ within distance η [1, 2, 3]. This leads to $z_i^- = -\frac{1}{a}\ln(\frac{1}{p_i-\eta} - 1)$ and $z_i^+ = -\frac{1}{a}\ln(\frac{1}{p_i+\eta} - 1)$.

DUAL FORMULATION

Lagrange multipliers allow to rewrite the problem in its dual form $\begin{cases} \min\limits_{\Gamma} \ \frac{1}{2}\Gamma^\top G\Gamma - \tilde{e}^\top\Gamma \\ f^\top\Gamma = 0 \end{cases}$, with

$$f^\top = [y^\top, \underbrace{-1\cdots-1}_{\text{n-m times}}, \underbrace{1\dots1}_{\text{n-m times}}] \ , \qquad 0 \leq \Gamma \leq [\underbrace{C\dots C}_{\text{n times}} \ \underbrace{\tilde{C}\dots\tilde{C}}_{\text{n-m times}} \ \underbrace{\tilde{C}\dots\tilde{C}}_{\text{n-m times}}]^\top$$

$$\tilde{e} = [\underbrace{1\dots1}_{\text{n times}} \underbrace{-z_{n+1}^+\cdots-z_m^+}_{\text{n-m times}} \underbrace{z_{n+1}^-\dots z_m^-}_{\text{n-m times}}] \ , \qquad G = \begin{pmatrix} K_1 & -K_2 & K_2 \\ -K_2^\top & K_3 & -K_3 \\ K_2^\top & -K_3 & K_3 \end{pmatrix}$$

with:

$$K_1 = (y_i y_j k(x_i, x_j))_{i,j=1\dots n}$$
$$K_2 = (k(x_i, x_j) y_i)_{i=1\dots n, j=n+1\dots m}$$
$$K_3 = (k(x_i, x_j))_{i,j=n+1\dots m}$$

This formulation is similar to the one in classical SVM, hence we can benefit from the current solvers

Examples

Numerical examples implementation is based on the SVM-KM Toolbox [4]. We use a RBF kernel with $C = \tilde{C} = 100$. We compare the classification performances and probabilistic predictions of the C-SVM and P-SVM approaches. In the 1^{st} case, probabilities are estimated by using Platt's scaling algorithm [5] while in the 2^{nd} case, probabilities are directly estimated as $P(y = 1|x) = \frac{1}{1+e^{-a(f(x)+b)}}$. Performances are evaluated by computing Accuracy (Acc) and Kullback Leibler distance (KL)

Probability estimation

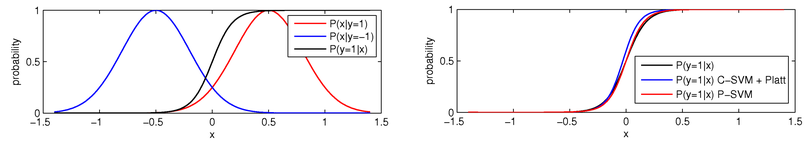
We generate two $\mathcal{N}(\mu, \sigma)$ unidimensional datasets, labelled '+1' and '-1' ($\sigma_{-1}^2{=}\ \sigma_1^2{=}0.3$, $\mu_{-1}{=}$-0.5 and $\mu_1{=}$+0.5). Let $(x_i^l)_{i=1\dots n^l}$ be the learning data set ($n^l{=}200$), $(x_i^t)_{i=1\dots n^t}$ the test set ($n^t{=}1000$). We compute the true probability $P(y_i = +1|x_i)$ for x_i to belong to class '+1'. Learning data are labelled in two ways:

a) 1^{st} dataset $(x_i^l, y_i^l)_{i=1\dots n^l}$ is used to train the C-SVM classifier. For $i = 1 \dots n^l$,

if $P(y_i^l = 1|x_i^l) > 0.5$, then $y_i^l = 1$,
if $P(y_i^l = 1|x_i^l) \leq 0.5$, then $y_i^l = -1$

b) 2^{nd} data set $(x_i^l, \hat{y}_i^l)_{i=1\dots n^l}$ is used to train the P-SVM algorithm. For $i = 1 \dots n^l$,

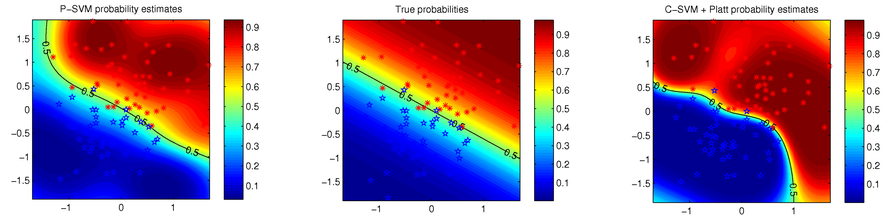
if $P(y_i^l = 1|x_i^l) > 1-\eta$, then $\hat{y}_i^l = 1$,
if $P(y_i^l = 1|x_i^l) < \eta$, then $\hat{y}_i^l = -1$,
otherwise $\hat{y}_i^l = P(y_i^l = 1|x_i^l)$.



True test data probabilities (black) and P-SVM estimations (red) are quasi-superimpoed (KL=0.2) whereas Platt's estimations are less accurate (KL=11.3).

Noise robustness

We generate two $\mathcal{N}(\mu, \sigma)$ 2D datasets ($\sigma_{-1}^2{=}\sigma_1^2{=}0.7$, $\mu_{-1}{=}$(-0.3, -0.5), $\mu_1{=}$(+0.3, +0.5)). We compute class '1' membership probability $P(y_i = 1|x_i)$ for each x_i^l of the learning data set. To simulate classification error, we artificially add a uniform noise (amplitude 0.1), to probabilities, such that for $i = 1 \dots n$, $\hat{P}(y_i = 1|x_i) = P(y_i = 1|x_i) + \delta_i$. We label learning data following the same scheme as described above.



Probability estimations of C-SVM and P-SVM over a grid using noisy learning data (uniform noise, amplitude 0.1). Noisy learning data are plotted in blue (class '-1') and red (class '1') stars.

Far from learning data points, both probability estimations are less accurate, this being directly linked to the choice of a gaussian kernel. However, P-SVM classification and probability estimations obtained for 1000 test points, are clearly more alike the ground truth ($\text{Acc}_{\text{P-SVM}} = 99\%$, $\text{KL}_{\text{P-SVM}} = 3.6$) than C-SVM ($\text{Acc}_{\text{C-SVM}} = 95\%$, $\text{KL}_{\text{C-SVM}} = 95$). C-SVM is sensitive to classification noise and is no more converging to the Bayes rule as seen in [6].

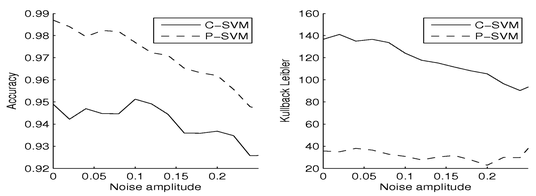


Figure shows the impact of noise amplitude on P-SVM and C-SVM classification performances (values are averaged over 30 random simulations). Even if noise increases, classifications and probability predictions performances of the P-SVM remain significantly higher than those of C-SVM.

Conclusion

Experimental results show that P-SVM can perform very well on simulated data for both discrimination and posterior probability estimation. This approach will be applied on clinical data to assess its usefulness in CAD for prostate cancer. This framework can also be generalized to other dataset involving quantitative data (e.g. to estimate a conditional cumulative distribution function).

References

[1] S. Rüping, *A Simple Method For Estimating Conditional Probabilities For SVMs,* In *LWA*, 2004.

[2] Y. Grandvalet, *A probabilistic interpretation of SVMs with an application to unbalanced classification,* In *NIPS*, 2006.

[3] S. Rueping, *SVM Classifier Estimation from Group Probabilities,* In *ICML*, 2010.

[4] S. Canu, SVM and kernel methods matlab toolbox, 2005.

[5] John C. Platt, *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,* In *MIT Press*, 1999.

[6] G. Stempfel, *Learning SVMs from Sloppily Labeled Data,* In *ICANN*, 2009.