Support Vector Machine for Classification and Regression

Ahlame Douzal
AMA-LIG, Université Joseph Fourier

Master 2R - MOSIG (2013)

November 25, 2013

Outline

1. Loss function, Separating Hyperplanes, Canonical Hyperplan
2. Hard, Soft and ν SVM
3. Multi-class SVM
4. ϵ-sensitive and ν SVR
5. Kernels and temporal kernels

Classifiers, Loss function

For binary classification

- Training Data: $(x_1, y_1), ..., (x_m, y_m) \in X \times \{\pm 1\}$
- Objective
  - To find a function f that will correctly classify unseen examples (x, y), $f : X \to \pm 1$

Classifiers, Loss function

Correctness is measured by means of the error risk, composed of:

- Empirical risk (estimated on the training set)

$$R_{emp} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}|f(x_i) - y_i|$$

- For the zero-one loss function:

$$c(x, y, f(x)) = \frac{1}{2}|f(x) - y|$$

the loss is 0 if (x, y) is classified correctly, 1 otherwise

- Even if $R_{emp}[f]$ is zero on the training set, it may not generalize well on unseen data

Classifiers, Loss function

- Error Risk (on new unknown observations)

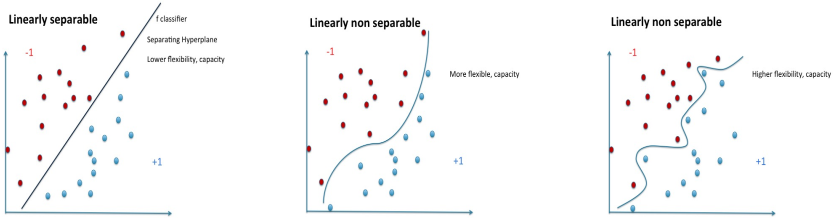
$$R[f] \;=\; \int \frac{1}{2}|f(x) - y| \, dP(x, y)$$

- P(x,y) generally unknown distribution,
- the problem remains to bound $R[f]$ (structural risk minimization)

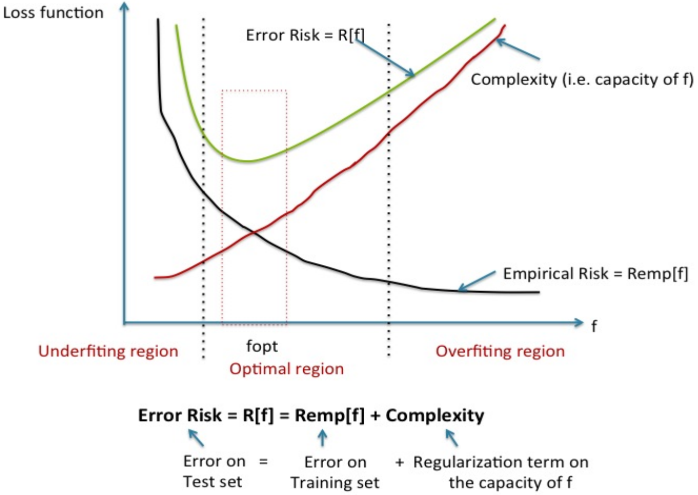
Classifiers, Loss function

- Complexity

- It measures the capacity of a family of classifiers to isolate ("shatter") observations
- VC-theory shows the need to restrict the set of functions f to the one that have suitable complexity for the amount of training data
-For example, capacity of LDA $<$ capacity of QDA



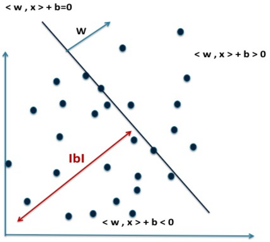
Classifiers, Loss function



Hyperplanes

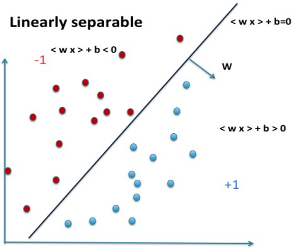
H a dot vectorial space $<,>$
$x_1, \ldots x_m$ m points of H
An hyperplan HP is defined:

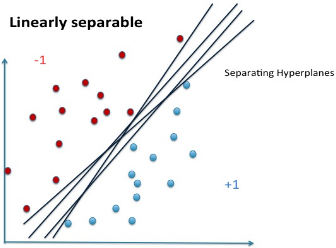
$$\{x \in H \;/\; < w, x > +b = 0\} \; w \in H, b \in \mathbb{R}$$



Separating Hyperplanes

- Binary classification
- Linearly separable points $x_1, ...x_m$ of H



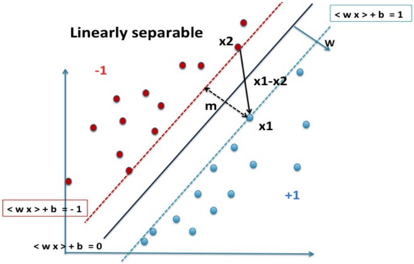


Canonical Hyperplan

Definition

The pair (w,b) is called a canonical hyperplan w.r.t. $x_1, ..., x_m \in H$, if it is scaled such that

$$\min_{i=1...m} |< w, x_i > +b| = 1 \quad (1)$$



Canonical Hyperplan

Let Hp_0, Hp_{+1} and Hp_{-1} be the three hyperplans as indicated in the above figure
Let x_1, x_2 be the closest points to Hp_0 (see Fig), then

$$< w, x_1 > + b = c > 0$$
$$< w, x_2 > + b = -c < 0$$

multiply each equations by a scale factor $\alpha = \frac{1}{c}$, thus

$$\begin{aligned} \alpha < w, x_1 > + \alpha\, b &= < w', x_1' > + b' = 1 \\ \alpha < w, x_2 > + \alpha\, b &= < w', x_2' > + b' = -1 \end{aligned}$$

Canonical Hyperplan

Margin value

- The closest point to the hyperplan has a distance of $\frac{1}{\|w\|}$

$$
\begin{aligned}
< w, x_1 > +b &= 1 & (2)\\
< w, x_2 > +b &= -1 & (3)\\
\text{from (2)-(3)} & \quad < w, (x_1 - x_2) >= 2 > \text{ and } \|w\| \, \|(x_1 - x_2)\| cos(\alpha) = 2 & (4)
\end{aligned}
$$

The margin $m = \|x_1 - x_2\| cos(\alpha)$ defines the orthogonal projection onto a line of w direction:

$$m = \frac{2}{\|w\|}, \text{and the closest point has a distance of } \frac{1}{\|w\|}$$

Remark: To best separate the classes, the problem becomes to determine the hyperplan that maximizes the margine m (i.e. minimizes $\frac{\|w\|}{2}$)

Support Vector Machine

- Let $(x_1, y_1), ..., (x_m, y_m)$ be m points, $x_i \in H$
- Assume a binary classification of **linearly separable** points (non separable to see later)
- Let HP be a separable hyperplan of direction w
- The trick: $y_i = +1$ (vs. $y_i = -1$) for points belonging to the side of direction w (vs. opposite direction to w)
- The decision function $f_{w,b}$ that gives the class label of a given x

$$f_{w,b}(x) = sign(< w, x > +b) = \{+1/or - 1\}$$

Support Vector Machine

SVM: Primal formalisation

- Among the set of separating hyperplans, the optimal *HP* is the one that maximizes the margin
- The problem can be formalized as a convex (unique solution) and quadratic optimization problem s.t. linear inequalities

$$\begin{aligned} \min_{w \in H, b \in \mathbb{R}} \quad & \frac{1}{2}\|w\| \\ s.t. \quad & y_i(< x_i, w > +b) \geq 1 \quad \forall i = 1, ..., m \end{aligned} \tag{5}$$

The associated Lagrangian $\mathcal{L}$ to minimize w.r.t. w and b, to maximize w.r.t. α_i

$$\mathcal{L}(w, b, \alpha) \quad = \quad \frac{1}{2}\|w\| - \sum_{i=1}^{m} \alpha_i(y_i(< x_i, w > +b) - 1) \tag{6}$$

Support Vector Machine

The derivatives $\frac{\partial \mathcal{L}}{b}$ and $\frac{\partial \mathcal{L}}{w}$ leads to

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \quad w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{7}$$

- $\forall x_i$ with $\alpha_i > 0$,
  - x_i define a support vector
  - x_i contributes to define the optimal plan
  - x_i involves on the canonical hyperplans
  - x_i contributes for the decision function
- $\forall x_i$ with $\alpha_i = 0$
  - x_i not considered for the decision function (sparsity)

Note that:

$$\forall\ i \in \{1, ..., m\} \quad \alpha_i \left(y_i \left(< x_i, w > +b\right) - 1\right) = 0$$

Support Vector Machine: Dual formalization

By substituting and replacing equations (7) in the Lagrangian given in (6) we obtain the SVM Dual formalization

$$
\begin{aligned}
\max_{\alpha \in \mathbb{R}^{\boldsymbol{m}}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \, \alpha_j \, y_i \, y_j \, < x_i, x_j > \\
s.t. \quad & \alpha_i \geq 0 \; , i = 1, ..., m \\
& \sum_{i=1}^{m} \alpha_i \, y_i = 0
\end{aligned}
\tag{8}
$$

The decision function

$$
f(x) \quad = \quad \text{sign} \left(\sum_{i=1}^{m} \alpha_i \, y_i \, < x, x_i > + b \right) \tag{9}
$$

For x_i limited to the support vectors.

Soft-Margin vs. Hard-margin SVM

- If non linearly separable data, there is no hard-margin solution
- Either linearly separable, hard-margin suffers of over fitting ($R_{Emp} = 0$) and worst generalization properties (high risk R)
- To ensure good generalization properties with lower R, one needs to find a larger margin and tolerate some samples to be within the margin or either miss-classified
- A regularization is thus used to relax on the empirical risk but by improving the generalization risk $R = R_{emp}$ + complexity
- For this, slack variables ξ_i are introduced to formalize the soft-margin SVM.

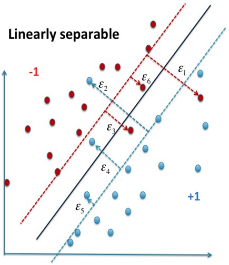
Soft-margin SVM

Primal formalization

$$\min_{w \in H, \xi \in \mathbb{R}^m, b \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m}\xi_i \tag{10}$$

$$\text{s.t.} \quad y_i(< x_i, w > +b) \geq 1 - \xi_i \quad \forall i = 1, ..., m$$

$$\xi_i \geq 0 \quad \forall i = 1, ..., m$$



Soft-margin SVM

$$
\begin{aligned}
\min_{w\in H,\xi\in\mathbb{R}^{\boldsymbol{m}},b\in\mathbb{R}} \quad & \frac{1}{2}\|w\|^2 + C\,\frac{1}{m}\sum_{i=1}^{m}\xi_i \\
s.t. \quad & y_i(< x_i, w > +b) \geq 1-\xi_i \quad \forall i=1,\dots,m \\
& \xi_i \geq 0 \quad \forall i = 1,\dots,m
\end{aligned}
$$

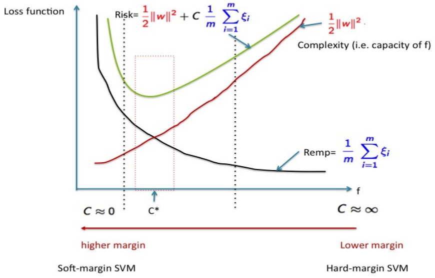
Some intuitions (1)

- $\forall\ x_i$ that is far from the margin and lying in the good side, the 2^{nd} constraint is always satisfied as $y_i\,(< x_i, w > +b) \geq 1$ and ξ_i which is not needed is set to 0 to minimize Eq. (11).
- $\forall\ x_i$ which is within the margin or lies in the wrong side, the constraint $y_i\,(< x_i, w > +b) \geq 1$ is violated, and $\xi_i > 0$ is involved to have a solution.

Soft-margin SVM

Some intuitions(2)

- The right term, called the hing-loss, measures the empirical risk induced by all the samples with $\xi_i > 0$
- The left term, called the regularization term, measures the complexity or the capacity of the model.
- The decrease of the left term, increases the margin, that decreases the capacity of the model and increases the hing-loss
- The minimization problem is a compromise, balanced by C, between the two left (complexity) / right (empirical risk) conflicting terms



Soft-marginSVM: Dual formalization

$$
\begin{aligned}
\max_{\alpha \in \mathbb{R}^m} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \, \alpha_j \, y_i \, y_j \; < x_i, x_j > \\
s.t. \quad & 0 \leq \alpha_i \leq \frac{C}{m} \, , i = 1, ..., m \\
& \sum_{i=1}^{m} \alpha_i \, y_i = 0
\end{aligned} \tag{11}
$$

Remarks:

- The constraint $\alpha_i \leq \frac{C}{m}$ ensures to bound the weight of a given support vector, to avoid over fitting, or that an outlier support vector takes too much importance in the decision function

ν-SVM

Some intuitions

- The parameter C in the soft margin-SVM is a compromise between the conflicting terms complexity and empirical risk
- Unfortunately we have no intuition about the meaning of C w.r.t. the data
- ν-SVM allows to substitute C by the parameter ν related to:
  - The number of errors
  - The number of support vectors

Primal formalization

$$\min_{w \in H, \xi \in \mathbb{R}^{\boldsymbol{m}}, \boldsymbol{b} \in \mathbb{R}, \varphi \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 - \nu\,\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i \tag{12}$$

$$\begin{aligned} s.t. \quad & y_i(< x_i, w > +b) \geq \rho - \xi_i \quad \forall i = 1, ..., m \\ & \xi_i \geq 0 \quad \forall i = 1, ..., m \\ & \rho \geq 0 \end{aligned}$$

ν-SVM

Interpretation of ρ

1. The classes are separated by a margin of $\frac{2\rho}{\|w\|^2}$
2. $\rho \in [0, 1]$ is a upper bound of the proportion of samples lying within the margin or in the wrong side (called the fraction of margin errors)
3. ρ is a lower bound of the proportion of support vectors

$$\text{The fraction of margin errors} \leq \rho \leq \text{The fraction of support vectors}$$

Remarks:

- The upper bound controls the sparsity (minimal number of support vectors)
- The lower bound controls the model precision (namely the maximal margin errors)
- The increase of ρ increases the margin, that allows the increase of the margin errors

ν-SVM

Dual formalization

$$
\begin{aligned}
\max_{\alpha \in \mathbb{R}^{m}} \quad & -\frac{1}{2} \sum_{i,j} \alpha_i \; \alpha_j \; y_i \; y_j \; < x_i, x_j > \\
s.t. \quad & 0 \leq \alpha_i \leq \frac{1}{m} \,, i = 1, ..., m \\
& \sum_{i=1}^{m} \alpha_i \; y_i = 0 \\
& \sum_{i=1}^{m} \alpha_i \geq 0
\end{aligned} \tag{13}
$$

Multi-class SVM

Let $S = \{(x_i, y_i) \;\; i = 1, ...m\}$, $y_i \in \{1, ..., K\}$. Two main approaches exist to deal with SMV on multi-classes.

1- One versus all approach

1. Generate K training sets $S_1, ..., S_K$:

$$S_k = \{(x_i, y_i^k) \; i = 1, ..., m\}$$
$$y_i^k = +1 \;\; if \;\; y_i = k \qquad y_i^k = -1 \;\; if \;\; y_i \neq k$$

2. For each training set S_k learn a binary SVM, with

$$g^k(x) = \sum_i^m \alpha_i \; y_i < x_i, x > +b$$
$$f^k(x) = sign(g^k(x)) \quad \text{the decision function}$$

3. Classification of a new sample x^*
   - Estimate $g^j(x^*) = max(g^1(x^*), ..., g^K(x^*))$
   - The class label is given by $f(x^*) = sign(g^j(x^*))$

Multi-class SVM: One versus all approach

Remarks

- For $g^j(x^*) > 0$, assign x^* to the *jth* class, otherwise the only decision is that x^* is not in the *jth* class
- Some samples may not be classified (for instance, $g^j(x^*) < 0$, many nearest maximal values for g)
- The K SVM's are trained on different sets $(S_1, ..., S_K)$ with functions $g^1, ..., g^K$ varying within different variation domains (non comparable), not suitable use of the max on the decision function
- Unbalanced classes in the training sets $(S_1, ..., S_K)$ small size for $+1$ larger for -1

Multi-class SVM: pairwise approach

2- Pairwise approach

1. Generate $K(K-1)$ Training sets for each couple of classes S_i, S_j
2. Learn a binary SVM per couple of classes, with g_{ij} the learned decision function
3. Assign a new sample x^* by a majority vote through the $k(K-1)$ decision functions $f_{ij}(x^*) = sign(g_{ij}(x^*))$

Remarks

- It leads to much more trained classifiers (limited if a large number of classes)
- The induced classes are expected to be smaller and more balanced
- We expect lower number of support vectors that for the One versus all approach

Support Vector Regression (SVR)

- Rather than dealing with outputs outputs $y = \{\pm 1\}$ in classification, regression estimation is concerned with estimating real-valued functions ($y \in \mathbb{R}$)
- *SVR* generalizes SV algorithm to the regression case
- SVR allows the estimation of the regression function by involving a part of the training (sparsity)
- The regression function is rarely linear; however, similarly to SVM, we first give the primal and dual formalizations for the case of a linear regression function, and show after how to extend the results to non linear regression

Support Vector Regression (SVR)

Definition

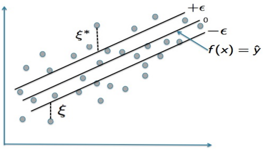
Let (x_i, y_i) $i = 1, ..., m$, $y_i \in \mathbb{R}$, the aim of SVR is the estimation of $\hat{y} = f(x)$ that minimizes the ϵ-insensitive Loss-function R^{ϵ}_{Emp}:

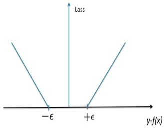
$$R^{\epsilon}_{Emp} = |f(x) - y|_{\epsilon} = max(0, |f(x) - y| - \epsilon)$$

Remarks

- The intuition behind the empirical risk is to be equal to 0 for an estimation error lower than ϵ and $|f(x) - y| - \epsilon$ if it is higher
- Case of Estimating a linear regression function $f(x) =< w, x > +b$
- Similarly, it remains to minimize R^{ϵ}_{Emp}, to not over fit maximize ϵ

Support Vector Regression ($\epsilon - SVR$)





Primal formalization

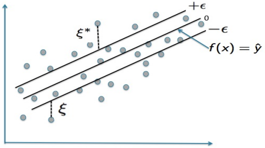
$$\min_{w \in H, \xi^{(*)} \in \mathbb{R}^m, b \in \mathbb{R}, \varphi \in \mathbb{R}} \quad \frac{1}{2}\|w\|^2 + C\frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{14}$$

$$\begin{aligned} s.t. \quad & (< x_i, w > + b) - y_i \leq \epsilon + \xi_i \quad \forall i = 1, ..., m \\ & y_i - (< x_i, w > + b) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \quad \forall i = 1, ..., m \end{aligned} \tag{15}$$

$\epsilon - SVR$: Primal formalization

$$
\begin{aligned}
\min_{w\in H,\xi^{(*)}\in\mathbb{R}^{\boldsymbol{m}},b\in\mathbb{R},\varphi\in\mathbb{R}} \quad & \frac{1}{2}\|w\|^2 \; + \; C\,\frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*) \qquad (16)\\
s.t. \quad & (< x_i, w > +b) - y_i \leq \epsilon + \xi_i \quad \forall i = 1, ..., m\\
& y_i - (< x_i, w > +b) \leq \epsilon + xi_i^*\\
& \xi_i, \xi_i^* \geq 0 \quad \forall i = 1, ..., m
\end{aligned}
$$

- For the samples with y_i above the tube, $\xi_i^* > 0$ ($\xi_i = 0$), samples are underestimated ($f(x_i) < y_i$)
- For the samples with y_i under the tube, $\xi_i > 0$ ($\xi_i^* = 0$), samples are overestimated ($f(x_i) > y_i$)
- For the remaining samples within the tube, $\xi_i^* = \xi_i = 0$, samples are well estimated ($|f(x_i) - y_i| \leq \epsilon$)



$\epsilon - SVR$

Some intuitions

- ϵ defines the margin around $f(x)$: $\epsilon = \frac{1}{2}\|w\|^2$
- Higher is ϵ, lower is $\|w\|^2$, and lower is the precision of the regression model
- Higher is ϵ, smoother is $f(x)$ and lower is the complexity of the model
- Lower is ϵ, less smoothed is $f(x)$, higher is the complexity, but higher is the risk to overfit
- For $\epsilon \approx 0$, the model is a hard linear regression (without a tube ϵ)

$\epsilon - SVR$: Dual formalization

Introducing Lagrange multipliers, on the primal form Eq. (16), one arrives at the following optimization problem (C and ϵ selected *a priori*)

$$\begin{aligned} \max_{\alpha,\alpha^* \in \mathbb{R}^m} \quad & -\epsilon \sum_{i=1}^{m}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)y_i \\ & -\frac{1}{2}\sum_{i,j}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) < x_i, x_j > \\ s.t. \quad & 0 \leq \alpha_i^*, \alpha_i \leq \frac{C}{m} \;\; \forall\, i = 1, ..., m \\ & \sum_{i=1}^{m}(\alpha_i^* - \alpha_i) \end{aligned} \tag{17}$$

The regression estimate

$$\begin{aligned} f(x) &= \sum_{i=1}^{m}(\alpha_i^* - \alpha_i) < x_i, x > + b \\ w &= \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\, x_i \end{aligned} \tag{18}$$

$\epsilon - SVR$: Dual formalization

Remarks

- α_i^* and α_i correspond to the weights of the support vectors that are, respectively, above, under the tube
- The support vectors (SV) are those samples with $\alpha_i^* > 0$ or $\alpha_i > 0$

Computing the Offset b

- To estimate b we refer to referring to the KKT(Karush-Kuhn-Tucker) conditions that state that at the point of the solution, the product between the dual variables and constraints has to vanish

$$\alpha_i(\epsilon + \xi_i - y_i + < w, x_i > +b) = 0 \quad (19)$$

$$\alpha_i(\epsilon + \xi_i^* + y_i - < w, x_i > -b) = 0 \quad (20)$$

$$(\frac{C}{m} - \alpha_i)\xi_i = 0 \quad (\frac{C}{m} - \alpha_i^*)\xi_i^* = 0 \quad (21)$$

$\epsilon - SVR$: Dual formalization

Useful derived conclusions

- Only samples (x_i, y_i) that lie outside the tube have $\alpha_i^{(*)} = \frac{C}{m}$ (as $\xi_i^{(*)} = 0$)
- $\alpha_i \, \alpha_i^* = 0$ (as the $i - th$ SV is either above or under the tube)
- $\alpha_i^{(*)} \in [0, \frac{C}{m}]$, $\xi_i^{(*)} = 0$, that is only for $SV's$ that lie within the tube

Thus the Offset b is,

$$\begin{aligned} b &= y_i - < w, x_i > -\epsilon \;\; for \;\; \alpha_i \in (0, \frac{C}{m}) \\ b &= y_i - < w, x_i > +\epsilon \;\; for \;\; \alpha_i^* \in (0, \frac{C}{m}) \end{aligned}$$

Remark

- This means, that any Lagrange multipliers $\alpha_i^{(*)} \in (0, \frac{C}{m})$ can be used to estimate b, it is safest to use one that is not too close to 0 or $\frac{C}{m}$

$\nu - SVR$

- ϵ of the $\epsilon - SVR$ is usfull if the desired accuracy can be specified beforhand
- In some cases, however, we just one to estimate y to be as accurate as possible without specifying an a priori level of accuracy
- For this, we refer to the $\nu - SVR$ that allows to compute automatically ϵ

Primal formalization

$$\begin{aligned} \min_{w \in H, \xi^{(*)} \in \mathbb{R}^m, b \in \mathbb{R}, \epsilon, b \in \mathbb{R}} \quad & \frac{1}{2}\|w\|^2 + C\left(\nu\epsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*)\right) \\ s.t. \quad & (< x_i, w > +b) - y_i \leq \epsilon + \xi_i \quad \forall i = 1, ..., m \\ & y_i - (< x_i, w > +b) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{aligned} \tag{22}$$

Intuitions

- If ϵ increases, the green term decreases (as less samples outside the tube), the function smoothness increases and the accuracy decreases
- If ϵ decreases, the brown term decreases, but the green term increases (as more samples outside the tube), the function is less smoothed and the the accuracy increases

$\nu - SVR$: Dual formalization

$$
\begin{aligned}
\max_{\alpha,\alpha^* \in \mathbb{R}^m} \quad & \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) y_i - \frac{1}{2} \sum_{i,j}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) < x_i, x_j > \\
s.t. \quad & 0 \leq \alpha_i^*, \alpha_i \leq \frac{C}{m} \;\; \forall\, i = 1, ..., m \\
& \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) \\
& \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) \leq C.\nu
\end{aligned} \tag{23}
$$

The regression estimate

$$
\begin{aligned}
f(x) &= \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) < x_i, x > + b \\
w &= \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\, x_i
\end{aligned} \tag{24}
$$

$\nu - SVR$:

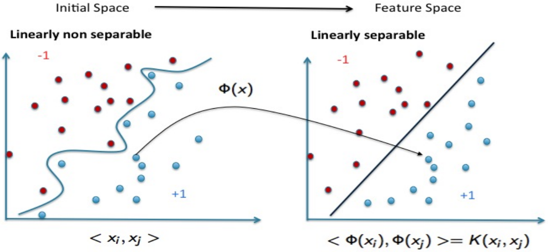
Interpretation of the a priori fixed $\nu \in [0, 1]$

The fraction of samples outside the tube(margin errors) $\leq \nu \leq$ The fraction of support vectors

- The upper bound controls the sparsity (minimal number of support vectors)
- The lower bound controls the model accuracy (namely the maximal margin errors)
- The increase of ν increases the margin, that increases the margin errors
- If ν increases, this allows for more samples outside the tube, appeals for more precision by decreasing ϵ and increasing the number of *SV*
- If ν decreases, this allows less samples outside the tube, it appeals for less precision and more sparsity by increasing ϵ and decreasing the number of *SV*

SVM and SVR: Non linearly separable data

- The above hard, soft, or ν SVM/SVR are developed for the case of linearly separable data
- To deal with non linearly separable data, the trick consists to embed data into high dimension space (called feature space), rendering the data linearly separable and the developed approaches applicable
- This is possible, by substituting all the cross-product used in the results by a kernel similarity measure (kernel trick)



Standard Kernels

- Polynomial: $k(x, x') = < x, x' >^d$
- Gaussian: $k(x, x') = exp(-\frac{\|x-x'\|}{2\sigma^2})$
- Sigmoid: $tanh(\mathcal{K}(x, x') + \Theta)$

with suitable choices of $d \in \mathbb{N}$, $\sigma, \mathcal{K}, \Theta \in \mathbb{R}$ empirically led to SV classifiers with similar accuracies as SV sets

Temporal Kernels

- The Global Alignment K_{GA} kernel (Cuturi et al. 2011) is defined as the exponentiated soft-minimum of all alignment distances:

$$\begin{aligned} DTW &= \min_{\pi \in A(n,m)} D_{x,y}(\pi) \\ D_{x,y} &= \sum_{i=1}^{|\pi|} \varphi(x_{\pi_{\mathbf{1}}(i)}, y_{\pi_{\mathbf{2}}(i)}) \\ K_{GA}(\mathbf{x},\mathbf{y}) &= \sum_{\pi \in A(n,m)} e^{-D_{\mathbf{x},\mathbf{y}}(\pi)} \\ &= \sum_{\pi \in A(n,m)} \prod_{i=1}^{|\pi|} k(x_{\pi_{\mathbf{1}}(i)}, y_{\pi_{\mathbf{2}}(i)}) \end{aligned}$$

where $k = exp^{-\varphi}$ a local similarity induced from the divergence φ

Temporal Kernels

- DTW kernel K_{DTW} (Haasdonk et al. 2004) a pseudo n.d. kernel

$$K_{DTW}(\mathbf{x}, \mathbf{y}) \quad = \quad e^{-\frac{1}{t} DTW(\mathbf{x},\mathbf{y})}$$

- DTW kernel DTW_{sc} with Sakoe-Chiba Constraints

$$DTW_{sc}(\mathbf{x}, \mathbf{y}) \quad = \quad \min_{\pi \in A(n,m)} D^{\gamma}_{\mathbf{x},\mathbf{y}}(\pi)$$

where the wrights $\gamma_{i,j}$ are defined as:

$$\gamma_{i,j} = \begin{array}{l} 1, \ \text{if} |i-j| < T \\ \infty, \ \text{otherwise} \end{array}$$

Temporal Kernels

- Dynamic Temporal Alignement Kernel K_{DTAK} (Shimodaira et al. 2002) consider a variant of the DTW to define the pseudo p.d. kernel

$$DTW_{DTAK}(\mathbf{x}, \mathbf{y}) \quad = \quad \max_{\pi \in A(n,m)} \sum_{i=1}^{|\pi|} k_\sigma(x_{\pi_1(i)}, y_{\pi_2(i)})$$

References

- B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144-152. ACM Press 1992.
- C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
- C. Cortes and V. Vapnik. Support-vector network. Machine Learning, 20:273-297,1995.
- R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classication. Journal of Machine Learning Research, 9:1871-1874, 2008. URL http://www.csie.ntu.edu.tw/ cjlin/papers/liblinear.pdf
- H.-T. Lin and C.-J. Lin. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University, 2003. URL http://www.csie.ntu.edu.tw/ cjlin/papers/tanh.pdf
- B. Schölkopf and A. J. Smola. Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.