Support Vector Machine for Classification and Regression

Ahlame Douzal

AMA-LIG, Université Grenoble Alpes

November 19, 2018

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: $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m) \in X \times \{\pm 1\}$
- Objective
  - To find a function f that will correctly classify unseen examples $\boldsymbol{x}$, $f : X \rightarrow \pm 1$

Classifiers, Loss function

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

- Empirical risk (estimated on the training set)

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

- For the zero-one loss function:

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

the loss is 0 if $(\boldsymbol{x}, y)$ 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)

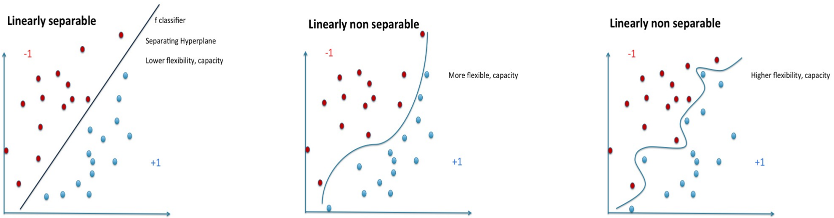
$$R[f] \;=\; \int \frac{1}{2}|f(\boldsymbol{x}) - y|\, dP(\boldsymbol{x}, y)$$

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

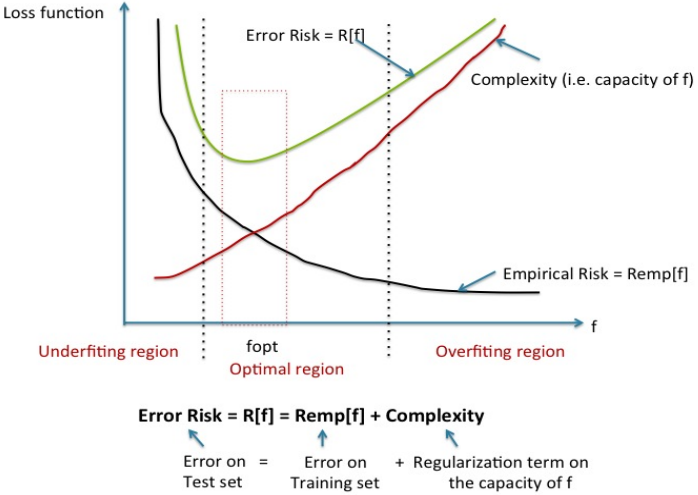
Classifiers, Loss function

- 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



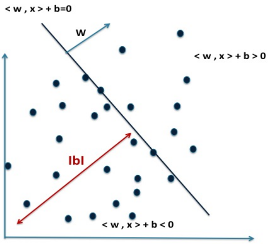
Classifiers, Loss function



Hyperplanes

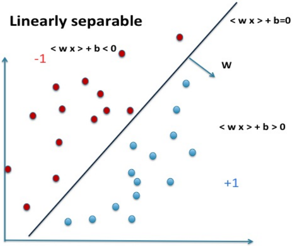
H a dot vectorial space $<,>$
$\boldsymbol{x}_1, ...\boldsymbol{x}_m$ m points of H
An hyperplan HP is defined:

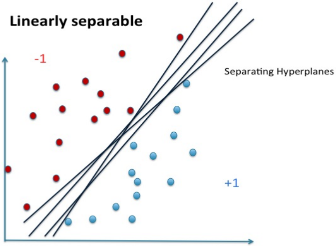
$$\{\boldsymbol{x} \in H \;/\; <\boldsymbol{w}, \boldsymbol{x}> + b = 0\} \; \boldsymbol{w} \in H, b \in \mathbb{R}$$



Separating Hyperplanes

- Binary classification
- Linearly separable points $\boldsymbol{x}_1, \ldots \boldsymbol{x}_m$ of H



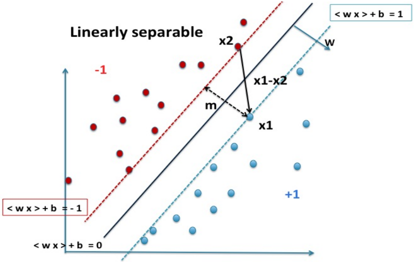


Canonical Hyperplan

Definition

The pair $(\boldsymbol{w}, b)$ is called a canonical hyperplan w.r.t. $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \in H$, if it is scaled such that

$$\min_{i=1\ldots m} |< \boldsymbol{w}, \boldsymbol{x}_i > +b| = 1 \qquad (1)$$



Canonical Hyperplan

Let Hp_0, Hp_{+1} and Hp_{-1} be the three hyperplans as indicated in the above figure
Let $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ be the closest points to Hp_0 (see Fig), then

$$< \boldsymbol{w}, \boldsymbol{x}_1 > + b = c > 0$$
$$< \boldsymbol{w}, \boldsymbol{x}_2 > + b = -c < 0$$

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

$$\begin{aligned} \alpha < \boldsymbol{w}, \boldsymbol{x}_1 > + \alpha\, b &= < \boldsymbol{w'}, \boldsymbol{x'}_1 > + b' = 1 \\ \alpha < \boldsymbol{w}, \boldsymbol{x}_2 > + \alpha\, b &= < \boldsymbol{w'}, \boldsymbol{x'}_2 > + b' = -1 \end{aligned}$$

Canonical Hyperplan

Margin value

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

$$
\begin{aligned}
< \boldsymbol{w}, \boldsymbol{x}_1 > +b &= 1 \quad (2)\\
< \boldsymbol{w}, \boldsymbol{x}_2 > +b &= -1 \quad (3)\\
\text{from (2)-(3)} \quad & < \boldsymbol{w}, (\boldsymbol{x}_1 - \boldsymbol{x}_2) >= 2 > \text{and} < \frac{\boldsymbol{w}}{\|w\|}, (\boldsymbol{x}_1 - \boldsymbol{x}_2) >= \frac{2}{\|\boldsymbol{w}\|} \quad (4)
\end{aligned}
$$

gives the orthogonal projection of $(\boldsymbol{x}_1 - \boldsymbol{x}_2)$ onto the line of direction $\boldsymbol{w}$. The distance of the closest point to the hyperplan (margin m) is then:

$$
m = \frac{1}{\|\boldsymbol{w}\|}
$$

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

Hard-margin Support Vector Machine

- Let $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m)$ be m points, $\boldsymbol{x}_i \in H$
- 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 $\boldsymbol{w}$ (vs. opposite direction to $\boldsymbol{w}$)
- The decision function $f_{\boldsymbol{w},b}$ that gives the class label of a given $\boldsymbol{x}$

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

Hard-margin 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

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

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

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

Hard-margin Support Vector Machine

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

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

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

Note that:

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

Hard-margin Support Vector Machine: Dual formalization

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

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

The decision function

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

For $\boldsymbol{x}_i$ 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 ($R_{Emp} \rightsquigarrow 0$) 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.

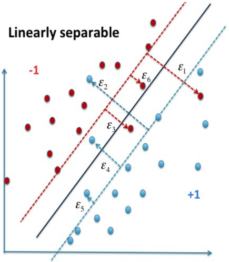
Soft-margin SVM

Primal formalization

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

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

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



Soft-margin SVM

$$\min_{\boldsymbol{w}\in H, \xi\in\mathbb{R}^m, b\in\mathbb{R}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\,\frac{1}{m}\sum_{i=1}^{m}\xi_i \tag{11}$$

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

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

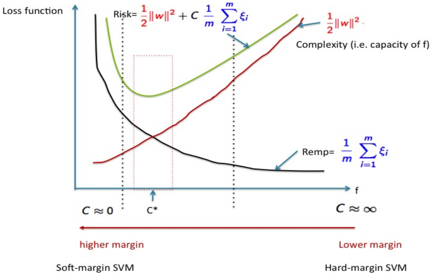
Some intuitions (1)

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

$$\max_{\alpha \in \mathbb{R}^m} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \, \alpha_j \, y_i \, y_j \; < \boldsymbol{x}_i, \boldsymbol{x}_j > \tag{12}$$

$$s.t. \quad 0 \leq \alpha_i \leq \frac{C}{m} \,, i = 1, ..., m$$

$$\sum_{i=1}^{m} \alpha_i \; y_i = 0$$

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

$$\begin{aligned} \min_{\mathbf{w}\in H, \xi\in\mathbb{R}^m, b\in\mathbb{R}, \rho\in\mathbb{R}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 - \nu\,\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i \\ s.t. \quad & y_i(<\mathbf{x}_i, \mathbf{w}> + b) \geq \rho - \xi_i \quad \forall i = 1, ..., m \\ & \xi_i \geq 0 \quad \forall i = 1, ..., m \\ & \rho \geq 0 \end{aligned} \tag{13}$$

ν-SVM

Interpretation of ρ

1. The classes are separated by a margin of $\frac{2\rho}{\|\mathbf{w}\|^2}$
2. $\nu \in [0, 1]$ 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. ν is a lower bound of the proportion 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 ν increases the margin, that allows the increase of the margin errors

ν-SVM

Dual formalization

$$\max_{\alpha \in \mathbb{R}^m} \quad -\frac{1}{2}\sum_{i,j} \alpha_i \; \alpha_j \; y_i \; y_j \; < \boldsymbol{x}_i, \boldsymbol{x}_j > \tag{14}$$

$$s.t. \quad 0 \leq \alpha_i \leq \frac{1}{m} \; , i = 1, ..., m$$

$$\sum_{i=1}^{m} \alpha_i \; y_i = 0$$

$$\sum_{i=1}^{m} \alpha_i \geq 0$$

Multi-class SVM

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

1- One versus all approach

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

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

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

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

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

Multi-class SVM: One versus all approach

Remarks

- For $g^j(\boldsymbol{x}^*) > 0$, assign $\boldsymbol{x}^*$ to the *jth* class, otherwise the only decision is that $\boldsymbol{x}^*$ is not in the *jth* class
- Some samples may not be classified (for instance, $g^j(\boldsymbol{x}^*) < 0$, 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 $\boldsymbol{x}^*$ by a majority vote through the $K(K-1)$ decision functions $f_{ij}(\boldsymbol{x}^*) = sign(g_{ij}(\boldsymbol{x}^*))$

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 than 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

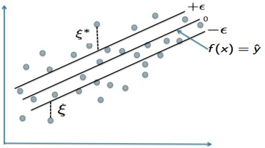
Let $(\boldsymbol{x}_i, y_i)$ $i = 1, ..., m$, $y_i \in \mathbb{R}$, the aim of SVR is the estimation of $\hat{y} = f(\boldsymbol{x})$ that minimizes the ϵ-insensitive Loss-function R^{ϵ}_{Emp}:

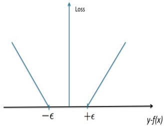
$$R^{\epsilon}_{Emp} = |f(\boldsymbol{x}) - y|_{\epsilon} = max(0, |f(\boldsymbol{x}) - y| - \epsilon)$$

Remarks

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

Support Vector Regression ($\epsilon - SVR$)





Primal formalization

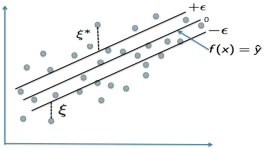
$$\min_{\mathbf{w}\in H, \xi^{(*)}\in\mathbb{R}^m, b\in\mathbb{R}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{15}$$

$$\begin{aligned} s.t. \quad & (<\mathbf{x}_i, \mathbf{w}> + b) - y_i \leq \epsilon + \xi_i \quad \forall i = 1, ..., m \\ & y_i - (<\mathbf{x}_i, \mathbf{w}> + b) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \quad \forall i = 1, ..., m \end{aligned} \tag{16}$$

$\epsilon - SVR$: Primal formalization

$$\begin{aligned} \min_{\mathbf{w}\in H, \xi^{(*)}\in\mathbb{R}^m, b\in\mathbb{R}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ s.t. \quad & (<\mathbf{x}_i, \mathbf{w}> + b) - y_i \leq \epsilon + \xi_i \quad \forall i = 1, ..., m \\ & y_i - (<\mathbf{x}_i, \mathbf{w}> + b) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \quad \forall i = 1, ..., m \end{aligned} \tag{17}$$

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



$\epsilon - SVR$

Some intuitions

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

$\epsilon - SVR$: Dual formalization

Introducing Lagrange multipliers, on the primal form Eq. (17), one arrives at the following optimization problem (C and ϵ selected *a priori*)

$$\begin{aligned}
\max_{\alpha,\alpha^* \in \mathbb{R}^m} \quad & -\epsilon \sum_{i=1}^{m}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{m}(\alpha_i^* - \alpha_i) y_i \\
& -\frac{1}{2}\sum_{i,j}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) < \mathbf{x}_i, \mathbf{x}_j > \\
s.t. \quad & 0 \leq \alpha_i^*, \alpha_i \leq \frac{C}{m} \;\; \forall\, i = 1, ..., m \\
& \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)
\end{aligned} \tag{18}$$

The regression estimate

$$\begin{aligned}
f(\mathbf{x}) &= \sum_{i=1}^{m}(\alpha_i^* - \alpha_i) < \mathbf{x}_i, \mathbf{x} > + b \\
\mathbf{w} &= \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\, \mathbf{x}_i
\end{aligned} \tag{19}$$

$\epsilon - SVR$: 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 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

$$\alpha_i(\epsilon + \xi_i - y_i + < \mathbf{w}, \mathbf{x}_i > + b) = 0 \tag{20}$$

$$\alpha_i(\epsilon + \xi_i^* + y_i - < \mathbf{w}, \mathbf{x}_i > - b) = 0 \tag{21}$$

$$(\frac{C}{m} - \alpha_i)\xi_i = 0 \quad (\frac{C}{m} - \alpha_i^*)\xi_i^* = 0 \tag{22}$$

$\epsilon - SVR$: Dual formalization

Useful derived conclusions

- Only samples $(\boldsymbol{x}_i, y_i)$ 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,

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

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

$$\begin{aligned} \min_{\mathbf{w}\in H, \xi^{(*)}\in\mathbb{R}^m, b\in\mathbb{R}, \epsilon\in\mathbb{R}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\nu\epsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*)\right) \\ s.t. \quad & (<\mathbf{x}_i, \mathbf{w}> + b) - y_i \leq \epsilon + \xi_i \quad \forall i = 1, ..., m \\ & y_i - (<\mathbf{x}_i, \mathbf{w}> + b) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \end{aligned} \tag{23}$$

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

$$
\begin{aligned}
\max_{\alpha,\alpha^* \in \mathbb{R}^m} \quad & \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) y_i - \frac{1}{2} \sum_{i,j}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) < \mathbf{x}_i, \mathbf{x}_j > \\
s.t. \quad & 0 \leq \alpha_i^*, \alpha_i \leq \frac{C}{m} \;\; \forall\, i = 1, ..., m \\
& \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) \\
& \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) \leq C.\nu
\end{aligned}
\tag{24}
$$

The regression estimate

$$
\begin{aligned}
f(\mathbf{x}) &= \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) < \mathbf{x}_i, \mathbf{x} > + b \\
\mathbf{w} &= \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\, \mathbf{x}_i
\end{aligned}
\tag{25}
$$

$\nu - SVR$:

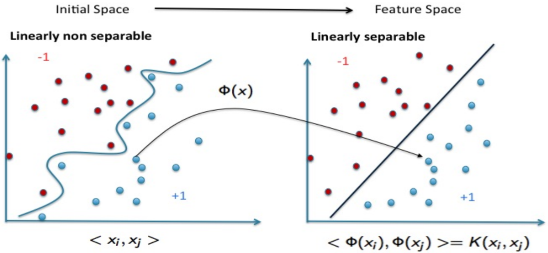
Interpretation of the a priori fixed $\nu \in [0, 1]$

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

with suitable choices of $d \in \mathbb{N}$, $\sigma, \kappa, \Theta \ \in \mathbb{R}$ 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:

$$
\begin{aligned}
DTW &= \min_{\pi \in A(n,m)} D_{x,y}(\pi) \\
D_{x,y} &= \sum_{i=1}^{|\pi|} \varphi(\boldsymbol{x}_{\pi_1(i)}, y_{\pi_2(i)}) \\
K_{GA}(\mathbf{x}, \mathbf{y}) &= \sum_{\pi \in A(n,m)} e^{-D_{x,y}(\pi)} \\
&= \sum_{\pi \in A(n,m)} \prod_{i=1}^{|\pi|} k(\boldsymbol{x}_{\pi_1(i)}, y_{\pi_2(i)})
\end{aligned}
$$

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

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

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

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

with $\gamma_{i,j}$ 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

$$DTW_{DTAK}(\mathbf{x}, \mathbf{y}) \quad = \quad \max x_{\pi \in A(n,m)} \sum_{i=1}^{|\pi|} k_{\sigma}(\boldsymbol{x}_{\pi_{\mathbf{1}}(i)}, y_{\pi_{\mathbf{2}}(i)})$$