Non-linear Support Vector Machines

Andrea Passerini
passerini@disi.unitn.it

Machine Learning
## Non-linear Support Vector Machines

### Non-linearly separable problems
- Hard-margin SVM can address linearly separable problems
- Soft-margin SVM can address linearly separable problems with outliers
- Non-linearly separable problems need a higher expressive power (i.e. more complex feature combinations)
- We do not want to lose the advantages of linear separators (i.e. large margin, theoretical guarantees)

### Solution
- Map input examples in a higher dimensional *feature space*
- Perform linear classification in this higher dimensional space
Non-linear Support Vector Machines

\[ \Phi : \mathcal{X} \rightarrow \mathcal{H} \]

- \( \Phi \) is a function mapping each example to a higher dimensional space \( \mathcal{H} \).
- Examples \( \mathbf{x} \) are replaced with their feature mapping \( \Phi(\mathbf{x}) \).
- The feature mapping should increase the expressive power of the representation (e.g., introducing features which are combinations of input features).
- Examples should be (approximately) linearly separable in the mapped space.
Feature map

Homogeneous
\( (d = 2) \)

\[
\Phi \left( \begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left( \begin{array}{c} x_1^2 \\ x_1 x_2 \\ x_2^2 \end{array} \right)
\]

Inhomogeneous
\( (d = 2) \)

\[
\Phi \left( \begin{array}{c} x_1 \\ x_2 \end{array} \right) = \left( \begin{array}{c} x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ x_2^2 \end{array} \right)
\]

Polynomial mapping

- Maps features to all possible conjunctions (i.e. products) of features:
  - of a certain degree \( d \) (homogeneous mapping)
  - up to a certain degree (inhomogeneous mapping)
Feature map

\[ \Phi \]
Non-linear Support Vector Machines

A linear separation (i.e. hyperplane) in feature space corresponds to a non-linear separation in input space, e.g.:

$$f(x) = w^T \Phi(x) + w_0$$

Non-linear SVM
Support Vector Regression

Rationale

- Retain combination of regularization and data fitting
- Regularization means *smoothness* (i.e. smaller weights, lower complexity) of the learned function
- Use a sparsifying loss to have few support vector
Support Vector Regression

\[ \ell(f(x), y) = |y - f(x)|_\epsilon = \begin{cases} 
    0 & \text{if } |y - f(x)| \leq \epsilon \\
    |y - f(x)| - \epsilon & \text{otherwise}
\end{cases} \]

- Tolerate small (\(\epsilon\)) deviations from the true value (i.e. no penalty)
- Defines an \(\epsilon\)-tube of insensitiveness around true values
- This also allows to trade off function complexity with data fitting (playing onn \(\epsilon\) value)
Support Vector Regression

**Optimization problem**

\[
\begin{align*}
\min_{w \in \mathcal{X}, w_0 \in \mathbb{R}, \xi, \xi^* \in \mathbb{R}^m} & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \\
\text{subject to} & \quad w^T \Phi(x_i) + w_0 - y_i \leq \epsilon + \xi_i \\
& \quad y_i - (w^T \Phi(x_i) + w_0) \leq \epsilon + \xi_i^* \\
& \quad \xi_i, \xi_i^* \geq 0
\end{align*}
\]

**Note**

- Two constraints for each example for the upper and lower sides of the tube
- Slack variables $\xi_i, \xi_i^*$ penalize predictions out of the $\epsilon$-insensitive tube

Non-linear SVM
We include constraints in the minimization function using Lagrange multipliers ($\alpha_i, \alpha_i^*, \beta_i, \beta_i^* \geq 0$):

$$
L = \frac{1}{2} ||\mathbf{w}||^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) - \sum_{i=1}^{m} (\beta_i \xi_i + \beta_i^* \xi_i^*) \\
- \sum_{i=1}^{m} \alpha_i (\epsilon + \xi_i + y_i - \mathbf{w}^T \Phi(\mathbf{x}_i) - w_0) \\
- \sum_{i=1}^{m} \alpha_i^* (\epsilon + \xi_i^* - y_i + \mathbf{w}^T \Phi(\mathbf{x}_i) + w_0)
$$
Dual formulation

Vanishing the derivatives wrt the primal variables we obtain:

\[
\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) \Phi(\mathbf{x}_i) = 0 \rightarrow \mathbf{w} = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) \Phi(\mathbf{x}_i)
\]

\[
\frac{\partial L}{\partial \mathbf{w}_0} = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0
\]

\[
\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \rightarrow \alpha_i \in [0, C]
\]

\[
\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \beta_i^* = 0 \rightarrow \alpha_i^* \in [0, C]
\]
Dual formulation

Substituting in the Lagrangian we get:

\[
\frac{1}{2} \sum_{i,j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \Phi(x_i)^T \Phi(x_j)
\]

\[
\sum_{i=1}^{m} \xi_i (C - \beta_i - \alpha_i) + \sum_{i=1}^{m} \xi_i^* (C - \beta_i^* - \alpha_i^*) = 0
\]

\[
-\epsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i (\alpha_i^* - \alpha_i) + w_0 \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0
\]

\[
- \sum_{i,j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \Phi(x_i)^T \Phi(x_j)
\]
Support Vector Regression

**Dual formulation**

\[
\max_{\alpha \in \mathbb{R}^m} - \frac{1}{2} \sum_{i,j=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \Phi(x_i)^T \Phi(x_j) \\
- \epsilon \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{m} y_i (\alpha_i^* - \alpha_i) \\
\text{subject to} \quad \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0 \\
\alpha_i, \alpha_i^* \in [0, C] \quad \forall i \in [1, m]
\]

**Regression function**

\[
f(x) = w^T \Phi(x) + w_0 = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \Phi(x_i)^T \Phi(x) + w_0
\]
Support Vector Regression

Karush-Khun-Tucker conditions (KKT)

- At the saddle point it holds that for all $i$:

$$
\alpha_i (\epsilon + \xi_i + y_i - \mathbf{w}^T \Phi(\mathbf{x}_i) - w_0) = 0
$$

$$
\alpha_i^* (\epsilon + \xi_i^* + y_i - \mathbf{w}^T \Phi(\mathbf{x}_i) + w_0) = 0
$$

$$
(C - \alpha_i) \xi_i = 0
$$

$$
(C - \alpha_i^*) \xi_i^* = 0
$$

Non-linear SVM
Support Vector Regression

Support Vectors

- All patterns within the $\epsilon$-tube, for which $|f(x_i) - y_i| < \epsilon$, have $\alpha_i, \alpha_i^* = 0$ and thus don’t contribute to the estimated function $f$.

- Patterns for which either $0 < \alpha_i < C$ or $0 < \alpha_i^* < C$ are on the border of the $\epsilon$-tube, that is $|f(x_i) - y_i| = \epsilon$. They are the unbound support vectors.

- The remaining training patterns are margin errors (either $\xi_i > 0$ or $\xi_i^* > 0$), and reside out of the $\epsilon$-insensitive region. They are bound support vectors, with corresponding $\alpha_i = C$ or $\alpha_i^* = C$. 

Non-linear SVM
Support Vectors

Non-linear SVM

\[ y - w^T \Phi(x) - w_0 = \epsilon \]
\[ y - w^T \Phi(x) - w_0 = 0 \]
\[ w^T \Phi(x) + w_0 - y = -\epsilon \]
Support Vector Regression: example for decreasing $\epsilon$

Non-linear SVM
Smallest Enclosing Hypersphere

<table>
<thead>
<tr>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Characterize a set of examples defining boundaries enclosing them</td>
</tr>
<tr>
<td>- Find smallest hypersphere in feature space enclosing data points</td>
</tr>
<tr>
<td>- Account for outliers paying a cost for leaving examples out of the sphere</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>- One-class classification: model a class when no negative examples exist</td>
</tr>
<tr>
<td>- Anomaly/novelty detection: detect test data falling outside of the sphere and return them as novel/anomalous (e.g. intrusion detection systems, Alzheimer’s patients monitoring)</td>
</tr>
</tbody>
</table>
Smallest Enclosing Hypersphere

Optimization problem

\[
\begin{align*}
\min_{R \in \mathbb{R}, \mathbf{o} \in \mathcal{H}, \xi \in \mathbb{R}^m} & \quad R^2 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to} & \quad \|\Phi(x_i) - \mathbf{o}\|^2 \leq R^2 + \xi_i \quad i = 1, \ldots, m \\
& \quad \xi_i \geq 0, \quad i = 1, \ldots, m
\end{align*}
\]

Note
- \( \mathbf{o} \) is the center of the sphere
- \( R \) is the radius which is minimized
- slack variables \( \xi_i \) gather costs for outliers

Non-linear SVM
Smallest Enclosing Hypersphere

Lagrangian \((\alpha_i, \beta_i \geq 0)\)

\[
L = R^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i (R^2 + \xi_i - \|\Phi(x_i) - o\|^2) - \sum_{i=1}^{m} \beta_i \xi_i
\]

Vanishing the derivatives wrt primal variables

\[
\frac{\partial L}{\partial R} = 2R(1 - \sum_{i=1}^{m} \alpha_i) = 0 \rightarrow \sum_{i=1}^{m} \alpha_i = 1
\]

\[
\frac{\partial L}{\partial o} = 2 \sum_{i=1}^{m} \alpha_i (\Phi(x_i) - o)(-1) = 0 \rightarrow o \sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} \alpha_i \Phi(x_i)
\]

\[
\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \rightarrow \alpha_i \in [0, C]
\]

Non-linear SVM
**Smallest Enclosing Hypersphere**

**Dual formulation**

\[
R^2 \left( 1 - \sum_{i=1}^{m} \alpha_i \right) + \sum_{i=1}^{m} \xi_i \left( C - \alpha_i - \beta_i \right) = 0
\]

\[
+ \sum_{i=1}^{m} \alpha_i (\Phi(x_i) - \sum_{j=1}^{m} \alpha_j \Phi(x_j))^T (\Phi(x_i) - \sum_{h=1}^{m} \alpha_h \Phi(x_h)) = 0
\]

**Non-linear SVM**
Smallest Enclosing Hypersphere

**Dual formulation**

\[
\sum_{i=1}^{m} \alpha_i (\Phi(x_i) - \sum_{j=1}^{m} \alpha_j \Phi(x_j))^T (\Phi(x_i) - \sum_{h=1}^{m} \alpha_h \Phi(x_h)) = \\
= \sum_{i=1}^{m} \alpha_i \Phi(x_i)^T \Phi(x_i) - \sum_{i=1}^{m} \alpha_i \Phi(x_i)^T \sum_{h=1}^{m} \alpha_h \Phi(x_h) \\
- \sum_{i=1}^{m} \alpha_i \sum_{j=1}^{m} \alpha_j \Phi(x_j)^T \Phi(x_i) + \sum_{i=1}^{m} \alpha_i \sum_{j=1}^{m} \alpha_j \Phi(x_j)^T \sum_{h=1}^{m} \alpha_h \Phi(x_h) = \\
= \sum_{i=1}^{m} \alpha_i \Phi(x_i)^T \Phi(x_i) - \sum_{i=1}^{m} \alpha_i \Phi(x_i)^T \sum_{j=1}^{m} \alpha_j \Phi(x_j)
\]

Non-linear SVM
Smallest Enclosing Hypersphere

Dual formulation

$$\max_{\alpha \in \mathbb{R}^m} \sum_{i=1}^{m} \alpha_i \Phi(x_i)^T \Phi(x_i) - \sum_{i,j=1}^{m} \alpha_i \alpha_j \Phi(x_i)^T \Phi(x_j)$$

subject to $$\sum_{i=1}^{m} \alpha_i = 1, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m.$$  

Distance function

- The distance of a point from the origin is:
  $$R^2(x) = \|\Phi(x) - \mathbf{0}\|^2$$
  $$= (\Phi(x) - \sum_{i=1}^{m} \alpha_i \Phi(x_i))^T (\Phi(x) - \sum_{j=1}^{m} \alpha_j \Phi(x_j))$$
  $$= \Phi(x)^T \Phi(x) - 2 \sum_{i=1}^{m} \alpha_i \Phi(x_i)^T \Phi(x) + \sum_{i,j=1}^{m} \alpha_i \alpha_j \Phi(x_i)^T \Phi(x_j)$$
Smallest Enclosing Hypersphere

Karush-Khun-Tucker conditions (KKT)

- At the saddle point it holds that for all $i$:

\[ \beta_i \xi_i = 0 \]

\[ \alpha_i (R^2 + \xi_i - \|\Phi(x_i) - o\|^2) = 0 \]

Support vectors

- Unbound support vectors ($0 < \alpha_i < C$), whose images lie on the surface of the enclosing sphere.
- Bound support vectors ($\alpha_i = C$), whose images lie outside of the enclosing sphere, which correspond to outliers.
- All other points ($\alpha = 0$) with images inside the enclosing sphere.

Non-linear SVM
Smallest Enclosing Hypersphere

Non-linear SVM
Smallest Enclosing Hypersphere

### Decision function

- The radius $R^*$ of the enclosing sphere can be computed using the distance function on any unbound support vector $x$:

$$R^2(x) = \Phi(x)^T \Phi(x) - 2 \sum_{i=1}^{m} \alpha_i \Phi(x_i)^T \Phi(x) + \sum_{i,j=1}^{m} \alpha_i \alpha_j \Phi(x_i)^T \Phi(x_j)$$

- A decision function for novelty detection could be:

$$f(x) = \text{sgn} \left( R^2(x) - (R^*)^2 \right)$$

- i.e. positive if the examples lays outside of the sphere and negative otherwise
Rationale

- Order examples by relevance (e.g. email urgency, movie rating)
- Learn scoring function predicting quality of example
- Constraint function to score $x_i$ higher than $x_j$ if it is more relevant (pairwise comparisons for training)
- Easily formalized as a support vector classification task
Support Vector Ranking

Optimization problem

\[
\begin{align*}
\min_{\mathbf{w} \in \mathcal{X}, w_0 \in \mathbb{R}, \xi_{i,j} \in \mathbb{R}} & \quad \frac{1}{2} ||\mathbf{w}||^2 + C \sum_{i,j} \xi_{i,j} \\
\text{subject to} & \quad \mathbf{w}^T \Phi(\mathbf{x}_i) - \mathbf{w}^T \Phi(\mathbf{x}_j) \geq 1 - \xi_{i,j} \\
& \quad \xi_{i,j} \geq 0 \\
& \quad \forall i, j : \mathbf{x}_i \prec \mathbf{x}_j
\end{align*}
\]

Note

- There is one constraint for each pair of examples having ordering information (\(\mathbf{x}_i \prec \mathbf{x}_j\) means the former is comes first in the ranking)
- Examples should be correctly ordered with a large margin

Non-linear SVM
Support vector classification on pairs

\[
\min_{w \in \mathcal{X}, w_0 \in \mathbb{R}, \xi_{i,j} \in \mathbb{R}} \quad \frac{1}{2} \|w\|^2 + C \sum_{i,j} \xi_{i,j}
\]

subject to

\[
y_{i,j} w^T (\Phi(x_i) - \Phi(x_j)) \geq 1 - \xi_{i,j}
\]

\[
\Phi(x_{ij})
\]

\[
\xi_{i,j} \geq 0
\]

\[
\forall i, j : x_i \prec x_j
\]

where labels are always positive \( y_{i,j} = 1 \)
Support Vector Ranking

Decision function

\[ f(x) = \mathbf{w}^T \phi(x) \]

- Standard support vector classification function (unbiased)
- Represents score of example for ranking it

Non-linear SVM
