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 loose 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

feature map

$$\Phi : \mathcal{X} \rightarrow \mathcal{H}$$

- Φ is a function mapping each example to a higher dimensional space $\mathcal{H}$
- Examples $\boldsymbol{x}$ are replaced with their feature mapping $\Phi(\boldsymbol{x})$
- 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
($d = 2$)

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$

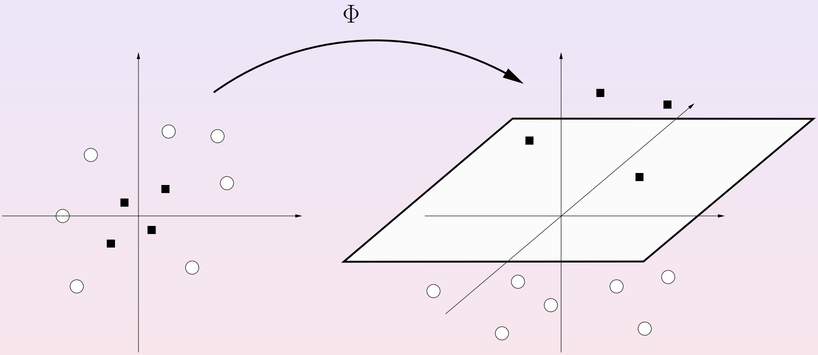
Inhomogeneous
($d = 2$)

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$

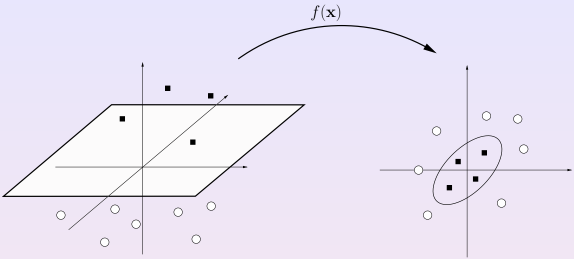
Polynomial mapping

- Maps features to all possible conjunctions (i.e. products) of features:
  1. *of a certain degree d* (*homogeneous* mapping)
  2. *up to a certain degree* (*inhomogeneous* mapping)

Feature map



Non-linear Support Vector Machines



Linear separation in feature space

- SVM algorithm is applied just replacing $\boldsymbol{x}$ with $\Phi(\mathbf{x})$:

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

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

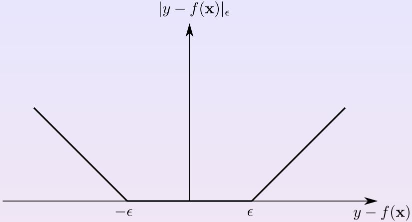
$$f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathrm{sgn}(w_1 x_1^2 + w_2 x_1 x_2 + w_3 x_2^2 + w_0)$$

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



ϵ-insensitive loss

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

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

Support Vector Regression

Optimization problem

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

Note

- Two constraints for each example for the upper and lower sides of the tube
- Slack variables ξ_i, ξ_i^* penalize predictions out of the ϵ-insensitive tube

Support Vector Regression

Lagrangian

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

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

Support Vector Regression

Dual formulation

- Vanishing the derivatives wrt the primal variables we obtain:

$$
\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\Phi(\mathbf{x}_i) = 0 \rightarrow \mathbf{w} = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\Phi(\mathbf{x}_i) \\
\frac{\partial L}{\partial w_0} &= \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0 \\
\frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0 \rightarrow \alpha_i \in [0, C] \\
\frac{\partial L}{\partial \xi_i^*} &= C - \alpha_i^* - \beta_i^* = 0 \rightarrow \alpha_i^* \in [0, C]
\end{aligned}
$$

Dual formulation

- Substituting in the Lagrangian we get:

$$\underbrace{\frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)}_{||\mathbf{w}||^2}$$

$$+\sum_{i=1}^{m}\xi_i\underbrace{(C-\beta_i-\alpha_i)}_{=0}+\sum_{i=1}^{m}\xi_i^*\underbrace{(C-\beta_i^*-\alpha_i^*)}_{=0}$$

$$-\epsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{m}y_i(\alpha_i^*-\alpha_i)+w_0\underbrace{\sum_{i=1}^{m}(\alpha_i-\alpha_i^*)}_{=0}$$

$$-\sum_{i,j=1}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$$

Support Vector Regression

Dual formulation

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \quad -\frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$$

$$-\epsilon\sum_{i=1}^{m}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{m} y_i(\alpha_i^* - \alpha_i)$$

$$\text{subject to} \quad \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$$

$$\alpha_i, \alpha_i^* \in [0, C] \quad \forall i \in [1, m]$$

Regression function

$$f(\mathbf{x}) = \mathbf{w}^T\Phi(\mathbf{x}) + w_0 = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}) + w_0$$

Support Vector Regression

Karush-Khun-Tucker conditions (KKT)

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

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

- Combined with $C - \alpha_i - \beta_i = 0, \alpha_i \geq 0, \beta_i \geq 0$ and $C - \alpha_i^* - \beta_i^* = 0, \alpha_i^* \geq 0, \beta_i^* \geq 0$ we get

$$\alpha_i \in [0, C] \qquad \alpha_i^* \in [0, C]$$

- and

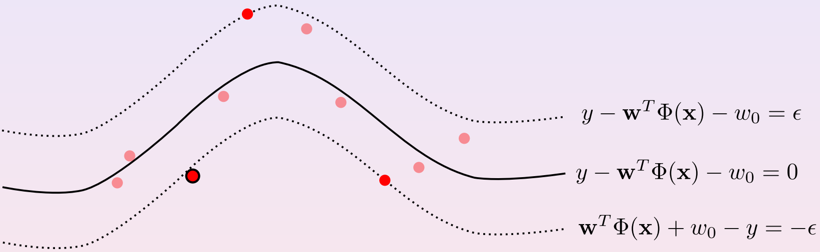
$$\alpha_i = C \textit{ if } \xi_i > 0 \qquad \alpha_i^* = C \textit{ if } \xi_i^* > 0$$

Support Vector Regression

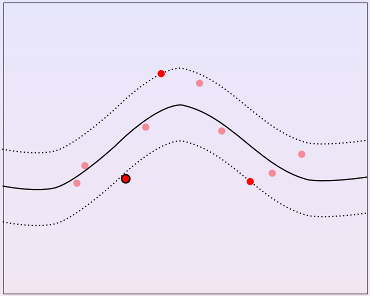
Support Vectors

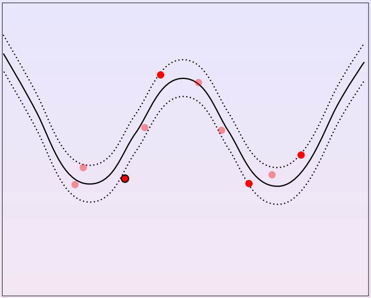
- All patterns within the ϵ-tube, for which $|f(\mathbf{x}_i) - y_i| < \epsilon$, have $\alpha_i, \alpha_i^* = 0$ and thus don't contribute to the estimated function f.
- Patterns for which either $0 < \alpha_i < C$ or $0 < \alpha_i^* < C$ are on the border of the ϵ-tube, that is $|f(\mathbf{x}_i) - y_i| = \epsilon$. 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 ϵ-insensitive region. They are bound support vectors, with corresponding $\alpha_i = C$ or $\alpha_i^* = C$.

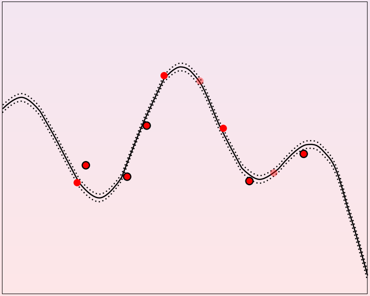
Support Vectors



Support Vector Regression: example for decreasing ϵ







References

Non linear SVM C. Burges, *A tutorial on support vector machines for pattern recognition*, Data Mining and Knowledge Discovery, 2(2), 121-167, 1998.

Support vector regression J.Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004 (Section 7.3.3)

Appendix

- Smallest enclosing hypersphere
- Support vector ranking

Smallest Enclosing Hypersphere

Rationale

- Characterize a set of examples defining boundaries enclosing them
- Find smallest hypersphere in feature space enclosing data points
- Account for outliers paying a cost for leaving examples out of the sphere

Usage

- One-class classification: model a class when no negative examples exist
- 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)

Smallest Enclosing Hypersphere

Optimization problem

$$
\begin{aligned}
\min_{R\in\mathbb{R},\boldsymbol{o}\in\mathcal{H},\boldsymbol{\xi}\in\mathbb{R}^m} \quad & R^2 + C\sum_{i=1}^{m} \xi_i \\
\text{subject to} \quad & ||\Phi(\mathbf{x}_i) - \boldsymbol{o}||^2 \leq R^2 + \xi_i \quad i = 1, \dots, m \\
& \xi_i \geq 0, \quad i = 1, \dots, m
\end{aligned}
$$

Note

- $\boldsymbol{o}$ is the center of the sphere
- R is the radius which is minimized
- slack variables ξ_i gather costs for outliers

Smallest Enclosing Hypersphere

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

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

Vanishing the derivatives wrt primal variables

$$\begin{aligned}
\frac{\partial L}{\partial R} &= 2R(1 - \sum_{i=1}^{m}\alpha_i) = 0 \rightarrow \sum_{i=1}^{m}\alpha_i = 1 \\
\frac{\partial L}{\partial \boldsymbol{o}} &= 2\sum_{i=1}^{m}\alpha_i(\Phi(x_i) - \boldsymbol{o})(-1) = 0 \rightarrow \boldsymbol{o}\underbrace{\sum_{i=1}^{m}\alpha_i}_{=1} = \sum_{i=1}^{m}\alpha_i\Phi(\mathbf{x}_i) \\
\frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0 \rightarrow \alpha_i \in [0, C]
\end{aligned}$$

Smallest Enclosing Hypersphere

Dual formulation

$$R^2 \underbrace{(1 - \sum_{i=1}^{m} \alpha_i)}_{=0} + \sum_{i=1}^{m} \xi_i \underbrace{(C - \alpha_i - \beta_i)}_{=0}$$

$$+ \sum_{i=1}^{m} \alpha_i (\Phi(\mathbf{x}_i) - \underbrace{\sum_{j=1}^{m} \alpha_j \Phi(\mathbf{x}_j)}_{\boldsymbol{o}})^T (\Phi(\mathbf{x}_i) - \underbrace{\sum_{h=1}^{m} \alpha_h \Phi(\mathbf{x}_h)}_{\boldsymbol{o}})$$

Smallest Enclosing Hypersphere

Dual formulation

$$\sum_{i=1}^{m}\alpha_i(\Phi(\mathbf{x}_i)-\sum_{j=1}^{m}\alpha_j\Phi(\mathbf{x}_j))^T(\Phi(\mathbf{x}_i)-\sum_{h=1}^{m}\alpha_h\Phi(\mathbf{x}_h))=$$

$$=\sum_{i=1}^{m}\alpha_i\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_i)-\sum_{i=1}^{m}\alpha_i\Phi(\mathbf{x}_i)^T\sum_{h=1}^{m}\alpha_h\Phi(\mathbf{x}_h)$$

$$-\sum_{i=1}^{m}\alpha_i\sum_{j=1}^{m}\alpha_j\Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i)+\underbrace{\sum_{i=1}^{m}\alpha_i}_{=1}\sum_{j=1}^{m}\alpha_j\Phi(\mathbf{x}_j)^T\sum_{h=1}^{m}\alpha_h\Phi(\mathbf{x}_h)=$$

$$=\sum_{i=1}^{m}\alpha_i\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_i)-\sum_{i=1}^{m}\alpha_i\Phi(\mathbf{x}_i)^T\sum_{j=1}^{m}\alpha_j\Phi(\mathbf{x}_j)$$

Smallest Enclosing Hypersphere

Dual formulation

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \quad \sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_i) - \sum_{i,j=1}^{m} \alpha_i \alpha_j \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$$

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

Distance function

- The distance of a point from the origin is:

$$\begin{aligned} R^2(x) &= ||\Phi(x) - \boldsymbol{o}||^2 \\ &= (\Phi(\mathbf{x}) - \sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x}_i))^T (\Phi(\mathbf{x}) - \sum_{j=1}^{m} \alpha_j \Phi(\mathbf{x}_j)) \\ &= \Phi(\mathbf{x})^T \Phi(\mathbf{x}) - 2\sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}) + \sum_{i,j=1}^{m} \alpha_i \alpha_j \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \end{aligned}$$

Smallest Enclosing Hypersphere

Karush-Khun-Tucker conditions (KKT)

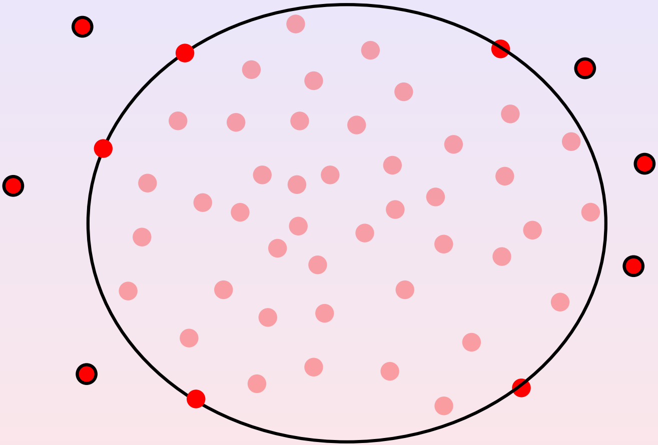
- At the saddle point it holds that for all i:

$$\beta_i \xi_i = 0$$
$$\alpha_i(R^2 + \xi_i - ||\Phi(\mathbf{x}_i) - \boldsymbol{o}||^2) = 0$$

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.

Smallest Enclosing Hypersphere



Smallest Enclosing Hypersphere

Decision function

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

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

- A decision function for novelty detection could be:

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

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

Support Vector Ranking

Rationale

- Order examples by relevance (e.g. email urgence, movie rating)
- Learn scoring function predicting quality of example
- Constraint function to score $\mathbf{x}_i$ higher than $\mathbf{x}_j$ if it is more relevant (pairwise comparisons for training)
- Easily formalized as a support vector classification task

Support Vector Ranking

Optimization problem

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

subject to

$$\mathbf{w}^T\Phi(\mathbf{x}_i) - \mathbf{w}^T\Phi(\mathbf{x}_j) \geq 1 - \xi_{i,j}$$

$$\xi_{i,j} \geq 0$$

$$\forall i,j : \boldsymbol{x}_i \prec \boldsymbol{x}_j$$

Note

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

Support Vector Ranking

Support vector classification on pairs

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

$$\text{subject to}$$

$$y_{i,j}\mathbf{w}^T \underbrace{(\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j))}_{\Phi(\mathbf{x}_{ij})} \geq 1 - \xi_{i,j}$$

$$\xi_{i,j} \geq 0$$

$$\forall i,j : \boldsymbol{x}_i \prec \boldsymbol{x}_j$$

- where labels are always positive $y_{i,j} = 1$

Support Vector Ranking

Decision function

$$f(\boldsymbol{x}) = \mathbf{w}^T \Phi(\mathbf{x})$$

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