Support Vector Machines

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

Ch. 14 of I2ML (Secs. 14.4, 14.7 – 14.9, and 14.11 – 14.14 excluded)

Relaxing the Constraints

- In practice, a separating hyperplane may not exist, possibly due to the fact that the data is not linearly separable or a high noise level which causes a large overlap of the classes.
- Even if a separating hyperplane exists, it is not always the best solution to the classification problem when there exist outliers in the data.
  - A mislabeled example can become an outlier which affects the location of the separating hyperplane.

Slack Variables

- A soft-margin SVM allows for the possibility of violating the inequality constraints

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1$$

by introducing slack variables

$$\xi_t \geq 0, \quad t = 1, \ldots, N$$

which store the deviation from the margin.

- Relaxed separation constraints:

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \xi_t$$

Penalty

- By making ξ_t large enough, the constraint on $(\mathbf{x}^t, r^t)$ can always be met.
- In order not to obtain the trivial solution where all ξ_t take on large values, we should penalize them in the objective function.
- Three cases for ξ_t:
  - $\xi_t = 0$: no problem with $\mathbf{x}^t$ (no penalty)
  - $0 < \xi_t < 1$: $\mathbf{x}^t$ lies on the right side of the hyperplane but in the margin (small penalty)
  - $\xi_t > 1$: $\mathbf{x}^t$ lies on the wrong side of the hyperplane (large penalty)
- Number of misclassifications: $\#\{\xi_t > 1\}$
- Number of nonseparable instances: $\#\{\xi_t > 0\}$
- Soft error as additional penalty term:

$$\sum_{t=1}^{N} \xi_t$$

Primal Optimization Problem

- Primal optimization problem:

$$\begin{aligned} \underset{\mathbf{w},\, w_0,\, \{\xi_t\}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{t=1}^{N}\xi_t \\ \text{subject to} \quad & r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \xi_t, \quad \forall t \\ & \xi_t \geq 0, \quad \forall t \end{aligned}$$

  where $C \geq 0$ is a regularization parameter (which trades off model complexity in terms of the number of support vectors and data misfit in terms of the number of nonseparable points).
- Both the misclassified instances and the ones in the margin are penalized for better generalization, though the latter ones would be correctly classified during testing.
- For the same reason as before, we will resort to the dual problem.

Lagrangian

- Lagrangian:

$$
\begin{aligned}
&\mathcal{L}(\mathbf{w}, w_0, \{\xi_t\}, \{\alpha_t\}, \{\mu_t\}) \\
&\quad = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{t=1}^{N}\xi_t - \sum_{t=1}^{N}\alpha_t\Big[r^t(\mathbf{w}^T\mathbf{x}^t + w_0) - 1 + \xi_t\Big] - \sum_{t=1}^{N}\mu_t\xi_t
\end{aligned}
$$

where the new Lagrange multipliers $\mu_t \geq 0$.

Eliminating Primal Variables

- Setting the gradients of $\mathcal{L}$ w.r.t. $\mathbf{w}$, w_0, and $\{\xi_t\}$ to 0:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{w} = \sum_t \alpha_t r^t \mathbf{x}^t \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \quad \Rightarrow \quad \sum_t \alpha_t r^t = 0 \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_t} = 0 \quad \Rightarrow \quad \mu_t = C - \alpha_t, \quad \forall t \tag{6}$$

- Plugging (4), (5), and (6) into $\mathcal{L}$ gives the objective function G to maximize for the dual problem:

$$G(\{\alpha_t\}) = -\frac{1}{2} \sum_t \sum_{t'} \alpha_t \alpha_{t'} r^t r^{(t')} (\mathbf{x}^t)^T \mathbf{x}^{(t')} + \sum_t \alpha_t$$

- Since $\mu_t \geq 0$, $\forall t$, (6) implies that $0 \leq \alpha_t \leq C$, $\forall t$.

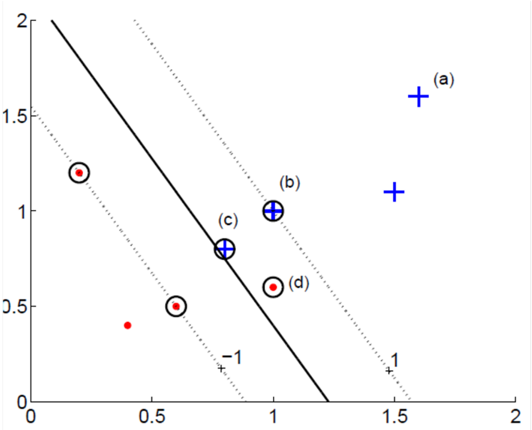
Dual Optimization Problem

- Dual optimization problem:

$$
\begin{aligned}
\underset{\{\alpha_t\}}{\text{maximize}} \quad & \sum_t \alpha_t - \frac{1}{2}\sum_t \sum_{t'} \alpha_t \alpha_{t'} r^t r^{(t')} (\mathbf{x}^t)^T \mathbf{x}^{(t')} \\
\text{subject to} \quad & \sum_t \alpha_t r^t = 0 \\
& 0 \leq \alpha_t \leq C, \quad \forall t
\end{aligned}
$$

- Similar to the hard-margin case (i.e., the separable case), instances that are not support vectors (lie on the correct side of the boundary with sufficient margin) vanish with $\alpha_t = 0$.
- The primal variables $\mathbf{w}$ and w_0 can be computed similarly based on the SVs.
  - The SVs have their $\alpha_t > 0$ and they define $\mathbf{w}$.
  - Of SVs, those whose $\alpha_t < C$ are the ones that are on the margin which can be used to calculate w_0 (they have $\xi_t = 0$ and satisfy $r^t(\mathbf{w}^T\mathbf{x}^t + w_0) = 1$).
  - Those instances that are in the margin or misclassified have their $\alpha_t = C$.

Soft-Margin Support Vector Machine



Support Vectors

- The nonseparable instances that we store as support vectors are the instances that we would have trouble correctly classifying if they were not in the training set; they would either be misclassified or classified correctly but not with enough confidence.
- An important result from Vapnik's statistical learning theory is that the expected test error rate has an upper bound which depends on the number of support vectors:

$$E_N[P(\text{error})] \leq \frac{E_N[\#\text{ of SVs}]}{N}$$

  where $E_N[\cdot]$ denotes the expectation over training sets of size N.
- It shows that the error rate depends on the number of support vectors and not on the input dimensionality.

Hinge Loss

- In the soft-margin SVM, we define an error ξ_t if the instance $(\mathbf{x}^t, r^t)$ is nonseparable, which can be described as a hinge loss as

$$L_{\text{hinge}}(y^t, r^t) = (1 - r^t y^t)_+ = \begin{cases} 0 & \text{if } r^t y^t \geq 1 \\ 1 - y^t r^t & \text{otherwise} \end{cases}$$

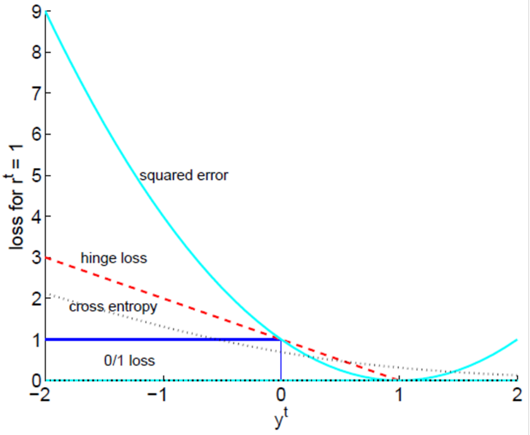
  where $y^t = \mathbf{w}^T \mathbf{x}^t + w_0$.

- The soft-margin SVM problem can be equivalently formulated as

$$\begin{aligned} \underset{\mathbf{w},\, w_0}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{t=1}^{N} (1 - r^t y^t)_+ \\ \text{subject to} \quad & y^t = \mathbf{w}^T \mathbf{x}^t + w_0, \quad \forall t \end{aligned}$$

- The hinge loss, again, reveals the nature of solution sparsity in SVM, i.e., predictions only depend on a subset of the training data.

More Loss Functions



Remark on SVMs

- The SVM problem can be case as convex programming problem (every local solution to a convex programming problem is a globally optimal solution), which is contrast to neural networks, where many local minima usually exist.
- In both training and testing, training data only appear in the form of dot products between vectors, which will become important later on.

Outline

Key Ideas of Kernel Methods

- Instead of defining a nonlinear model in the original (input) space, the problem is mapped to a new (feature) space by performing a nonlinear transformation using suitably chosen basis functions.
- A linear model is then applied in the new space.
- This approach can be used in both classification and regression problems.
- In the particular case of support vector machines, it leads to certain simplifications, where the basis functions are often defined implicitly via defining kernel functions directly.

Basis Functions

- Basis Functions:

$$\mathbf{z} = \phi(\mathbf{x}) \quad \text{where } z_j = \phi_j(\mathbf{x}),\ j = 1, \ldots, k$$

  mapping from the d-dimensional $\mathbf{x}$-space to the k-dimensional $\mathbf{z}$-space.

- Discriminant function:

$$\begin{aligned} g(\mathbf{z}) =& \mathbf{w}^T \mathbf{z} + w_0 \\ g(\mathbf{x}) =& \mathbf{w}^T \phi(\mathbf{x}) + w_0 = \sum_{j=1}^{k} w_j \phi_j(\mathbf{x}) + w_0 \end{aligned}$$

- Usually, $k \gg d$, N (in fact k can even be infinite). The dual form is preferred because its complexity depends on N but that of the primal form depends on k.

Primal Optimization Problem

- We use the general case of soft-margin nonlinear SVM because we have no guarantee that the problem is linearly separable in this new space.
- Primal optimization problem:

$$
\begin{aligned}
\underset{\mathbf{w},\, w_0,\, \{\xi_t\}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{t=1}^{N}\xi_t \\
\text{subject to} \quad & r^t(\mathbf{w}^T\phi(\mathbf{x}^t) + w_0) \geq 1 - \xi_t, \quad \forall t \\
& \xi_t \geq 0, \quad \forall t
\end{aligned}
$$

  where $C \geq 0$.

- We will resort to the dual problem.

Lagrangian

- Lagrangian:

$$\mathcal{L}(\mathbf{w}, \{\xi_t\}, \{\alpha_t\}, \{\mu_t\}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{t=1}^{N}\xi_t - \sum_{t=1}^{N}\alpha_t\Big[r^t(\mathbf{w}^T\phi(\mathbf{x}^t) + w_0) - 1 + \xi_t\Big] - \sum_{t=1}^{N}\mu_t\xi_t$$

where the Lagrange multipliers $\alpha_t,\ \mu_t \geq 0$.

Dual Optimization Problem – I

- Setting the gradients of $\mathcal{L}$ w.r.t. $\mathbf{w}$, w_0, and $\{\xi_t\}$ to 0:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{w} = \sum_t \alpha_t r^t \phi(\mathbf{x}^t) \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \quad \Rightarrow \quad \sum_t \alpha_t r^t = 0 \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_t} = 0 \quad \Rightarrow \quad \mu_t = C - \alpha_t, \quad \forall t \tag{9}$$

- Plugging (7) and (8) into $\mathcal{L}$ gives the objective function G for the dual problem:

$$G(\{\alpha_t\}) = -\frac{1}{2} \sum_t \sum_{t'} \alpha_t \alpha_{t'} r^t r^{(t')} \phi(\mathbf{x}^t)^T \phi(\mathbf{x}^{(t')}) + \sum_t \alpha_t$$

Dual Optimization Problem – II

- Dual optimization problem:

$$
\begin{aligned}
\underset{\{\alpha_t\}}{\text{maximize}} \quad & \sum_t \alpha_t - \frac{1}{2} \sum_t \sum_{t'} \alpha_t \alpha_{t'} r^t r^{(t')} \phi(\mathbf{x}^t)^T \phi(\mathbf{x}^{(t')}) \\
\text{subject to} \quad & \sum_t \alpha_t r^t = 0 \\
& 0 \leq \alpha_t \leq C, \ \forall t
\end{aligned}
$$

Kernel Functions – I

- In kernel SVM, we have $K(\mathbf{x}^t, \mathbf{x}^{(t')}) \equiv \phi(\mathbf{x}^t)^T \phi(\mathbf{x}^{(t')})$ which is a kernel function (a.k.a. positive definite kernel, Mercer kernel, or reproducing kernel).

$$\begin{aligned} \underset{\{\alpha_t\}}{\text{maximize}} \quad & \sum_t \alpha_t - \frac{1}{2} \sum_t \sum_{t'} \alpha_t \alpha_{t'} r^t r^{(t')} K(\mathbf{x}^t, \mathbf{x}^{(t')}) \\ \text{subject to} \quad & \sum_t \alpha_t r^t = 0 \\ & 0 \leq \alpha_t \leq C, \ \forall t \end{aligned}$$

- Instead of mapping two instances $\mathbf{x}^t$ and $\mathbf{x}^{(t')}$ to the $\mathbf{z}$-space and doing a dot product there, we directly apply the kernel function in the original $\mathbf{x}$-space.
- Kernel matrix (a.k.a. Gram matrix):

$$\mathbf{K} = \left[K(\mathbf{x}^t, \mathbf{x}^{(t')}) \right]_{t,t'=1}^{N}$$

which, like a covariance matrix, is symmetric and positive semidefinite.

Kernel Functions – II

- Solution:

$$\mathbf{w} = \sum_t \alpha_t r^t \mathbf{z}^t = \sum_{\mathbf{x}^t \in \mathcal{SV}} \alpha_t r^t \phi(\mathbf{x}^t)$$

- Discriminant function:

$$g(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + w_0 = \sum_{\mathbf{x}^t \in \mathcal{SV}} \alpha_t r^t \phi(\mathbf{x}^t)^T \phi(\mathbf{x}) + w_0 = \sum_{\mathbf{x}^t \in \mathcal{SV}} \alpha_t r^t K(\mathbf{x}^t, \mathbf{x}) + w_0$$

where the kernel function also shows up in the discriminant.

Some Common Kernel Functions – I

- Polynomial kernel:

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + 1)^q$$

where q is the degree.
E.g., when $q = 2$ and $d = 2$,

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') =& (\mathbf{x}^T \mathbf{x}' + 1)^2 \\ =& (x_1 x_1' + x_2 x_2' + 1)^2 \\ =& 1 + 2x_1 x_1' + 2x_2 x_2' + 2x_1 x_1' x_2 x_2' + (x_1)^2 (x_1')^2 + (x_2)^2 (x_2')^2 \end{aligned}$$

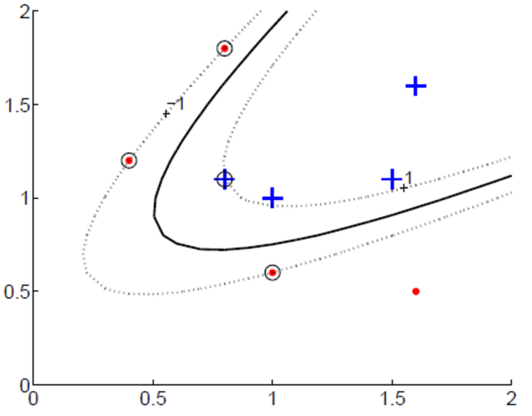
which corresponds to the inner product of the basis function

$$\phi(\mathbf{x}) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, (x_1)^2, (x_2)^2\right)^T$$

When $q = 1$, we have the linear kernel corresponding to the original formulation.

Some Common Kernel Functions – II

- Polynomial kernel of degree 2:



Some Common Kernel Functions – III

- Radial basis function (RBF) kernel (or Gaussian radial kernel):

$$K(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2s^2}\right]$$

  which is a spherical kernel where $\mathbf{x}'$ is the center and s, supplied by the user, defines the radius.
- The feature space of the RBF kernel has an infinite number of dimensions.
- It can be generalized to

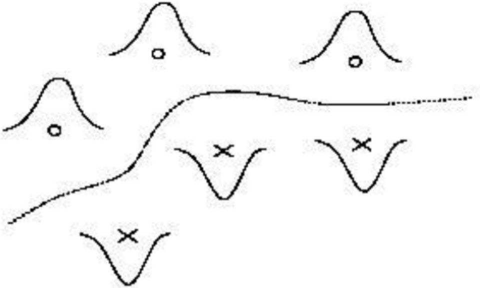
$$K(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{\mathcal{D}(\mathbf{x}, \mathbf{x}')}{2s^2}\right]$$

  where $\mathcal{D}(\cdot, \cdot)$ is some distance function.
- When taking the Mahalanobis distance, we have the Mahalanobis kernel

$$K(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{(\mathbf{x} - \mathbf{x}')^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{x}')}{2s^2}\right]$$

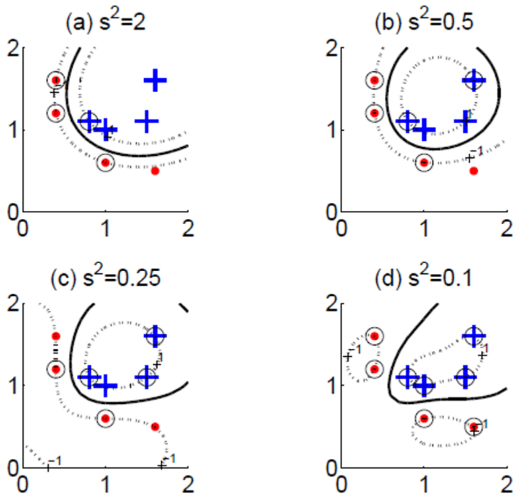
Some Common Kernel Functions – IV

- Discriminant function with RBF kernel: amounts to putting bumps of various sizes on the training set



Some Common Kernel Functions – V

- Gaussian kernel with different spread values, s^2:



Some Common Kernel Functions – VI

- Sigmoidal kernel (or hyperbolic tangent kernel):

  $$K(\mathbf{x}, \mathbf{x}') = \tanh(\kappa \mathbf{x}^T \mathbf{x}' + \theta)$$

  which, strictly speaking, is not positive semidefinite for certain parameter values κ and θ.
- This is similar to multilayer perceptrons that we discussed in last lecture.

Outline

t_2 Loss Function

- We start with a linear model for regression as

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

and we have used the squared loss in ordinary linear regression

$$E_2^t(r^t, f(\mathbf{x}^t)) = |r^t - f(\mathbf{x}^t)|^2$$

- Total loss:

$$E_2 = \sum_t E_2^t(r^t, f(\mathbf{x}^t)) = \sum_t |r^t - f(\mathbf{x}^t)|^2$$

- Squared regression (or least squares regression):

$$\underset{\mathbf{w},\, w_0}{\text{minimize}} \quad \frac{1}{N} \sum_{t=1}^{N} |r^t - f(\mathbf{x}^t)|^2$$

ϵ-Insensitive Loss Function – I

- In order for the sparseness property of support vectors in SVM for classification to carry over to regression, we do not use the squared loss but the ϵ-insensitive loss function:

$$E_\epsilon^t(r^t, f(\mathbf{x}^t)) = (|r^t - f(\mathbf{x}^t)| - \epsilon)_+ = \begin{cases} 0 & \text{if } |r^t - f(\mathbf{x}^t)| \leq \epsilon \\ |r^t - f(\mathbf{x}^t)| - \epsilon & \text{otherwise} \end{cases}$$

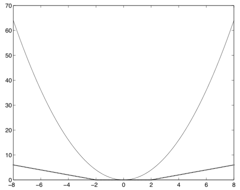
- Two characteristics:
  - Errors are tolerated up to a threshold of ϵ, i.e., no loss for point lying inside an ϵ-tube around the prediction.
  - Errors beyond ϵ have a linear (rather than quadratic) effect so that the model is more more tolerant to noise and robust against noise.
- Total loss:

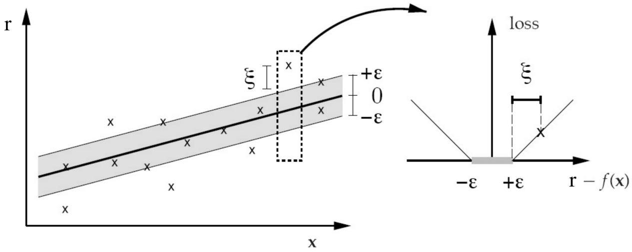
$$E_\epsilon = \sum_t E_\epsilon^t(r^t, f(\mathbf{x}^t)) = \sum_t (|r^t - f(\mathbf{x}^t)| - \epsilon)_+$$

- Tube regression:

$$\underset{\mathbf{w},\, w_0}{\text{minimize}} \quad \frac{1}{N} \sum_{t=1}^{N} (|r^t - f(\mathbf{x}^t)| - \epsilon)_+$$

ϵ-Insensitive Loss Function – II





Support Vector Regression

- Support vector (machine) regression (SVR) is given as

$$\underset{\mathbf{w},\, w_0}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t (|r^t - f(\mathbf{x}^t)| - \epsilon)_+$$

  where C trades off the model complexity (i.e., the flatness of the model) and data misfit.
- The value of ϵ determines the width of the tube (a smaller value indicates a lower tolerance for error) and also affects the number of support vectors and, consequently, the solution sparsity.
  - If ϵ is decreased, the boundary of the tube is shifted inward. Therefore, more datapoints are around the boundary indicating more support vectors.
  - Similarly, increasing ϵ will result in fewer points around the boundary.
- A convex problem, but not a standard QP.
- We will rewrite it to a form similar to SVM which can be QP-solvable.

Primal Optimization Problem

- We introduce slack variables ξ_t^+ and ξ_t^- to account for deviations out of the ϵ-zone.
- Primal optimization problem:

$$
\begin{aligned}
\underset{\mathbf{w},\, w_0,\, \{\xi_t^+\},\, \{\xi_t^-\}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t (\xi_t^+ + \xi_t^-) \\
\text{subject to} \quad & r^t - (\mathbf{w}^T\mathbf{x}^t + w_0) \le \epsilon + \xi_t^+, \quad \forall t \\
& (\mathbf{w}^T\mathbf{x}^t + w_0) - r^t \le \epsilon + \xi_t^-, \quad \forall t \\
& \xi_t^+, \xi_t^- \ge 0, \quad \forall t
\end{aligned}
$$

  which is a standard QP.
- Two types of slack variables:
  - ξ_t^+: for positive deviation such that $r^t - (\mathbf{w}^T\mathbf{x}^t + w_0) > \epsilon$.
  - ξ_t^-: for negative deviation such that $(\mathbf{w}^T\mathbf{x}^t + w_0) - r^t > \epsilon$.
- If $r^t - (\mathbf{w}^T\mathbf{x}^t + w_0) \le \epsilon$ and $(\mathbf{w}^T\mathbf{x}^t + w_0) - r^t \le \epsilon$, then $\xi_t^+ = \xi_t^- = 0$, contributing no cost to the objective function.

Lagrangian

- Similar to SVM for classification, the optimization problem for SVR can also be rewritten in the dual form.
- Lagrangian:

$$
\begin{aligned}
&\mathcal{L}(\mathbf{w}, w_0, \{\xi_t^+\}, \{\xi_t^-\}, \{\alpha_t^+\}, \{\alpha_t^-\}, \{\mu_t^+\}, \{\mu_t^-\}) \\
=&\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t(\xi_t^+ + \xi_t^-) \\
&-\sum_t \alpha_t^+\left[\epsilon + \xi_t^+ - r^t + (\mathbf{w}^T\mathbf{x}^t + w_0)\right] - \sum_t \alpha_t^-\left[\epsilon + \xi_t^- + r^t - (\mathbf{w}^T\mathbf{x}^t + w_0)\right] \\
&-\sum_t(\mu_t^+\xi_t^+ + \mu_t^-\xi_t^-)
\end{aligned}
$$

where $\alpha_t^+, \alpha_t^-, \mu_t^+, \mu_t^- > 0$.

Eliminating Primal Variables

- Setting the gradients of $\mathcal{L}$ w.r.t. $\mathbf{w}$, w_0, $\{\xi_t^+\}$, and $\{\xi_t^-\}$ to 0:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{w} = \sum_t (\alpha_t^+ - \alpha_t^-)\mathbf{x}^t \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = 0 \quad \Rightarrow \quad \sum_t (\alpha_t^+ - \alpha_t^-) = 0 \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_t^+} = 0 \quad \Rightarrow \quad \mu_t^+ = C - \alpha_t^+, \quad \forall t \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_t^-} = 0 \quad \Rightarrow \quad \mu_t^- = C - \alpha_t^-, \quad \forall t \tag{13}$$

- Plugging (9), (10), (11), and (12) into $\mathcal{L}$ gives the objective function G for the dual problem:

$$G(\{\alpha_t^+\}, \{\alpha_t^-\}) = -\frac{1}{2}\sum_t \sum_{t'} (\alpha_t^+ - \alpha_t^-)(\alpha_{t'}^+ - \alpha_{t'}^-)(\mathbf{x}^t)^T \mathbf{x}^{(t')}$$
$$- \epsilon \sum_t (\alpha_t^+ + \alpha_t^-) + \sum_t r^t (\alpha_t^+ - \alpha_t^-)$$

Dual Optimization Problem – I

- Dual optimization problem:

$$
\begin{aligned}
\underset{\{\alpha_t^+\},\{\alpha_t^-\}}{\text{maximize}} \quad & -\frac{1}{2}\sum_t \sum_{t'} (\alpha_t^+ - \alpha_t^-)(\alpha_{t'}^+ - \alpha_{t'}^-)(\mathbf{x}^t)^T \mathbf{x}^{(t')} \\
& - \epsilon \sum_t (\alpha_t^+ + \alpha_t^-) + \sum_t r^t (\alpha_t^+ - \alpha_t^-) \\
\text{subject to} \quad & \sum_t (\alpha_t^+ - \alpha_t^-) = 0 \\
& 0 \le \alpha_t^+ \le C, \ \forall t \\
& 0 \le \alpha_t^- \le C, \ \forall t
\end{aligned}
$$

- Instances in the ϵ-tube ($\alpha_t^+ = \alpha_t^- = 0$) are instances fitted with enough precision.
- The support vectors satisfy either $\alpha_t^+ > 0$ or $\alpha_t^- > 0$ and are of two types.
  - instances on the boundary of the ϵ-tube (either $0 < \alpha_t^+ < C$ or $0 < \alpha_t^- < C$), and we use these to calculate w_0
  - instances outside the ϵ-tube are instances for which we do not have a good fit (either $\alpha_t^+ = C$ or $\alpha_t^- = C$)

Dual Optimization Problem – II

- We have the fitted line as a weighted sum of the support vectors:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \sum_{\mathbf{x}^t \in \mathcal{SV}} (\alpha_t^+ - \alpha_t^-)(\mathbf{x}^t)^T\mathbf{x} + w_0$$

- Due to the sparseness property of the ϵ-insensitive loss function, only a small fraction of the training instances are support vectors which are used in defining the regression function (like the discriminant function for classification).
- Nonlinear (kernel) extension is possible by introducing appropriate kernel functions.

SVR