E0 270 Machine Learning **Lecture 5 (Jan 22, 2013)**

Support Vector Machines for Classification and Regression

Lecturer: Shivani Agarwal

Disclaimer: These notes are a *brief* summary of the topics covered in the lecture. They are not a substitute for the full lecture.

Outline

- Linearly separable data: Hard margin SVMs
- Non-linearly separable data: Soft margin SVMs
- Loss minimization view
- Support vector regression (SVR)

1 Linearly Separable Data: Hard Margin SVMs

In this lecture we consider linear **support vector machines** (SVMs); we will consider nonlinear extensions in the next lecture. Let $\mathcal{X} = \mathbb{R}^d$, and consider a binary classification task with $\mathcal{Y} = \widehat{\mathcal{Y}} = \{\pm 1\}$. A training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (\mathbb{R}^d \times \{\pm 1\})^m$ is said to be **linearly separable** if there exists a linear classifier $h_{\mathbf{w},b}(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$ which classifies all examples in S correctly, i.e. for which $y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0 \ \forall i \in [m]$. For example, Figure 1 (left) shows a training sample in $\mathbb{R}^2$ that is linearly separable, together with two possible linear classifiers that separate the data correctly (note that the decision surface of a linear classifier in 2 dimensions is a line, and more generally in $d > 2$ dimensions is a hyperplane). Which of the two classifiers is likely to give better generalization performance?

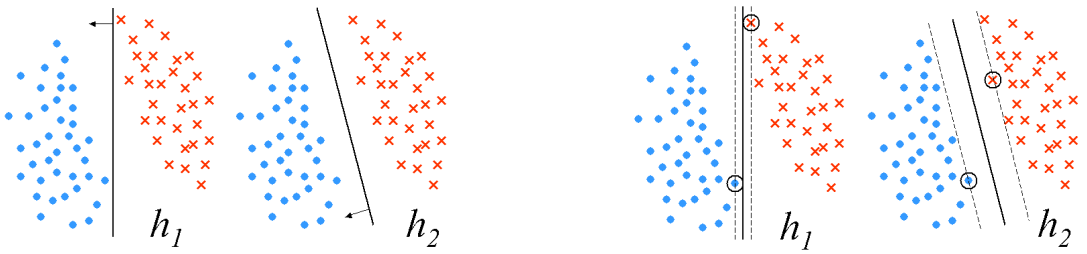

Figure 1: **Left:** A linearly separable data set, with two possible linear classifiers that separate the data. Blue circles represent class label 1 and red crosses -1; the arrow represents the direction of positive classification. **Right:** The same data set and classifiers, with margin of separation shown.

Although both classifiers separate the data, the distance or margin with which the separation is achieved is different; this is shown in Figure 1 (right). For the rest of this section, assume that the training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ is linearly separable; in this setting, the SVM algorithm selects the **maximum margin** linear classifier, i.e. the linear classifier that separates the training data with the largest margin. More precisely, define the **(geometric) margin of a linear classifier** $h_{\mathbf{w},b}(\mathbf{x}) = \textbf{sign}(\mathbf{w}^\top \mathbf{x} + b)$ **on an example** $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ as

$$\gamma_i = \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|_2} \,. \tag{1}$$

Note that the distance of $\mathbf{x}_i$ from the hyperplane $\mathbf{w}^\top \mathbf{x}+b = 0$ is given by $\frac{|\mathbf{w}^\top \mathbf{x}_i+b|}{\|\mathbf{w}\|_2}$; therefore the above margin on $(\mathbf{x}_i, y_i)$ is simply a signed version of this distance, with a positive sign if the example is classified correctly and negative otherwise. The **(geometric) margin of $h_{\mathbf{w},b}$ on the sample** $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ is then defined as the minimal margin on examples in S:

$$\gamma = \min_{i \in [m]} \gamma_i \,. \tag{2}$$

Given a linearly separable training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (\mathbb{R}^d \times \{\pm 1\})^m$, the **hard margin SVM** algorithm finds a linear classifier that maximizes the above margin on S. In particular, any linear classifier that separates S correctly will have margin $\gamma > 0$; without loss of generality, we can represent any such classifier by some $(\mathbf{w}, b)$ such that

$$\min_{i \in [m]} y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1 \,. \tag{3}$$

The margin of such a classifier on S then becomes simply

$$\gamma = \min_{i \in [m]} \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|_2} = \frac{1}{\|\mathbf{w}\|_2} \,. \tag{4}$$

Thus, maximizing the margin becomes equivalent to minimizing the norm $\|\mathbf{w}\|_2$ subject to the constraints in Eq. (3), which can be written as the following optimization problem:

$$\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|_2^2 \tag{5}$$

subject to

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \ \geq \ 1 \,, \quad i = 1, \ldots, m. \tag{6}$$

This is a convex **quadratic program** (QP) and can in principle be solved directly. However it is useful to consider the dual of the above problem, which sheds light on the structure of the solution and also facilitates the extension to nonlinear classifiers which we will see in the next lecture. Note that the above optimization problem involves a convex objective and convex (in fact linear) inequality constraints, so that strong duality holds and therefore solving the dual problem is equivalent to solving the above primal problem. Introducing Lagrange multipliers $\alpha_i \geq 0$ $(i = 1, \ldots, m)$ for the inequality constraints above gives the Lagrangian function

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^m \alpha_i (1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) \,. \tag{7}$$

We would like to maximize this w.r.t. $\boldsymbol{\alpha}$ and minimize w.r.t. $(\mathbf{w}, b)$. Setting the derivatives of $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$ w.r.t. $\mathbf{w}$ and b to zero gives:

$$\mathbf{w} \ = \ \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \tag{8}$$

$$\sum_{i=1}^m \alpha_i y_i \ = \ 0 \,. \tag{9}$$

This leads to the following dual problem:

$$\max_{\boldsymbol{\alpha}} \ -\frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^\top \mathbf{x}_j) + \sum_{i=1}^m \alpha_i \tag{10}$$

subject to

$$\sum_{i=1}^m \alpha_i y_i \ = \ 0 \tag{11}$$

$$\alpha_i \ \geq \ 0 \,, \quad i = 1, \ldots, m. \tag{12}$$

This is again a convex QP (in the m variables α_i) and can be solved efficiently using numerical optimization methods. On obtaining the solution $\widehat{\boldsymbol{\alpha}}$ to the above dual problem, the weight vector $\widehat{\mathbf{w}}$ corresponding to the maximal margin classifier can be obtained via Eq. (8):

$$\widehat{\mathbf{w}} = \sum_{i=1}^{m} \widehat{\alpha}_i y_i \mathbf{x}_i \,.$$

Now, by the complementary slackness condition in the KKT conditions, we have for each $i \in [m]$,

$$\widehat{\alpha}_i \big(1 - y_i(\widehat{\mathbf{w}}^\top \mathbf{x}_i + \widehat{b})\big) = 0 \,.$$

This gives

$$\widehat{\alpha}_i > 0 \implies 1 - y_i(\widehat{\mathbf{w}}^\top \mathbf{x}_i + \widehat{b}) = 0 \,.$$

In other words, $\widehat{\alpha}_i$ is positive only for training points $\mathbf{x}_i$ that lie on the margin, i.e. that are closest to the separating hyperplane; these points are called the **support vectors**. For all other training points $\mathbf{x}_i$, we have $\widehat{\alpha}_i = 0$. Thus the solution for $\widehat{\mathbf{w}}$ can be written as a linear combination of just the support vectors; specifically, if we define

$$\mathrm{SV} = \big\{ i \in [m] : \widehat{\alpha}_i > 0 \big\} \,,$$

then we have

$$\widehat{\mathbf{w}} = \sum_{i \in \mathrm{SV}} \widehat{\alpha}_i y_i \mathbf{x}_i \,.$$

Moreover, for all $i \in \mathrm{SV}$, we have

$$1 - y_i(\widehat{\mathbf{w}}^\top \mathbf{x}_i + \widehat{b}) = 0 \quad \text{or} \quad y_i - (\widehat{\mathbf{w}}^\top \mathbf{x}_i + \widehat{b}) = 0 \,.$$

This allows us to obtain $\widehat{b}$ from any of the support vectors; in practice, for numerical stability, one generally averages over all the support vectors, giving

$$\widehat{b} = \frac{1}{|\mathrm{SV}|} \sum_{i \in \mathrm{SV}} (y_i - \widehat{\mathbf{w}}^\top \mathbf{x}_i) \,.$$

In order to classify a new point $\mathbf{x} \in \mathbb{R}^d$ using the learned classifier, one then computes

$$h_{\widehat{\mathbf{w}},\widehat{b}}(\mathbf{x}) = \mathrm{sign}(\widehat{\mathbf{w}}^\top \mathbf{x} + \widehat{b}) = \mathrm{sign}\bigg(\sum_{i \in \mathrm{SV}} \widehat{\alpha}_i y_i (\mathbf{x}_i^\top \mathbf{x}) + \widehat{b} \bigg) \,. \tag{13}$$

2 Non-Linearly Separable Data: Soft Margin SVMs

The above derivation assumed the existence of a linear classifier that can correctly classify all examples in a given training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$. But what if the sample is not linearly separable?

In this case, one needs to allow for the possibility of errors in classification. This is usually done by relaxing the constraints in Eq. (6) through the introduction of slack variables $\xi_i \geq 0$ ($i = 1, \ldots, m$), and requiring only that

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \,, \quad i = 1, \ldots, m. \tag{14}$$

An extra cost for errors can be assigned as follows:

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{m} \xi_i \tag{15}$$

subject to

$$\begin{aligned} y_i(\mathbf{w}^\top \mathbf{x}_i + b) &\geq 1 - \xi_i \,, \quad i = 1, \ldots, m && (16) \\ \xi_i &\geq 0 \,, \quad i = 1, \ldots, m. && (17) \end{aligned}$$

Thus, whenever $y_i(\mathbf{w}^\top \mathbf{x}_i + b) < 1$, we pay an associated cost of $C\xi_i = C(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$ in the objective function; a classification error occurs when $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0$, or equivalently when $\xi_i \geq 1$. The parameter

$C > 0$ controls the tradeoff between increasing the margin (minimizing $\|\mathbf{w}\|_2$) and reducing the errors (minimizing ξ_i): a large value of C keeps the errors small at the cost of a reduced margin; a small value of C allows for more errors while increasing the margin on the remaining examples. Forming the dual of the above problem as before leads to the same convex QP as in the linearly separable case, except that the constraints in Eq. (12) are replaced by

$$0 \leq \alpha_i \leq C \qquad i = 1, \ldots, m. \tag{18}$$

The solution for $\widehat{\mathbf{w}}$ is obtained similarly to the linearly separable case:

$$\widehat{\mathbf{w}} = \sum_{i=1}^{m} \widehat{\alpha}_i y_i \mathbf{x}_i \,.$$

In this case, the complementary slackness conditions yield for each $i \in [m]$:

$$\begin{aligned} \widehat{\alpha}_i\big(1 - \widehat{\xi}_i - y_i(\widehat{\mathbf{w}}^\top \mathbf{x}_i + \widehat{b})\big) &= 0 \\ (C - \widehat{\alpha}_i)\widehat{\xi}_i &= 0\,. \end{aligned}$$

This gives

$$\begin{aligned} \widehat{\alpha}_i > 0 \quad &\Longrightarrow \quad 1 - \widehat{\xi}_i - y_i(\widehat{\mathbf{w}}^\top \mathbf{x}_i + \widehat{b}) = 0 \\ \widehat{\alpha}_i < C \quad &\Longrightarrow \quad \widehat{\xi}_i = 0\,. \end{aligned}$$

In particular, this gives

$$0 < \widehat{\alpha}_i < C \quad \Longrightarrow \quad 1 - y_i(\widehat{\mathbf{w}}^\top \mathbf{x}_i + \widehat{b}) = 0\,;$$

these are the points on the margin. Thus here we have three types of support vectors with $\widehat{\alpha}_i > 0$ (see Figure 2):

$$\begin{aligned} \mathrm{SV}_1 &= \big\{i \in [m] : 0 < \widehat{\alpha}_i < C\big\} \\ \mathrm{SV}_2 &= \big\{i \in [m] : \widehat{\alpha}_i = C,\ \widehat{\xi}_i < 1\big\} \\ \mathrm{SV}_3 &= \big\{i \in [m] : \widehat{\alpha}_i = C,\ \widehat{\xi}_i \geq 1\big\}\,. \end{aligned}$$

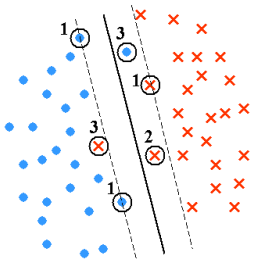

Figure 2: Three types of support vectors in the non-separable case.

SV_1 contains margin support vectors ($\widehat{\xi}_i = 0$; these lie on the margin and are correctly classified); SV_2 contains non-margin support vectors with $0 < \widehat{\xi}_i < 1$ (these are correctly classified, but lie within the margin); SV_3 contains non-margin support vectors with $\widehat{\xi}_i \geq 1$ (these correspond to classification errors).

Let

$$\mathrm{SV} = \mathrm{SV}_1 \cup \mathrm{SV}_2 \cup \mathrm{SV}_3\,.$$

Then we have

$$\widehat{\mathbf{w}} = \sum_{i \in \mathrm{SV}} \widehat{\alpha}_i y_i \mathbf{x}_i\,.$$

Moreover, we can use the margin support vectors in SV_1 to compute $\widehat{b}$:

$$\widehat{b} = \frac{1}{|\mathrm{SV}_1|} \sum_{i \in \mathrm{SV}_1} (y_i - \widehat{\mathbf{w}}^\top \mathbf{x}_i)\,.$$

The above formulation of the SVM algorithm for the general (nonseparable) case is often called the **soft margin SVM**.

3 Loss Minimization View

An alternative motivation for the (soft margin) SVM algorithm is in terms of minimizing the **hinge loss** on the training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$. Specifically, define $\ell_{\text{hinge}} : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$ as

$$\ell_{\text{hinge}}(y, f) = (1 - yf)_+ \,, \tag{19}$$

where $z_+ = \max(0, z)$. This loss is convex in f and upper bounds the 0-1 loss much as the logistic loss does. Now consider learning a linear classifier that minimizes the empirical hinge loss, plus an L_2 regularization term:

$$\min_{\mathbf{w}, b} \; \frac{1}{m} \sum_{i=1}^{m} \big(1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big)_+ + \lambda \|\mathbf{w}\|_2^2 \,. \tag{20}$$

Introducing slack variables ξ_i $(i = 1, \ldots, m)$, we can re-write this as

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \; \frac{1}{m} \sum_{i=1}^{m} \xi_i + \lambda \|\mathbf{w}\|_2^2 \tag{21}$$

subject to

$$\begin{aligned} \xi_i &\geq 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \,, \quad i = 1, \ldots, m & (22) \\ \xi_i &\geq 0 \,, \quad i = 1, \ldots, m. & (23) \end{aligned}$$

This is equivalent to the soft margin SVM (with $C = \frac{1}{2\lambda m}$); in other words, the soft margin SVM algorithm derived earlier effectively performs L_2-regularized empirical hinge loss minimization (with $\lambda = \frac{1}{2Cm}$)!

4 Support Vector Regression (SVR)

Consider now a regression problem with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \widehat{\mathcal{Y}} = \mathbb{R}$. Given a training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (\mathbb{R}^d \times \mathbb{R})^m$, the **support vector regression** (SVR) algorithm minimizes an L_2-regularized form of the **ϵ-insensitive loss** $\ell_\epsilon : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, defined as

$$\begin{aligned} \ell_\epsilon(y, \widehat{y}) &= \big(|\widehat{y} - y| - \epsilon\big)_+ & (24) \\ &= \begin{cases} 0 & \text{if } |\widehat{y} - y| \leq \epsilon \\ |\widehat{y} - y| - \epsilon & \text{otherwise.} \end{cases} & (25) \end{aligned}$$

This yields

$$\min_{\mathbf{w}, b} \; \frac{1}{m} \sum_{i=1}^{m} \big(\big|(\mathbf{w}^\top \mathbf{x}_i + b) - y_i\big| - \epsilon\big)_+ + \lambda \|\mathbf{w}\|_2^2 \,. \tag{26}$$

Introducing slack variables ξ_i, ξ_i^* $(i = 1, \ldots, m)$ and writing $\lambda = \frac{1}{2Cm}$ for appropriate $C > 0$, we can re-write this as

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \; \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \tag{27}$$

subject to

$$\begin{aligned} \xi_i &\geq y_i - (\mathbf{w}^\top \mathbf{x}_i + b) - \epsilon \,, \quad i = 1, \ldots, m & (28) \\ \xi_i^* &\geq (\mathbf{w}^\top \mathbf{x}_i + b) - y_i - \epsilon \,, \quad i = 1, \ldots, m & (29) \\ \xi_i, \xi_i^* &\geq 0 \,, \quad i = 1, \ldots, m. & (30) \end{aligned}$$

This is again a convex QP that can in principle be solved directly; again, it useful to consider the dual, which helps to understand the structure of the solution and facilitates the extension to nonlinear SVR. We

leave the details as an exercise; the resulting dual problem has the following form:

$$\max_{\boldsymbol{\alpha}} \quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(\mathbf{x}_i^\top\mathbf{x}_j)+\sum_{i=1}^{m}y_i(\alpha_i-\alpha_i^*)-\epsilon\sum_{i=1}^{m}(\alpha_i+\alpha_i^*) \tag{31}$$

subject to

$$\sum_{i=1}^{m}(\alpha_i-\alpha_i^*) = 0 \tag{32}$$

$$0\le\alpha_i\le C\,, \qquad i=1,\ldots,m. \tag{33}$$

$$0\le\alpha_i^*\le C\,, \qquad i=1,\ldots,m. \tag{34}$$

This is again a convex QP (in the $2m$ variables α_i,α_i^*); the solution $\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\alpha}}^*$ can be used to find the solution $\widehat{\mathbf{w}}$ to the primal problem as follows:

$$\widehat{\mathbf{w}}=\sum_{i=1}^{m}(\widehat{\alpha}_i-\widehat{\alpha}_i^*)\mathbf{x}_i\,.$$

In this case, the complementary slackness conditions yield for each $i\in[m]$:

$$\begin{aligned}\widehat{\alpha}_i\big(\widehat{\xi}_i-y_i+(\widehat{\mathbf{w}}^\top\mathbf{x}_i+\widehat{b})+\epsilon\big) &= 0\\ \widehat{\alpha}_i^*\big(\widehat{\xi}_i^*+y_i-(\widehat{\mathbf{w}}^\top\mathbf{x}_i+\widehat{b})+\epsilon\big) &= 0\\ (C-\widehat{\alpha}_i)\widehat{\xi}_i &= 0\\ (C-\widehat{\alpha}_i^*)\widehat{\xi}_i^* &= 0\,.\end{aligned}$$

Analysis of these conditions shows that for each i, either $\widehat{\alpha}_i$ or $\widehat{\alpha}_i^*$ (or both) must be zero. For points inside the **ϵ-tube** around the learned linear function, i.e. for which $|(\widehat{\mathbf{w}}^\top\mathbf{x}_i+\widehat{b})-y_i|<\epsilon$, we have both $\widehat{\alpha}_i=\widehat{\alpha}_i^*=0$. The remaining points constitute two types of support vectors:

$$\begin{aligned}\mathrm{SV}_1 &= \big\{i\in[m]: 0<\widehat{\alpha}_i<C \text{ or } 0<\widehat{\alpha}_i^*<C\big\}\\ \mathrm{SV}_2 &= \big\{i\in[m]: \widehat{\alpha}_i=C \text{ or } \widehat{\alpha}_i^*=C\big\}\,.\end{aligned}$$

SV_1 contains support vectors on the tube boundary (with $\widehat{\xi}_i=\widehat{\xi}_i^*=0$); SV_2 contains support vectors outside the tube (with $\widehat{\xi}_i>0$ or $\widehat{\xi}_i^*>0$). Taking

$$\mathrm{SV}=\mathrm{SV}_1\cup\mathrm{SV}_2\,,$$

we then have

$$\widehat{\mathbf{w}}=\sum_{i\in\mathrm{SV}}(\widehat{\alpha}_i-\widehat{\alpha}_i^*)\mathbf{x}_i\,.$$

As before, the boundary support vectors in SV_1 can be used to compute $\widehat{b}$, which gives

$$\widehat{b}=\frac{1}{|\mathrm{SV}_1|}\Bigg(\sum_{i:0<\widehat{\alpha}_i<C}(y_i-\widehat{\mathbf{w}}^\top\mathbf{x}_i-\epsilon)+\sum_{i:0<\widehat{\alpha}_i^*<C}(\widehat{\mathbf{w}}^\top\mathbf{x}_i-y_i-\epsilon)\Bigg)\,.$$

The prediction for a new point $\mathbf{x}\in\mathbb{R}^d$ is then made via

$$f_{\widehat{\mathbf{w}},\widehat{b}}(\mathbf{x})=\widehat{\mathbf{w}}^\top\mathbf{x}+\widehat{b}=\sum_{i\in\mathrm{SV}}(\widehat{\alpha}_i-\widehat{\alpha}_i^*)(\mathbf{x}_i^\top\mathbf{x})+\widehat{b}\,.$$

In practice, the parameter C in SVM and the parameters C and ϵ in SVR are generally selected by cross-validation on the training sample (or using a separate validation set). An alternative parametrization of the SVM and SVR optimization problems, termed ν-SVM and ν-SVR, makes use of a different parameter ν that directly bounds the fraction of training examples that end up as support vectors.

Exercise. Derive the dual of the SVR optimization problem above.

Exercise. Derive an alternative formulation of the SVR optimization problem that makes use of a single slack variable ξ_i for each data point rather than two slack variables ξ_i,ξ_i^*. Show that this leads to the same solution as above.

Exercise. Derive a regression algorithm that given a training sample S, minimizes on S the L_2-regularized **absolute loss** $\ell_{\mathrm{abs}}:\mathbb{R}\times\mathbb{R}\to\mathbb{R}_+$, given by $\ell_{\mathrm{abs}}(y,\widehat{y})=|\widehat{y}-y|$, over all linear functions.