15 Support Vector Machines

In this chapter and the next we discuss a very useful machine learning tool: the support vector machine paradigm (SVM) for learning linear predictors in high dimensional feature spaces. The high dimensionality of the feature space raises both sample complexity and computational complexity challenges.

The SVM algorithmic paradigm tackles the sample complexity challenge by searching for "large margin" separators. Roughly speaking, a halfspace separates a training set with a large margin if all the examples are not only on the correct side of the separating hyperplane but also far away from it. Restricting the algorithm to output a large margin separator can yield a small sample complexity even if the dimensionality of the feature space is high (and even infinite). We introduce the concept of margin and relate it to the regularized loss minimization paradigm as well as to the convergence rate of the Perceptron algorithm.

In the next chapter we will tackle the computational complexity challenge using the idea of *kernels*.

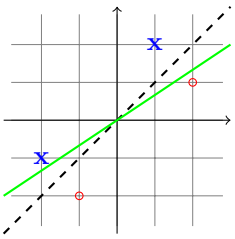
15.1 Margin and Hard-SVM

Let $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ be a training set of examples, where each $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. We say that this training set is linearly separable, if there exists a halfspace, $(\mathbf{w}, b)$, such that $y_i = \text{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$ for all i. Alternatively, this condition can be rewritten as

$$\forall i \in [m], \quad y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0.$$

All halfspaces $(\mathbf{w}, b)$ that satisfy this condition are ERM hypotheses (their 0-1 error is zero, which is the minimum possible error). For any separable training sample, there are many ERM halfspaces. Which one of them should the learner pick?

Consider, for example, the training set described in the picture that follows.



While both the dashed-black and solid-green hyperplanes separate the four examples, our intuition would probably lead us to prefer the black hyperplane over the green one. One way to formalize this intuition is using the concept of *margin*.

The margin of a hyperplane with respect to a training set is defined to be the minimal distance between a point in the training set and the hyperplane. If a hyperplane has a large margin, then it will still separate the training set even if we slightly perturb each instance.

We will see later on that the true error of a halfspace can be bounded in terms of the margin it has over the training sample (the larger the margin, the smaller the error), regardless of the Euclidean dimension in which this halfspace resides.

Hard-SVM is the learning rule in which we return an ERM hyperplane that separates the training set with the largest possible margin. To define Hard-SVM formally, we first express the distance between a point $\mathbf{x}$ to a hyperplane using the parameters defining the halfspace.

CLAIM 15.1 *The distance between a point* $\mathbf{x}$ *and the hyperplane defined by* $(\mathbf{w}, b)$ *where* $\|\mathbf{w}\| = 1$ *is* $|\langle \mathbf{w}, \mathbf{x} \rangle + b|$.

Proof The distance between a point $\mathbf{x}$ and the hyperplane is defined as

$$\min\{\|\mathbf{x} - \mathbf{v}\| : \langle \mathbf{w}, \mathbf{v} \rangle + b = 0\}.$$

Taking $\mathbf{v} = \mathbf{x} - (\langle \mathbf{w}, \mathbf{x} \rangle + b)\mathbf{w}$ we have that

$$\langle \mathbf{w}, \mathbf{v} \rangle + b = \langle \mathbf{w}, \mathbf{x} \rangle - (\langle \mathbf{w}, \mathbf{x} \rangle + b)\|\mathbf{w}\|^2 + b = 0,$$

and

$$\|\mathbf{x} - \mathbf{v}\| = |\langle \mathbf{w}, \mathbf{x} \rangle + b| \, \|\mathbf{w}\| = |\langle \mathbf{w}, \mathbf{x} \rangle + b|.$$

Hence, the distance is at most $|\langle \mathbf{w}, \mathbf{x} \rangle + b|$. Next, take any other point $\mathbf{u}$ on the hyperplane, thus $\langle \mathbf{w}, \mathbf{u} \rangle + b = 0$. We have

$$\begin{aligned}
\|\mathbf{x} - \mathbf{u}\|^2 &= \|\mathbf{x} - \mathbf{v} + \mathbf{v} - \mathbf{u}\|^2 \\
&= \|\mathbf{x} - \mathbf{v}\|^2 + \|\mathbf{v} - \mathbf{u}\|^2 + 2\langle \mathbf{x} - \mathbf{v}, \mathbf{v} - \mathbf{u} \rangle \\
&\geq \|\mathbf{x} - \mathbf{v}\|^2 + 2\langle \mathbf{x} - \mathbf{v}, \mathbf{v} - \mathbf{u} \rangle \\
&= \|\mathbf{x} - \mathbf{v}\|^2 + 2(\langle \mathbf{w}, \mathbf{x} \rangle + b)\langle \mathbf{w}, \mathbf{v} - \mathbf{u} \rangle \\
&= \|\mathbf{x} - \mathbf{v}\|^2,
\end{aligned}$$

where the last equality is because $\langle \mathbf{w}, \mathbf{v} \rangle = \langle \mathbf{w}, \mathbf{u} \rangle = -b$. Hence, the distance

between $\mathbf{x}$ and $\mathbf{u}$ is at least the distance between $\mathbf{x}$ and $\mathbf{v}$, which concludes our proof. □

On the basis of the preceding claim, the closest point in the training set to the separating hyperplane is $\min_{i\in[m]} |\langle \mathbf{w}, \mathbf{x}_i\rangle + b|$. Therefore, the Hard-SVM rule is

$$\operatorname*{argmax}_{(\mathbf{w},b):\|\mathbf{w}\|=1} \min_{i\in[m]} |\langle \mathbf{w}, \mathbf{x}_i\rangle + b| \quad \text{s.t.} \quad \forall i,\ y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) > 0.$$

Whenever there is a solution to the preceding problem (i.e., we are in the separable case), we can write an equivalent problem as follows (see Exercise 1):

$$\operatorname*{argmax}_{(\mathbf{w},b):\|\mathbf{w}\|=1} \min_{i\in[m]} y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b). \tag{15.1}$$

Next, we give another equivalent formulation of the Hard-SVM rule as a quadratic optimization problem.[1]

Hard-SVM

input: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$

solve:

$$(\mathbf{w}_0, b_0) = \operatorname*{argmin}_{(\mathbf{w},b)} \|\mathbf{w}\|^2 \text{ s.t. } \forall i,\ y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 \tag{15.2}$$

output: $\hat{\mathbf{w}} = \frac{\mathbf{w}_0}{\|\mathbf{w}_0\|}, \quad \hat{b} = \frac{b_0}{\|\mathbf{w}_0\|}$

The lemma that follows shows that the output of hard-SVM is indeed the separating hyperplane with the largest margin. Intuitively, hard-SVM searches for $\mathbf{w}$ of minimal norm among all the vectors that separate the data and for which $|\langle \mathbf{w}, \mathbf{x}_i\rangle + b| \geq 1$ for all i. In other words, we enforce the margin to be 1, but now the units in which we measure the margin scale with the norm of $\mathbf{w}$. Therefore, finding the largest margin halfspace boils down to finding $\mathbf{w}$ whose norm is minimal. Formally:

LEMMA 15.2 *The output of Hard-SVM is a solution of Equation (15.1).*

Proof Let $(\mathbf{w}^\star, b^\star)$ be a solution of Equation (15.1) and define the margin achieved by $(\mathbf{w}^\star, b^\star)$ to be $\gamma^\star = \min_{i\in[m]}\ y_i(\langle \mathbf{w}^\star, \mathbf{x}_i\rangle + b^\star)$. Therefore, for all i we have

$$y_i(\langle \mathbf{w}^\star, \mathbf{x}_i\rangle + b^\star) \geq \gamma^\star$$

or equivalently

$$y_i(\langle \tfrac{\mathbf{w}^\star}{\gamma^\star}, \mathbf{x}_i\rangle + \tfrac{b^\star}{\gamma^\star}) \geq 1.$$

Hence, the pair $(\frac{\mathbf{w}^\star}{\gamma^\star}, \frac{b^\star}{\gamma^\star})$ satisfies the conditions of the quadratic optimization

[1] A quadratic optimization problem is an optimization problem in which the objective is a convex quadratic function and the constraints are linear inequalities.

problem given in Equation (15.2). Therefore, $\|\mathbf{w}_0\| \leq \|\frac{\mathbf{w}^\star}{\gamma^\star}\| = \frac{1}{\gamma^\star}$. It follows that for all i,

$$y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) = \frac{1}{\|\mathbf{w}_0\|} y_i(\langle \mathbf{w}_0, \mathbf{x}_i \rangle + b_0) \geq \frac{1}{\|\mathbf{w}_0\|} \geq \gamma^\star.$$

Since $\|\hat{\mathbf{w}}\| = 1$ we obtain that $(\hat{\mathbf{w}}, \hat{b})$ is an optimal solution of Equation (15.1). □

15.1.1 The Homogenous Case

It is often more convenient to consider homogenous halfspaces, namely, halfspaces that pass through the origin and are thus defined by $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$, where the bias term b is set to be zero. Hard-SVM for homogenous halfspaces amounts to solving

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad \forall i, \; y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1. \tag{15.3}$$

As we discussed in Chapter 9, we can reduce the problem of learning nonhomogenous halfspaces to the problem of learning homogenous halfspaces by adding one more feature to each instance of $\mathbf{x}_i$, thus increasing the dimension to $d + 1$.

Note, however, that the optimization problem given in Equation (15.2) does not regularize the bias term b, while if we learn a homogenous halfspace in $\mathbb{R}^{d+1}$ using Equation (15.3) then we regularize the bias term (i.e., the $d+1$ component of the weight vector) as well. However, regularizing b usually does not make a significant difference to the sample complexity.

15.1.2 The Sample Complexity of Hard-SVM

Recall that the VC-dimension of halfspaces in $\mathbb{R}^d$ is $d + 1$. It follows that the sample complexity of learning halfspaces grows with the dimensionality of the problem. Furthermore, the fundamental theorem of learning tells us that if the number of examples is significantly smaller than d/ϵ then no algorithm can learn an ϵ-accurate halfspace. This is problematic when d is very large.

To overcome this problem, we will make an additional assumption on the underlying data distribution. In particular, we will define a "separability with margin γ" assumption and will show that if the data is separable with margin γ then the sample complexity is bounded from above by a function of $1/\gamma^2$. It follows that even if the dimensionality is very large (or even infinite), as long as the data adheres to the separability with margin assumption we can still have a small sample complexity. There is no contradiction to the lower bound given in the fundamental theorem of learning because we are now making an additional assumption on the underlying data distribution.

Before we formally define the separability with margin assumption, there is a scaling issue we need to resolve. Suppose that a training set $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ is separable with a margin γ, namely, the maximal objective value of Equation (15.1) is at least γ. Then, for any positive scalar $\alpha > 0$, the training set

$S' = (\alpha \mathbf{x}_1, y_1), \ldots, (\alpha \mathbf{x}_m, y_m)$ is separable with a margin of $\alpha\gamma$. That is, a simple scaling of the data can make it separable with an arbitrarily large margin. It follows that in order to give a meaningful definition of margin we must take into account the scale of the examples as well. One way to formalize this is using the definition that follows.

DEFINITION 15.3 Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. We say that $\mathcal{D}$ is separable with a (γ, ρ)-margin if there exists $(\mathbf{w}^\star, b^\star)$ such that $\|\mathbf{w}^\star\| = 1$ and such that with probability 1 over the choice of $(\mathbf{x}, y) \sim \mathcal{D}$ we have that $y(\langle \mathbf{w}^\star, \mathbf{x} \rangle + b^\star) \geq \gamma$ and $\|\mathbf{x}\| \leq \rho$. Similarly, we say that $\mathcal{D}$ is separable with a (γ, ρ)-margin using a homogenous halfspace if the preceding holds with a halfspace of the form $(\mathbf{w}^\star, 0)$.

In the advanced part of the book (Chapter 26), we will prove that the sample complexity of Hard-SVM depends on $(\rho/\gamma)^2$ and is independent of the dimension d. In particular, Theorem 26.13 in Section 26.3 states the following:

THEOREM 15.4 *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the (γ, ρ)-separability with margin assumption using a homogenous halfspace. Then, with probability of at least $1 - \delta$ over the choice of a training set of size m, the 0-1 error of the output of Hard-SVM is at most*

$$\sqrt{\frac{4\,(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}.$$

Remark 15.1 (Margin and the Perceptron) In Section 9.1.2 we have described and analyzed the Perceptron algorithm for finding an ERM hypothesis with respect to the class of halfspaces. In particular, in Theorem 9.1 we upper bounded the number of updates the Perceptron might make on a given training set. It can be shown (see Exercise 2) that the upper bound is exactly $(\rho/\gamma)^2$, where ρ is the radius of examples and γ is the margin.

15.2 Soft-SVM and Norm Regularization

The Hard-SVM formulation assumes that the training set is linearly separable, which is a rather strong assumption. Soft-SVM can be viewed as a relaxation of the Hard-SVM rule that can be applied even if the training set is not linearly separable.

The optimization problem in Equation (15.2) enforces the hard constraints $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ for all i. A natural relaxation is to allow the constraint to be violated for some of the examples in the training set. This can be modeled by introducing nonnegative slack variables, $\xi_1, \ldots, \xi_m$, and replacing each constraint $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ by the constraint $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i$. That is, ξ_i measures by how much the constraint $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ is being violated. Soft-SVM jointly minimizes the norm of $\mathbf{w}$ (corresponding to the margin) and the average of ξ_i (corresponding to the violations of the constraints). The tradeoff between the two

terms is controlled by a parameter λ. This leads to the Soft-SVM optimization problem:

Soft-SVM

input: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
parameter: $\lambda > 0$
solve:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \left(\lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i \right) \quad \text{s.t. } \forall i, \quad y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \tag{15.4}$$

output: $\mathbf{w}, b$

We can rewrite Equation (15.4) as a regularized loss minimization problem. Recall the definition of the hinge loss:

$$\ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}, y)) = \max\{0, 1 - y(\langle \mathbf{w}, \mathbf{x} \rangle + b)\}.$$

Given $(\mathbf{w}, b)$ and a training set S, the averaged hinge loss on S is denoted by $L_S^{\text{hinge}}((\mathbf{w}, b))$. Now, consider the regularized loss minimization problem:

$$\min_{\mathbf{w},b} \left(\lambda \|\mathbf{w}\|^2 + L_S^{\text{hinge}}((\mathbf{w}, b)) \right). \tag{15.5}$$

CLAIM 15.5 *Equation (15.4) and Equation (15.5) are equivalent.*

Proof Fix some $\mathbf{w}, b$ and consider the minimization over $\boldsymbol{\xi}$ in Equation (15.4). Fix some i. Since ξ_i must be nonnegative, the best assignment to ξ_i would be 0 if $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ and would be $1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$ otherwise. In other words, $\xi_i = \ell^{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}_i, y_i))$ for all i, and the claim follows. □

We therefore see that Soft-SVM falls into the paradigm of regularized loss minimization that we studied in the previous chapter. A Soft-SVM algorithm, that is, a solution for Equation (15.5), has a bias toward low norm separators. The objective function that we aim to minimize in Equation (15.5) penalizes not only for training errors but also for large norm.

It is often more convenient to consider Soft-SVM for learning a homogenous halfspace, where the bias term b is set to be zero, which yields the following optimization problem:

$$\min_{\mathbf{w}} \left(\lambda \|\mathbf{w}\|^2 + L_S^{\text{hinge}}(\mathbf{w}) \right), \tag{15.6}$$

where

$$L_S^{\text{hinge}}(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}_i \rangle\}.$$

15.2.1 The Sample Complexity of Soft-SVM

We now analyze the sample complexity of Soft-SVM for the case of homogenous halfspaces (namely, the output of Equation (15.6)). In Corollary 13.8 we derived a generalization bound for the regularized loss minimization framework assuming that the loss function is convex and Lipschitz. We have already shown that the hinge loss is convex so it is only left to analyze the Lipschitzness of the hinge loss.

CLAIM 15.6 *Let $f(\mathbf{w}) = \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\}$. Then, f is $\|\mathbf{x}\|$-Lipschitz.*

Proof It is easy to verify that any subgradient of f at $\mathbf{w}$ is of the form $\alpha\mathbf{x}$ where $|\alpha| \leq 1$. The claim now follows from Lemma 14.7. □

Corollary 13.8 therefore yields the following:

COROLLARY 15.7 *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0, 1\}$, where $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\| \leq \rho\}$. Consider running Soft-SVM (Equation (15.6)) on a training set $S \sim \mathcal{D}^m$ and let $A(S)$ be the solution of Soft-SVM. Then, for every $\mathbf{u}$,*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L^{\text{hinge}}_{\mathcal{D}}(A(S))] \;\leq\; L^{\text{hinge}}_{\mathcal{D}}(\mathbf{u}) + \lambda\|\mathbf{u}\|^2 + \frac{2\rho^2}{\lambda m}.$$

Furthermore, since the hinge loss upper bounds the $0-1$ loss we also have

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L^{0-1}_{\mathcal{D}}(A(S))] \;\leq\; L^{\text{hinge}}_{\mathcal{D}}(\mathbf{u}) + \lambda\|\mathbf{u}\|^2 + \frac{2\rho^2}{\lambda m}.$$

Last, for every $B > 0$, if we set $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ then

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L^{0-1}_{\mathcal{D}}(A(S))] \;\leq\; \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L^{\text{hinge}}_{\mathcal{D}}(A(S))] \;\leq\; \min_{\mathbf{w}:\|\mathbf{w}\|\leq B} L^{\text{hinge}}_{\mathcal{D}}(\mathbf{w}) + \sqrt{\frac{8\rho^2 B^2}{m}}.$$

We therefore see that we can control the sample complexity of learning a halfspace as a function of the norm of that halfspace, independently of the Euclidean dimension of the space over which the halfspace is defined. This becomes highly significant when we learn via embeddings into high dimensional feature spaces, as we will consider in the next chapter.

Remark 15.2 The condition that $\mathcal{X}$ will contain vectors with a bounded norm follows from the requirement that the loss function will be Lipschitz. This is not just a technicality. As we discussed before, separation with large margin is meaningless without imposing a restriction on the scale of the instances. Indeed, without a constraint on the scale, we can always enlarge the margin by multiplying all instances by a large scalar.

15.2.2 Margin and Norm-Based Bounds versus Dimension

The bounds we have derived for Hard-SVM and Soft-SVM do not depend on the dimension of the instance space. Instead, the bounds depend on the norm of the

examples, ρ, the norm of the halfspace B (or equivalently the margin parameter γ) and, in the nonseparable case, the bounds also depend on the minimum hinge loss of all halfspaces of norm $\leq B$. In contrast, the VC-dimension of the class of homogenous halfspaces is d, which implies that the error of an ERM hypothesis decreases as $\sqrt{d/m}$ does. We now give an example in which $\rho^2 B^2 \ll d$; hence the bound given in Corollary 15.7 is much better than the VC bound.

Consider the problem of learning to classify a short text document according to its topic, say, whether the document is about sports or not. We first need to represent documents as vectors. One simple yet effective way is to use a *bag-of-words* representation. That is, we define a dictionary of words and set the dimension d to be the number of words in the dictionary. Given a document, we represent it as a vector $\mathbf{x} \in \{0,1\}^d$, where $x_i = 1$ if the i'th word in the dictionary appears in the document and $x_i = 0$ otherwise. Therefore, for this problem, the value of ρ^2 will be the maximal number of distinct words in a given document.

A halfspace for this problem assigns weights to words. It is natural to assume that by assigning positive and negative weights to a few dozen words we will be able to determine whether a given document is about sports or not with reasonable accuracy. Therefore, for this problem, the value of B^2 can be set to be less than 100. Overall, it is reasonable to say that the value of $B^2\rho^2$ is smaller than 10,000.

On the other hand, a typical size of a dictionary is much larger than 10,000. For example, there are more than 100,000 distinct words in English. We have therefore shown a problem in which there can be an order of magnitude difference between learning a halfspace with the SVM rule and learning a halfspace using the vanilla ERM rule.

Of course, it is possible to construct problems in which the SVM bound will be worse than the VC bound. When we use SVM, we in fact introduce another form of inductive bias – we prefer large margin halfspaces. While this inductive bias can significantly decrease our estimation error, it can also enlarge the approximation error.

15.2.3 The Ramp Loss*

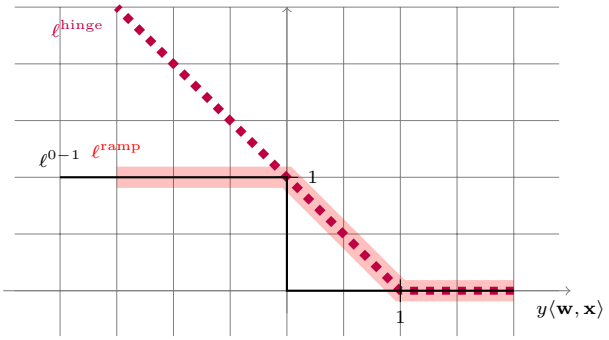
The margin-based bounds we have derived in Corollary 15.7 rely on the fact that we minimize the hinge loss. As we have shown in the previous subsection, the term $\sqrt{\rho^2 B^2/m}$ can be much smaller than the corresponding term in the VC bound, $\sqrt{d/m}$. However, the approximation error in Corollary 15.7 is measured with respect to the hinge loss while the approximation error in VC bounds is measured with respect to the $0-1$ loss. Since the hinge loss upper bounds the $0-1$ loss, the approximation error with respect to the $0-1$ loss will never exceed that of the hinge loss.

It is not possible to derive bounds that involve the estimation error term $\sqrt{\rho^2 B^2/m}$ for the $0-1$ loss. This follows from the fact that the $0-1$ loss is scale

insensitive, and therefore there is no meaning to the norm of $\mathbf{w}$ or its margin when we measure error with the 0−1 loss. However, it is possible to define a loss function that on one hand it is scale sensitive and thus enjoys the estimation error $\sqrt{\rho^2 B^2/m}$ while on the other hand it is more similar to the 0−1 loss. One option is the *ramp loss*, defined as

$$\ell^{\text{ramp}}(\mathbf{w}, (\mathbf{x}, y)) = \min\{1, \ell^{\text{hinge}}(\mathbf{w}, (\mathbf{x}, y))\} = \min\{1\,,\, \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\}\}.$$

The ramp loss penalizes mistakes in the same way as the 0−1 loss and does not penalize examples that are separated with margin. The difference between the ramp loss and the 0−1 loss is only with respect to examples that are correctly classified but not with a significant margin. Generalization bounds for the ramp loss are given in the advanced part of this book (see Appendix 26.3).



The reason SVM relies on the hinge loss and not on the ramp loss is that the hinge loss is convex and, therefore, from the *computational* point of view, minimizing the hinge loss can be performed efficiently. In contrast, the problem of minimizing the ramp loss is computationally intractable.

15.3 Optimality Conditions and "Support Vectors"*

The name "Support Vector Machine" stems from the fact that the solution of hard-SVM, $\mathbf{w}_0$, is supported by (i.e., is in the linear span of) the examples that are exactly at distance $1/\|\mathbf{w}_0\|$ from the separating hyperplane. These vectors are therefore called *support vectors*. To see this, we rely on **Fritz John optimality conditions**.

THEOREM 15.8 *Let $\mathbf{w}_0$ be as defined in Equation (15.3) and let $I = \{i : |\langle \mathbf{w}_0, \mathbf{x}_i\rangle| = 1\}$. Then, there exist coefficients $\alpha_1, \ldots, \alpha_m$ such that*

$$\mathbf{w}_0 = \sum_{i \in I} \alpha_i \mathbf{x}_i.$$

The examples $\{\mathbf{x}_i : i \in I\}$ are called *support vectors*.

The proof of this theorem follows by applying the following lemma to Equation (15.3).

LEMMA 15.9 (Fritz John) *Suppose that*

$$\mathbf{w}^\star \in \operatorname*{argmin}_{\mathbf{w}} f(\mathbf{w}) \quad s.t. \;\; \forall i \in [m],\; g_i(\mathbf{w}) \leq 0,$$

where $f, g_1, \ldots, g_m$ *are differentiable. Then, there exists* $\boldsymbol{\alpha} \in \mathbb{R}^m$ *such that* $\nabla f(\mathbf{w}^\star) + \sum_{i \in I} \alpha_i \nabla g_i(\mathbf{w}^\star) = \mathbf{0}$*, where* $I = \{i : g_i(\mathbf{w}^\star) = 0\}$*.*

15.4 Duality*

Historically, many of the properties of SVM have been obtained by considering the *dual* of Equation (15.3). Our presentation of SVM does not rely on duality. For completeness, we present in the following how to derive the dual of Equation (15.3).

We start by rewriting the problem in an equivalent form as follows. Consider the function

$$g(\mathbf{w}) \;=\; \max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \;=\; \begin{cases} 0 & \text{if } \forall i, y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1 \\ \infty & \text{otherwise} \end{cases} .$$

We can therefore rewrite Equation (15.3) as

$$\min_{\mathbf{w}} \left(\|\mathbf{w}\|^2 + g(\mathbf{w}) \right). \tag{15.7}$$

Rearranging the preceding we obtain that Equation (15.3) can be rewritten as the problem

$$\min_{\mathbf{w}} \max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right). \tag{15.8}$$

Now suppose that we flip the order of min and max in the above equation. This can only decrease the objective value (see Exercise 4), and we have

$$\begin{aligned} &\min_{\mathbf{w}} \max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right) \\ &\geq \max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{w}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right). \end{aligned}$$

The preceding inequality is called *weak duality*. It turns out that in our case, *strong duality* also holds; namely, the inequality holds with equality. Therefore, the *dual* problem is

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{w}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right). \tag{15.9}$$

We can simplify the dual problem by noting that once $\boldsymbol{\alpha}$ is fixed, the optimization

problem with respect to $\mathbf{w}$ is unconstrained and the objective is differentiable; thus, at the optimum, the gradient equals zero:

$$\mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \;=\; 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i .$$

This shows us that the solution must be in the linear span of the examples, a fact we will use later to derive SVM with kernels. Plugging the preceding into Equation (15.9) we obtain that the dual problem can be rewritten as

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \left(\frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left\langle \sum_j \alpha_j y_j \mathbf{x}_j , \mathbf{x}_i \right\rangle \right) \right) . \tag{15.10}$$

Rearranging yields the dual problem

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m : \boldsymbol{\alpha} \geq \mathbf{0}} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_j , \mathbf{x}_i \rangle \right) . \tag{15.11}$$

Note that the dual problem only involves inner products between instances and does not require direct access to specific elements within an instance. This property is important when implementing SVM with kernels, as we will discuss in the next chapter.

15.5 Implementing Soft-SVM Using SGD

In this section we describe a very simple algorithm for solving the optimization problem of Soft-SVM, namely,

$$\min_{\mathbf{w}} \left(\frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y \langle \mathbf{w}, \mathbf{x}_i \rangle\} \right) . \tag{15.12}$$

We rely on the SGD framework for solving regularized loss minimization problems, as described in Section 14.5.3.

Recall that, on the basis of Equation (14.15), we can rewrite the update rule of SGD as

$$\mathbf{w}^{(t+1)} = -\frac{1}{\lambda t} \sum_{j=1}^{t} \mathbf{v}_j ,$$

where $\mathbf{v}_j$ is a subgradient of the loss function at $\mathbf{w}^{(j)}$ on the random example chosen at iteration j. For the hinge loss, given an example $(\mathbf{x}, y)$, we can choose $\mathbf{v}_j$ to be $\mathbf{0}$ if $y \langle \mathbf{w}^{(j)}, \mathbf{x} \rangle \geq 1$ and $\mathbf{v}_j = -y\,\mathbf{x}$ otherwise (see Example 14.2). Denoting $\boldsymbol{\theta}^{(t)} = -\sum_{j<t} \mathbf{v}_j$ we obtain the following procedure.

SGD for Solving Soft-SVM

goal: Solve Equation (15.12)
parameter: T
initialize: $\boldsymbol{\theta}^{(1)} = \mathbf{0}$
for $t = 1, \ldots, T$
  Let $\mathbf{w}^{(t)} = \frac{1}{\lambda t}\boldsymbol{\theta}^{(t)}$
  Choose i uniformly at random from $[m]$
  If $(y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle < 1)$
    Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + y_i \mathbf{x}_i$
  Else
    Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$
output: $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}$

15.6 Summary

SVM is an algorithm for learning halfspaces with a certain type of prior knowledge, namely, preference for large margin. Hard-SVM seeks the halfspace that separates the data perfectly with the largest margin, whereas soft-SVM does not assume separability of the data and allows the constraints to be violated to some extent. The sample complexity for both types of SVM is different from the sample complexity of straightforward halfspace learning, as it does not depend on the dimension of the domain but rather on parameters such as the maximal norms of $\mathbf{x}$ and $\mathbf{w}$.

The importance of dimension-independent sample complexity will be realized in the next chapter, where we will discuss the embedding of the given domain into some high dimensional feature space as means for enriching our hypothesis class. Such a procedure raises computational and sample complexity problems. The latter is solved by using SVM, whereas the former can be solved by using SVM with kernels, as we will see in the next chapter.

15.7 Bibliographic Remarks

SVMs have been introduced in (Cortes & Vapnik 1995, Boser, Guyon & Vapnik 1992). There are many good books on the theoretical and practical aspects of SVMs. For example, (Vapnik 1995, Cristianini & Shawe-Taylor 2000, Schölkopf & Smola 2002, Hsu, Chang & Lin 2003, Steinwart & Christmann 2008). Using SGD for solving soft-SVM has been proposed in Shalev-Shwartz et al. (2007).