ELEG/CISC 867: Advanced Machine Learning **Spring 2019**

Lecture 12: Support Vector Machine

Lecturer: Xiugang Wu *05/07/2019, 05/09/2019*

Support Vector Machine (SVM) is a type of "large margin" classifier, which seeks for a halfspace that separates a training set with a large margin, i.e. 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).

1 Margin

Recall that the hypothesis class of halfspaces $\mathcal{H}_{\text{HS}}$ is given by

$$\mathcal{H}_{\text{HS}} = \{\mathbf{x} \mapsto \text{sgn}(\mathbf{w}^T\mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

In the separable case, i.e. $\mathbb{P}(Y = \text{sgn}(\mathbf{w}^T\mathbf{X} + b)) = 1$ for some $(\mathbf{w}, b)$, the ERM rule for learning halfspaces reduces to finding $(\mathbf{w}, b)$ such that

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0, \ \forall i \in [1:n].$$

Note that for a separable training sample there are many ERM halfspaces, so which one should we pick? Hard-SVM addresses this issue by picking an ERM hyperplane that separates the training set with the largest possible margin; here 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.

Formally, for a point $\mathbf{x}$ and hyperplane $L = \{\mathbf{v} : \mathbf{w}^T\mathbf{v} + b = 0\}$, the distance $d(\mathbf{x}, L)$ between $\mathbf{x}$ and L is defined as

$$d(\mathbf{x}, L) = \min\{\|\mathbf{x} - \mathbf{v}\| : \mathbf{v} \in L\}.$$

Assuming $\mathbf{v}$ is some point on the plane L, we have

$$\begin{aligned} d(\mathbf{x}, L) &= \left| \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x} - \mathbf{v}) \right| \\ &= \frac{1}{\|\mathbf{w}\|} \left| \mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{v} \right| \\ &= \frac{1}{\|\mathbf{w}\|} \left| \mathbf{w}^T\mathbf{x} + b - (\mathbf{w}^T\mathbf{v} + b) \right| \\ &= \frac{1}{\|\mathbf{w}\|} \left| \mathbf{w}^T\mathbf{x} + b \right|, \end{aligned}$$

which further simplifies to $\frac{1}{\|\mathbf{w}\|}$ if $\left|\mathbf{w}^T\mathbf{x} + b\right| = 1$. Therefore, for a linearly separable training set, a separating hyperplane $(\mathbf{w}, b)$ satisfies

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0, \forall i \in [1:n]$$

and has margin γ given by

$$\gamma = \min_{i \in [1:n]} \frac{1}{\|\mathbf{w}\|} \left| \mathbf{w}^T\mathbf{x}_i + b \right| = \min_{i \in [1:n]} \frac{1}{\|\mathbf{w}\|} y_i(\mathbf{w}^T\mathbf{x}_i + b).$$

Note that if we scale $(\mathbf{w}, b)$ such that

$$\min_{i\in[1:n]} y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1,$$

then the margin is simply given by

$$\gamma = \frac{1}{\|\mathbf{w}\|}.$$

The closest examples, i.e. the $\mathbf{x}_i$'s that attain the above minimum, are called support vectors.

2 Hard-SVM

Hard-SVM seeks for the separating plane with the largest margin, i.e.,

$$\operatorname*{argmax}_{(\mathbf{w},b)} \frac{1}{\|\mathbf{w}\|} \quad \text{s.t.} \quad \min_{i\in[1:n]} y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1.$$

Writing the above optimization problem as an Quadratic Program (QP), we obtain

$$\operatorname*{argmin}_{(\mathbf{w},b)} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad \forall i \in [1:n].$$

This leads to the algorithm of hard-SVM.

Algorithm 1 Hard-SVM

1: **input:** Training sample $z^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$

2: **solve:**

$$(\mathbf{w}_0, b_0) = \operatorname*{argmin}_{(\mathbf{w},b)} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad \forall i \in [1:n]$$

3: **output:** Separating plane $(\mathbf{w}_0, b_0)$, margin $\gamma = \frac{1}{\|\mathbf{w}_0\|}$

2.1 Sample Complexity of Hard-SVM

Recall that the VC dimension of halfspaces in $\mathbb{R}^d$ is $d + 1$. Therefore, the sample complexity of learning halfspaces grows with the dimensionality d, which is problematic in high-dimensional setups. To address this issue, 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 $1/\gamma^2$. Hence, 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.

We now formalize this idea. First note that to formally define the separability with margin assumption, there is a scaling issue we need to resolve. This is because simply scaling the data from $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ to $\{(\alpha\mathbf{x}_i, y_i)\}_{i=1}^n$ will result in an α times larger margin, but obviously the sample complexity wouldn't become smaller just because of this scaling trick. Therefore, when defining the separability with margin assumption, we need to take into account the scale of $\mathbf{x}$. One way to formalize this is to use the following definition.

Definition 2.1 *We say that P over $\mathbb{R}^d \times \{\pm 1\}$ is separable with a (γ, ρ)-margin if there exists some $(\mathbf{w}, b)$ such that*

$$P(\|\mathbf{x}\| \leq \rho \text{ and } Y(\mathbf{w}^T\mathbf{X} + b)/\|\mathbf{w}\| \geq \gamma) = 1.$$

Another way to look at the above definition is to realize that what really matters in determining the sample complexity is the relative margin with respect to the scale of $\mathbf{x}$, i.e. γ/ρ, rather than the absolute margin value γ. Indeed, one can show that if the underlying data distribution is separable with a (γ, ρ)-margin, then the sample complexity grows with $(\rho/\gamma)^2$.

Theorem 2.1 *If the underlying data distribution P is separable with a (γ, ρ)-margin, then the sample complexity of hard-SVM satisfies*

$$n(\epsilon, \delta) \leq \frac{8}{\epsilon^2}\left(2(\rho/\gamma)^2 + \log(2/\delta)\right).$$

Note that the above result implies that even if the dimensionality d is very high, as long as ρ/γ is small, i.e. the (relative) margin is large, we still enjoy a low sample complexity for learning SVM. This is not contradictory to the VC theory which says the sample complexity for learning halfspaces grow with d, because here we have made an additional assumption on the underlying distribution P and it is this additional assumption that helps to reduce the sample complexity.

3 Soft-SVM

To apply hard-SVM, the training set needs to be linearly separable. What if the training set is not linearly separable? One way to deal with this issue is to embed the domain set $\mathcal{X}$ into a (higher dimensional) feature space $\mathcal{F}$ so that the examples in the $\mathcal{F}$ space are separable; we will discuss this approach in detail in the next lecture. Another way to address this non-separability issue is to use the following soft-SVM algorithm.

Algorithm 2 Soft-SVM

1: **input:** Training sample $z^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $\lambda > 0$

2: **solve:**

$$\min_{(\mathbf{w},b,\xi)} \lambda\|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \quad \forall i \in [1:n]$$

3: **output:** solution $(\mathbf{w}, b)$

Compared to hard-SVM which enforces the constraint $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1,\ \forall i \in [1:n]$, soft-SVM allows this constraint to be violated and uses ξ_i to measure how much it is violated; it then jointly minimizes the the norm of $\mathbf{w}$ (corresponding to the margin) and the average of ξ_i (corresponding to the violations of the constraints).

Using the hinge loss $\ell_{\text{hinge}}((\mathbf{w}, b), (\mathbf{x}, y)) = \max\{0, 1 - y(\mathbf{w}^T\mathbf{x} + b)\}$, one can rewrite the soft-SVM as RLM under the hinge loss with Tikhonov regularization:

$$\min_{(\mathbf{w},b)} L_{\text{hinge}}((\mathbf{w}, b), z^n) + \lambda\|\mathbf{w}\|^2.$$

To see the equivalence between the above RLM formulation and soft-SVM, note the following:

- Under the hinge loss, the empirical risk is given by

$$L_{\text{hinge}}((\mathbf{w}, b), z^n) = \frac{1}{n}\sum_{i=1}^n \max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\}.$$

- For any fixed $(\mathbf{w}, b)$ pair, to minimize $\frac{1}{n}\sum_{i=1}^n \xi_i$ in soft-SVM, we would choose $\xi_i = 0$ if $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$ and $\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)$ if $y_i(\mathbf{w}^T\mathbf{x}_i + b) < 1$, i.e. $\xi_i = \max\{0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\}$.