210709 Machine Learning **Spring 2021**

Lecture 7. Support Vector Machine I

Lecturer: Jie Wang Date: April 30, 2021

The major references of this lecture are [2, 1].

1 Introduction

Suppose that we are given a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathcal{C} = \{-1, 1\}$, $i = 1, 2, \ldots, n$. Support vector machine (SVM) tries to find a linear function $f : \mathbb{R}^d \to \mathbb{R}$ in the form of

$$f(X; \mathbf{w}, b) = b + \sum_{j=1}^{d} w_j X_j,$$

such that

$$y_i = \mathbf{sign}\,(f(\mathbf{x}_i; \mathbf{w}, b)).$$

To fit the data, we need to put all the positive training instances in the positive half space and the negative training instances in the negative half space.

2 SVM for Linearly Separable Cases

2.1 Maximum Margin

To illustrate the idea of SVM, we consider a simple case where the training samples are ***linearly separable***, that is, we can find a ***hyperplane***—which separates the feature space into two ***halfspaces***: the positive halfspace and the negative halfspace—such that positive and negative data instances fall into the positive and negative halfspaces, respectively.

Definition 1. Let $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{w} \neq 0$, and $b \in \mathbb{R}$. A linear classifier that takes the form of

$$f(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \tag{1}$$

defines a hyperplane (its 0-level set)

$$H_f = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}; \mathbf{w}, b) = 0\},$$

separating the feature space into two halfspaces: the positive halfspace

$$H_f^+ = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}; \mathbf{w}, b) > 0\},$$

and the negative halfspace

$$H_f^- = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}; \mathbf{w}, b) < 0\},.$$

Thus, linearly separable indeed assumes the existence of a hyperplane H_f (specified by a linear classifier f) such that all positive (negative) labeled data instances belong to the positive (negative) half space H_f^+ (H_f^-). In other words, the labels of the data instances share the same sign with the halfspaces they fall into. This leads to a concise definition of linearly separable.

Definition 2. A training sample is linearly separable if there exists $(\hat{\mathbf{w}}, \hat{b})$ such that

$$y_i = \mathbf{sign}\,(f(\mathbf{x}_i; \hat{\mathbf{w}}, \hat{b})),\, \forall\, i \in [n], \tag{2}$$

which is equivalent to

$$y_i f(\mathbf{x}_i; \hat{\mathbf{w}}, \hat{b}) > 0,\, \forall\, i \in [n], \tag{3}$$

where $[n] = \{1, \ldots, n\}$.

In this section, we assume that the training sample is linearly separable.

Assumption 1. *The training sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$ is linearly separable.*

However, we can find infinitely many hyperplanes such that the inequality in (3) holds. Which one shall we choose? The SVM classifier makes the decision based on the notion of ***geometric margin***.

Definition 3. Suppose that we have a data sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$. The ***geometric margin*** $\gamma_f(\mathbf{x}_i)$ of a linear classifier

$$f(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

at a point $\mathbf{x}_i$ is its ***signed Euclidean distance*** to the hyperplane $\{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$:

$$\gamma_f(\mathbf{x}_i) = \frac{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}{\|\mathbf{w}\|}.$$

The ***geometric margin*** γ_f of a linear classifier f for a sample $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_i^n$ is the minimum geometric margin over the points in the sample, that is

$$\gamma_f = \min_{i \in [n]} \gamma_f(\mathbf{x}_i).$$

Remark 1. The geometric margin of a data instance to a hyperplane can be *negative*, which implies that it falls into the wrong side of the hyperplane. Given a training sample, a ***negative*** geometric margin implies that some of the data instances are ***misclassified***.

SVM looks for the hyperplane which maximizes the geometric margin, and thus it is known as the ***maximum margin classifier***. Specifically, we can model SVM by the following optimization problem:

$$\max_{\mathbf{w},b} \gamma_f = \max_{\mathbf{w},b} \min_{i \in [n]} \frac{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}{\|\mathbf{w}\|} = \max_{\mathbf{w},b} \frac{1}{\|\mathbf{w}\|} \left(\min_{i \in [n]} y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \right). \tag{4}$$

Remark 2. The problem in (4) is challenging to solve. One obvious reason is that the variables are mixtures of continuous variables ($\mathbf{w}$,b) and discrete variables (the index i), leaving many optimization methods we are familiar with out of our options. However, ***a surprising fact*** is that, the problem in (4) is equivalent to a convex optimization problem, which can be readily solved by many popular methods.

2.2 The Convex Version

We show that we can transform the problem in (4) to a convex optimization problem.

Step 1: Reducing the search space

Recall that the problem in (4) has two sets of variables: the continuous variables $(\mathbf{w}, b)$ and the discrete variable $i \in [n]$. We can see that the domain of the problem in (4) is

$$D = D_1 \times D_2,$$

where

$$D_1 = \{(\mathbf{w}, b) : \mathbf{w} \in \mathbb{R}^d, \mathbf{w} \neq 0, b \in \mathbb{R}\} \text{ and } D_2 = \{i : i = 1, 2, \dots, n\}.$$

Notice that, the value of the objective function in problem (4), i.e., the geometric margin γ_f, is unchanged if we multiply $(\mathbf{w}, b)$ by a *positive* scalar (why positive?), that is

$$\gamma_f = \gamma_{\lambda f}, \, \forall \, \gamma > 0.$$

In other words, for any $(\mathbf{w}, b) \in \mathbb{R}^{d+1}$ with $\mathbf{w} \neq 0$, all points of the ray $\{\lambda(\mathbf{w}, b) : \lambda > 0\}$ share the same value of the geometric margin. Thus, for any ray in $\mathbb{R}^{d+1}$ (except the two rays going upside and downside), we can consider only one single point of it. But which one shall we pick? Here comes the first trick in deriving SVM: we pick $(\mathbf{w}, b)$ that satisfies the constraint as follows.

$$\min_i \, y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1. \tag{5}$$

This transforms the problem in (4) to

$$\begin{aligned} &\max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|}, \\ &\text{s.t. } \min_i \, y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1. \end{aligned} \tag{6}$$

Remark 3. Notice that, what we are looking for is indeed a separating hyperplane defined by a linear classifier. However, different linear classifiers may specify the same separting hyperplanes. For example, it is easy to see that $H_f = H_{\lambda f}$ for any $\lambda > 0$. Thus, for a set of linear classifiers that define the same separating hyperplanes, we can only consider one of them. This is the geometric intuition behind the transformation from (4) to (6).

Step 2: Transforming the objective function to a convex function

In view of the problem in (6), we can see that maximizing $1/\|\mathbf{w}\|$ is equivalent to minimizing $\|\mathbf{w}\|$. Thus, we can transform (6) as follows.

$$\begin{aligned} &\min_{\mathbf{w}, b} \frac{1}{2}\|\mathbf{w}\|^2, \\ &\text{s.t. } \min_i \, y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1. \end{aligned} \tag{7}$$

Remark 4. We note that, though the problems in (6) and (7) are similar to each other, the former is NOT equivalent to the latter. The key difference is that, the problem in (6) does not allow $\mathbf{w} = 0$, while the problem (7) does.

Question 1. Under which cases, the problem in (7) admits optimal solutions in the form of $(0, b)$?

Step 3: Relaxing the constraints

The constraint in problem (7) is in the form of a minimization problem, which is difficult to deal with. However, we can relax the constraint (5) by requiring that

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \forall\, i \in [n],$$

Then, the problem in (7) changes to

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2, \quad \text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1. \tag{8}$$

The problem in (8) is the commonly-seem formulation of SVM for the linearly separable data samples. Though we arrive at (8) by relaxing the constraint in (7), we can show that the problems (7) and (8) are equivalent to each other, that is, one of the constraints in (8) must hold as an equality at its optimal solution.

Question 2.

1. Show there is at least one of the constraints holds as an equality at the optimum.
2. Show there exist at least one positive **and** negative samples such that the equality holds at the optimum.
3. Can we remove the inequalities that hold strictly at the optimum without affecting the solution?

Definition 4. Given a linear classifier in the form of (1), the ***marginal hyperplanes*** are

$$H_f(1) = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 1\} \text{ and } H_f(-1) = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = -1\}.$$

The ***support vectors*** are the data instances on the marginal hyperplanes, i.e.,

$$\{\mathbf{x} : |\langle \mathbf{w}, \mathbf{x} \rangle + b| = 1, \, \mathbf{x} \in \mathcal{D}\}.$$

3 SVM for Non-separable Cases

In most real applications, the training data instances are not linearly separable, that is, for any hyperplane H_f, there exists $\mathbf{x}_i \in \mathcal{D}$ such that

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) < 0.$$

Thus, the constraints in (8) can not hold simultaneously. To address this problem, we introduce a set of nonnegative ***slack variables*** $\{\xi_i\}_{i=1}^n$ to relax the constraints as

$$y_i(\langle \mathbf{w}, \mathbf{x} \rangle + b) \geq 1 - \xi_i, \, i \in [n].$$

We can see that the value of ξ_i measures the vector $\mathbf{x}_i$'s violation of the corresponding inequality $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$. To limit the violations over all data instances, we add a penalty to the objective function in (8), which leads to

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i, \quad \text{s.t. } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \, \xi_i \geq 0, \, i \in [n]. \tag{9}$$

The problem in (9) is a widely-used version of SVM for non-separable cases.

Question 3. For a linearly separable data sample, shall we arrive at the same separating hyperplane by solving the problems in (8) and (9), respectively?

References

[1] D. P. Bertsekas. *Nonlinear Programming, 3ed.* Athena Scientific, 2016.

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning, 2ed.* The MIT Press, 2018.