Lecture-03: Support Vector Machines – Separable Case

1 Linear binary classification

Let the input space be $\mathcal{X} = \mathbb{R}^N$ for the number of dimensions $N \geqslant 1$, the output space $\mathcal{Y} = \{-1, 1\}$, and the target function be some mapping $c : \mathcal{X} \to \mathcal{Y}$. We will also use $\|x\|$ to denote the 2-norm of $x \in \mathcal{X}$.

Definition 1.1. For a vector $w \in \mathbb{R}^N$ and scalar $b \in \mathbb{R}$, we define a hyperplane as a set of points

$$E_{w,b} \triangleq \left\{ x \in \mathbb{R}^N : \frac{\langle w, x \rangle}{\|w\|} = -\frac{b}{\|w\|} \right\}.$$

Definition 1.2. Distance of a point from a set is defined as $d(x, A) \triangleq \min\{d(x, y) : y \in A\}$. For $x, y \in \mathbb{R}^N$, the distance $d(x, y) \triangleq \|x - y\|^2$.

Lemma 1.3. *For a vector $w \in \mathbb{R}^N$ and $b \in \mathbb{R}$, we have $d(0, E_{w,b}) = -\frac{b}{\|w\|}$.*

Proof. For a vector w, we define a unit vector $u \triangleq \frac{w}{\|w\|}$. It follows that $x_0 \triangleq -\frac{b}{\|w\|} u$ lies on the hyperplane $E_{w,b}$, which is parallel to the unit vector u and at distance $d(0, x_0) = -\frac{b}{\|w\|}$ from the origin. Any point $x \in E_{w,b}$ on the hyperplane can be written as a sum of two orthogonal vectors $x = x_0 + x - x_0$ where $\langle x - x_0, w \rangle = 0$. Therefore, $d(0, x)^2 = d(0, x_0)^2 + d(x_0, x)^2 \geqslant d(0, x_0^2)$, and hence $d(0, E_{w,b}) = d(0, x_0)$. □

Remark 1. A hyperplane $E_{w,b} = \left\{ x \in \mathbb{R}^N : \langle w, x \rangle + b = 0 \right\}$ is defined in terms of the unit vector $w / \|w\|$ and its distance $-b / \|w\|$ from the origin.

Lemma 1.4. *The distance of any point $x \in \mathbb{R}^N$ to a hyperplane $E_{w,b}$ is given by $d(x, E_{w,b}) = \frac{|\langle w, x \rangle + b|}{\|w\|}$.*

Proof. Let $u = w / \|w\|$ be the unit vector in the direction of w. Any point y on a hyperplane $E_{w,b}$, can be written as sum of two orthogonal vectors $y = x_0 + y - x_0$ where $\langle y - x_0, x_0 \rangle = 0$ and $x_0 = -\frac{b}{\|w\|} u$. Any point $x \in \mathbb{R}^N$ can be represented as $x = \langle x, u \rangle u + v$, such that $\langle v, w \rangle = 0$. Therefore,

$$d(x, E_{w,b})^2 = \min_{y \in E_{w,b}} d(x, y)^2 = \min_{y \in E_{w,b}} d(x_0 + y - x_0, \langle x, u \rangle u + v)^2 \geqslant \left(\frac{\langle x, w \rangle + b}{\|w\|} \right)^2.$$

□

Remark 2. The distance of a point $x \in \mathbb{R}^N$ from the hyperplane $E_{w,b}$ is given by $d(x, E_{w,b})$. If $\langle w, x \rangle + b > 0$, then the point x lies above the hyperplane $E_{w,b}$, and if $\langle w, x \rangle + b < 0$, then point x lies below the hyperplane $E_{w,b}$.

Assumption 1.5. We are given a training sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ consisting of m labeled training examples $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, where each example x_i is generated *i.i.d.* by a fixed but unknown distribution D, and the label $y_i = c(x_i)$ for an unknown concept $c : \mathcal{X} \to \mathcal{Y}$.

Assumption 1.6. We define the hypothesis set as a collection of separating hyperplanes

$$H \triangleq \left\{ x \mapsto \text{sign}(\langle w, x \rangle + b) : w \in \mathbb{R}^N, b \in \mathbb{R} \right\}.$$

Remark 3. Any hypothesis $h \in H$ is identified by the pair (w, b) such that $h(x) = \text{sign}(\langle w, x \rangle + b)$ for all $x \in \mathbb{R}^N$. A hypothesis $h \in H$ labels positively all points falling on one side of the hyperplane $E_{w,b} \triangleq \left\{ x \in \mathbb{R}^N : \langle w, x \rangle + b = 0 \right\}$ and labels negatively all others. This problem is referred to as **linear binary classification problem**.

Remark 4. For the Hamming loss function $(y, y') \mapsto \mathbb{1}_{\{y \neq y'\}}$, the generalization error is $R(h) = P\{h(x) \neq c(x)\}$. The objective is to select an $h \in H$ such that the generalization error $R(h)$ is minimized.

2 SVMs — separable case

Support vector machines are one of the most theoretically well-motivated and practically most effective binary classification algorithms. We first introduce this algorithm for separable datasets, then present its general version for non-separable datasets.

Assumption 2.1. Consider a training sample of size m denoted by $z \in (\mathcal{X} \times \mathcal{Y})^m$ and define the disjoint sets $T_{-1} \triangleq \{i \in [m] : y_i = -1\}$ and $T_1 \triangleq \{i \in [m] : y_i = 1\}$. We assume that T_1, T_{-1} are non empty and can be separated by a hyperplane. That is, there exists a hyperplane $E_{w,b}$ such that $[m] = T_{-1} \cup T_1$ and

$$T_1 = \{i \in [m] : h(x_i) = \langle w, x_i \rangle + b > 0\}, \qquad T_{-1} = \{i \in [m] : h(x_i) = \langle w, x_i \rangle + b < 0\}.$$

Remark 5. For such a hyperplane $E_{w,b}$, we have $y_i h(x_i) > 0$ for all $i \in [m]$.

Remark 6. Let $E_{w,b}$ be one of infinite such planes. Which hyperplane should a learning algorithm select? The solution E_{w^*,b^*} returned by the SVM algorithm is the hyperplane with the maximum *margin*, or the distance to the closest points, and is thus known as the *maximum-margin hyperplane*.

2.1 Primal optimization problem

Assumption 2.1 confirms the existence of at least one pair (w, b) such that $\langle w, x_i \rangle + b \neq 0$ for all $i \in [m]$. For any training sample z and hyperplane $E_{w,b}$, we define the minimum $d_0(z, w, b) \triangleq \min_{i \in [m]} |\langle w, x_i \rangle + b|$. In terms of this minimum, we can write the margin, the minimum distance of sample z to the hyperplane $E_{w,b}$, as

$$\rho(w, b) \triangleq \min_{i \in [m]} d(z_i, E_{w,b}) = \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|} = \frac{d_0(z, w, b)}{\|w\|}.$$

Correct classification is achieved for a labeled example (x_i, y_i) when label $y_i = \text{sign}(\langle w, x_i \rangle + b)$. Since $|\langle w, x_i \rangle + b| \geqslant d_0(z, w, b)$ for all labeled example $z_i = (x_i, y_i)$, a correct classification is achieved when

$$y_i(\langle w, x_i \rangle + b) \geqslant d_0(z, w, b) \text{ for all } i \in [m].$$

SVM finds the maximum margin hyperplane that solves the following problem

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \frac{\|w\|^2}{d_0(z, w, b)^2} \\ \text{subject to:} \quad & y_i(\langle w, x_i \rangle + b) \geqslant d_0(z, w, b) \text{ for all } i \in [m]. \end{aligned}$$

We observe that the solution to this optimization problem remains unchanged to scaling of (w, b). Normalizing the pair (w, b) by $d_0(z, w, b)$, we get the *canonical hyperplane* $E_{w,b}$ such that $\min_{i \in [m]} |\langle w, x_i \rangle + b| = 1$. The *marginal hyperplanes* are defined to be parallel to the separating hyperplane and passing through the closest points on the negative or positive sides. Since they are parallel to the separating hyperplane, they admit the same normal vector w. By the definition of a canonical representation, for a point x on a marginal hyperplane, $|\langle w, x \rangle + b| = 1$, and thus the marginal hyperplanes are $\langle w, x \rangle + b = \pm 1$. Hence our original problem statement translates to finding (w, b) so as to maximize the margin ρ such that all points are correctly separated is equivalent to

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ \text{subject to:} \quad & y_i(\langle w, x_i \rangle + b) \geqslant 1 \text{ for all } i \in [m]. \end{aligned} \tag{1}$$

The objective function $F : w \mapsto \frac{1}{2} \|w\|^2$ is infinitely differentiable, its gradient is $\nabla_w(F) = w$ and its Hessian is the identity matrix $\nabla^2 F(w) = \mathbf{I}$ with strictly positive eigenvalues. Therefore, $\nabla^2 F(w) \succ 0$ and F is strictly convex. The constraints are all defined by the affine functions $g_i : (w, b) \mapsto 1 - y_i(\langle w, x_i \rangle + b)$ and are thus qualified. Thus the optimization problem in (1) has a unique solution, and can be solved by a *quadratic program*.

2.2 Support vectors

In this section, we will show that the normal vector w to the resulting hyperplane is a linear combination of some feature vectors, referred to as *support vectors*. Consider the dual variables $\alpha_i \geqslant 0$ for all $i \in [m]$

associated to the m affine constraints and let $\alpha \triangleq (\alpha_i : i \in [m])$. Then, we can define the Lagrangian for all canonical pairs $(w,b) \in \mathbb{R}^{N+1}$ and Lagrange dual variables $\alpha \in \mathbb{R}^m_+$ as

$$\mathcal{L}(w,b,\alpha) \triangleq \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i [y_i(\langle w, x_i\rangle + b) - 1].$$

Since the primal problem in (1) has convex cost function with affine constraints, E_{w^*,b^*} is the optimal separating cannonical hyperplane if and only if there exists $\alpha^* \in \mathbb{R}^m_+$ that satisfies the following three KKT conditions:

$$\nabla_w \mathcal{L}|_{w=w^*} = w^* - \sum_{i=1}^{m} \alpha_i y_i x_i = 0, \qquad \nabla_b \mathcal{L}|_{b=b^*} = -\sum_{i=1}^{m} \alpha_i^* y_i = 0, \qquad \alpha_i^* [y_i(\langle w^*, x_i\rangle + b^*) - 1] = 0.$$

Remark 7. The complementary condition implies that $\alpha_i^* = 0$ if the labeled points are not on the supporting hyperplane, i.e. $y_i(\langle w^*, x_i\rangle + b^*) \neq 1$.

Definition 2.2 (Support vectors). We can define the **support vectors** as the examples or feature vectors for which the corresponding Lagrange variable $\alpha_i^* \neq 0$, i.e.

$$S \triangleq \{i \in [m] : \alpha_i^* \neq 0\} \subseteq \{i \in [m] : y_i(\langle w^*, x_i\rangle + b^*) = 1\}.$$

Remark 8. The optimal primal variables (w^*, b^*) in the SVM solution are the stationary points of the associated Lagrangian, and hence we can write the normal vector as a linear combination of support vectors, i.e.

$$w^* = \sum_{i=1}^{m} \alpha_i^* y_i x_i = \sum_{i \in S} \alpha_i^* y_i x_i. \tag{2}$$

Remark 9. Support vectors completely determine the maximum-margin hyperplane solution. Vectors not lying on the marginal hyperplane do not affect the definition of these hyperplanes.

Remark 10. The slope of the hyperplane w^* is unique but the support vectors are not unique. A hyperplane is sufficiently determined by $N+1$ points in N dimensions. Thus, when more than $N+1$ points lie on a marginal hyperplane, different choices are possible for the $N+1$ support vectors.

Remark 11. We have expressed the normal vector w^* for the optimal hyperplane, in terms of the optimal dual variable $\alpha^* \in \mathbb{R}^m_+$. We have not yet found the optimal dual variables, or the normalized distance b^*.

2.3 Dual optimization problem

In this section, we will show that the hypothesis $h \in H$ and distance b can be expressed as inner products. To this end, we look at the the dual form of the constrained primal optimization problem (1). Recall that the dual function $F(\alpha) = \inf_{w,b} \mathcal{L}(w,b,\alpha)$. The Lagrangian $\mathcal{L}$ is minimized at the optimal primal variables (w^*, b^*) such that $\nabla_w \mathcal{L}(w^*, b^*) = \nabla_b \mathcal{L}(w^*, b^*) = 0$ to write the optimal normal vector $w^* = \sum_{i=1}^m \alpha_i y_i x_i$ in terms of the dual variables $\alpha \in \mathbb{R}^m_+$ as expressed in (2), together with the constraint $\sum_{i=1}^m \alpha_i y_i = 0$.

Definition 2.3 (Gram matrix). For a labeled sample $z \in (\mathcal{X} \times \mathcal{Y})^m$, we can define a Gram matrix $A \in \mathbb{R}^{m \times m}$ defined by the (i,j)th entries $A_{ij} \triangleq \langle y_i x_i, y_j x_j \rangle$ for all $i, j \in [m]$.

Remark 12. The matrix A is the Gram matrix associated with vectors $(y_1 x_1, \ldots, y_m x_m)$ and hence is positive semidefinite. We can easily check that for any $\alpha \in \mathbb{R}^m$, we have

$$\alpha^T A \alpha = \sum_{i,j \in [m]} \langle \alpha_i y_i x_i, \alpha_j y_j x_j \rangle = \left\| \sum_{i \in [m]} \alpha_i y_i x_i \right\|^2 \geqslant 0.$$

Substituting $w^* = \sum_{i=1}^m \alpha_i y_i x_i$, $\sum_{i=1}^m \alpha_i y_i = 0$, and the definition of Gram matrix A, in the Lagrangian $\mathcal{L}(w^*, b^*, \alpha)$, we can write the dual function as $F(\alpha) = \mathcal{L}(w^*, b^*, \alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \alpha_i A_{ij} \alpha_j$. Therefore, we can write the dual SVM optimization problem as

$$\max_{\alpha} \ \|\alpha\|_1 - \frac{1}{2}\alpha^T A \alpha \tag{3}$$
$$\text{subject to:} \ \ \alpha_i \geqslant 0, \text{ for all } i \in [m], \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0.$$

The objective function $G : \alpha \mapsto \|\alpha\|_1 - \frac{1}{2}\alpha^T A \alpha$ is infinitely differentiable, and its Hessian is given by $\nabla^2 G = -A \preceq 0$, and hence G is a concave function. Since the constraints are affine and convex, the dual maximization problem (3) is equivalent to a convex optimization problem. Since G is a quadratic function of Lagrange variables α, this dual optimization problem is also a quadratic program, as in the case of the primal optimization. Since the constraints are affine, they are qualified and strong duality holds. Thus, the primal and dual problems are equivalent, i.e., the solution α^* of the dual problem (3) can be used directly to determine the hypothesis returned by SVMs. Using (2) for the normal to the supporting hyperplane, we can write the hypothesis

$$h(x) = \text{sign}(\langle w^*, x \rangle + b^*) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i^* y_i \langle x_i, x \rangle + b^*\right).$$

For any support vector x_i for $i \in S$, we have $y_i = \langle w^*, x_i \rangle + b^*$, and hence we can write for all $j. \in S$

$$b^* = y_j - \sum_{i=1}^{m} \alpha_i^* y_i \langle x_i, x_j \rangle. \tag{4}$$

Combining the above two results, we get for any $j \in S$

$$h(x) = \text{sign}\left(\sum_{i \in S} \alpha_i^* y_i \langle x_i, x - x_j \rangle + y_j\right).$$

Remark 13. The hypothesis solution depends only on inner products between vectors and not directly on the vectors themselves.

Remark 14. Since (4) holds for all $i \in S$, that is for all i such that $\alpha_i^* \neq 0$, we can write

$$0 = \sum_{i=1}^{m} \alpha_i^* y_i b^* = \sum_{i=1}^{m} \alpha_i^* y_i^2 - \sum_{i,j=1}^{m} \alpha_i^* A_{i,j} \alpha_j^* = \sum_{i=1}^{m} \alpha_i^* - \|w^*\|^2.$$

That is, we can write the optimal margin ρ as $\rho^2 = \frac{1}{\|w^*\|^2} = \frac{1}{\|\alpha^*\|_1}$.