Lecture-02: Binary Classification

1 Definitions

Definition 1.1 (Input space). The set if all possible *examples* or *instances* is called the **input space** and denoted by $\mathcal{X} \subseteq \mathbb{R}^N$ with $N \geqslant 1$.

Definition 1.2 (Output space). The set of all possible *labels* or *targets* is called the **output space** and denoted by $\mathcal{Y}$. For binary classification, $\mathcal{Y} = \{0,1\}$ or $\{-1,1\}$.

Definition 1.3 (Concepts). A mapping from input space to output space is called a **concept** and denoted by $c : \mathcal{X} \to \mathcal{Y}$. When $\mathcal{Y} = \{0,1\}$, any concept c can be identified by the set $A_c = \{x \in \mathcal{X} : c(x) = 1\}$ such that $c(x) = \mathbb{1}_{\{x \in A_c\}} = \mathbb{1}_{\{A_c\}}(x)$.

The set of all true concepts is called the **concept class** and denoted by C.

Example 1.4. The set of all triangles, rectangles, circles, lines in the plane are all examples of concept classes.

Definition 1.5 (Hypothesis). The set of all possible candidate concepts is called the **hypothesis class** and denoted by $H \subseteq (\mathcal{Y}')^{\mathcal{X}}$. A *consistent* hypothesis set contains the concept to learn, and an *inconsistent* hypothesis set doesn't contain it.

Assumptions 1.6. All examples in $\mathcal{X}$ are identically and independently distributed (*iid*) with a fixed but unknown underlying distribution D.

Definition 1.7 (Sample). We have a **sample** $S = (x_i \in \mathcal{X} : i \in [m])$ of size m generated *iid* according to the distribution D. For a concept $c : \mathcal{X} \to \mathcal{Y}'$, we have a **labeled sample** $T = ((x_i, y_i) \in \mathcal{X} \times \mathcal{Y}' : i \in [m])$ such that $y_i = c(x_i)$.

Definition 1.8 (Generalization error). Given a hypothesis $h \in H$, target concept c, and an underlying distribution D from which an example X is generated *iid*, the generalization error or risk of hypothesis h is defined as

$$R(h) \triangleq P[h(X) \neq c(X)] = \mathbb{E}\mathbb{1}_{\{h(X) \neq c(X)\}}.$$

Definition 1.9 (Supervised learning). The **supervised learning** is selection of a hypothesis $h_T \in H$ to minimize the generalization error with respect to c. That is,

$$h_T = \arg\min_{h \in H} R(h).$$

Remark 1. The generalization error of a hypothesis is not directly accessible to the learner since both the distribution D and concept c are unknown. However, one can measure the *empirical error* of a hypothesis on the labeled sample T.

Definition 1.10 (Empirical error). For a hypothesis $h \in H$, a target concept $c \in C$, and a sample $S = (x_i \in \mathcal{X} : i \in [m])$ The *empirical error* is defined as

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{h(x_i) \neq c(x_i)\}}.$$

Remark 2. The empirical error of a hypothesis is the average error over the sample S, while the generalization error is the expected error based on the distribution D. We see that $\mathbb{E}\hat{R}(h) = R(h)$ by the linearity of expectations. We will see later that $\hat{R}(h) \approx R(h)$ with high probability.

2 Support Vector Machines

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

2.1 Linear 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 $c : \mathcal{X} \to \mathcal{Y}$.

Assumptions 2.1. We define the hypothesis set as a collection of separating hyperplanes

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

The labeled training sample of size m denoted by $T = ((x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [m])$ where $y_i = c(x_i)$ for each $i \in [m]$, and each example of the sample is generated *iid* by the distribution D.

The objective is to select an $h \in H$ such that the generalization error $R_D(h)$ is minimized, where

$$R_D(h) = P\left[h(x) \neq c(x)\right].$$

Remark 3. Any hypothesis $h \in H$ is of the form $x \mapsto \text{sign}(\langle w, x \rangle + b)$ and labels positively all points falling on one side of the hyperplane $\langle w, x \rangle + b = 0$ and labels negatively all others. This problem is referred to as **linear classification problem**.

3 SVMs — separable case

Assumptions 3.1. The training sample T can be separated into two non-empty sets by a hyperplane. In other words, there exists a hyperplane $\langle w, x \rangle + b = 0$ such that $T = T_1 \cup T_2$ and $T_1 \cap T_2 = \emptyset$ where

$$T_1 = \left\{(x, y) \in T : \langle w, x \rangle + b > 0\right\}, \qquad T_2 = \left\{(x, y) \in T : \langle w, x \rangle + b < 0\right\}.$$

Let $\langle w, x \rangle + b = 0$ be one of infinitely such planes. Which hyperplane should a learning algorithm select? The solution 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**.

3.1 Primal optimization problem

The assumption above confirms the existence of at least one pair (w, b) such that $\langle w, x \rangle + b \neq 0$. We can normalize the pair (w, b) by the scalar $\min_{(x,y)\in T} |\langle w, x \rangle + b|$, such that if the closest point is $x_0 \in S$, then

$$|\langle w, x_0 \rangle + b| = 1.$$

We define this representation of the hyperplane $\langle w, x \rangle + b = 0$ as the **canonical hyperplane**. The distance of any point $x_0 \in \mathbb{R}^N$ to a hyperplane is given by

$$d(x_0, \langle w, x \rangle + b = 0) = \frac{|\langle w, x \rangle + b|}{\|w\|}.$$

This is due to the fact that $w/\|w\|$ is the unit vector normal to the hyperplane, and the distance of the hyperplane from the origin is $-b/\|w\|$. Hence, the distance of any point $x_0 \in \mathbb{R}^N$ from the hyperplane $\langle w, x \rangle + b = 0$ is given by the distance between it's projection to the unit vector $w/\|w\|$ and $-b/\|w\|$ which equals

$$\left| \left\langle \frac{w}{\|w\|}, x_0 \right\rangle + \frac{b}{\|w\|} \right| = \frac{|\langle w, x_0 \rangle + b|}{\|w\|}.$$

Let ρ be the minimum distance of any point to the plane, i.e

$$\rho \triangleq \min_{(x,y)\in T} \frac{|\langle w, x \rangle + b|}{\|w\|} = \frac{1}{\|w\|}.$$

The maximizing the margin is equivalent to minimizing the norm $\|w\|$ or $\frac{1}{2}\|w\|^2$.

We can show in a figure the margin for a maximum-margin hyperplane with a canonical representation (w, b). We also see the **marginal hyperplanes**, 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 $\langle w, x \rangle + b = \pm 1$. Correct classification is achieved for a

labeled point $(x,y) \in T$ when $y = \text{sign}(\langle w,x\rangle + b)$. Since $|\langle w,x\rangle + b| \geqslant 1$ for all labeled points $(x,y) \in T$ by the definition of canonical hyperplanes, a correct classification is achieved when

$$y(\langle w,x\rangle + b) \geqslant 1 \text{ for all } (x,y) \in T.$$

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

$$\begin{aligned} \min_{w,b} \ & \frac{1}{2}\|w\|^2 \\ \text{subject to:} \ & y(\langle w,x\rangle + b) \geqslant 1 \text{ for all } (x,y) \in T. \end{aligned} \tag{1}$$

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 $\nabla^2 F(w) = I$ 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 $g_i : (w,b) \mapsto 1 - y(\langle w,x\rangle + b)$ and are thus qualified. Thus the optimization problem in (1) has a unique solution, and can be solved by a **quadratic program**.

3.2 Support vectors

Consider the Lagrange 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 variables $\alpha \in \mathbb{R}^m_+$ as

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

The KKT conditions are obtained by setting the gradient of the Lagrangian with respect to the primal variables w and b to zero, and by writing the complementary conditions:

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

This implies the SVM solution (w^*,b^*), the weight vector $w^* = \sum_{i=1}^m \alpha_i y_i x_i$ is a linear combination of the training examples $(x_1,\ldots,x_m)$, with that vector appearing in the summation if $\alpha_i \neq 0$.

Definition 3.2 (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.

$$V(S) = \{x_i \in S : \alpha_i \neq 0\} \subseteq \{x_i \in S : \langle w,x_i\rangle + b = 1\}.$$

By the complementarity condition, if $x_i \in V$, then $y_i(\langle w,x_i\rangle + b) = 1$, and hence the support vectors lie on the marginal hyperplanes $\langle w,x_i\rangle + b = \pm 1$. That is,

$$w^* = \sum_{x_i \in V} \alpha_i y_i x_i. \tag{2}$$

Remark 4. Support vectors completely determine the maximum-margin hyperplane solution. Vectors not lying on the marginal hyperplane or $V(S)$ do not affect the definition of these hyperplanes.

Remark 5. 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.

3.3 Dual optimization problem

To derive the dual form of the constrained primal optimization problem (1), we substitute the definition of optimal vector w^* in terms of the dual variables as expressed in (2) and apply the constraint $\sum_{i=1}^m \alpha_i y_i = 0$ into the Lagrangian, to get

$$\mathcal{L} = \frac{1}{2}\left\|\sum_{i=1}^m \alpha_i y_i x_i\right\|^2 - \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle x_i,x_j\rangle + \sum_{i=1}^m \alpha_i = \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \alpha_i A_T(ij) \alpha_j,$$

where the matrix $A_T = (y_i \langle x_i, x_j \rangle y_j : i,j \in [m])$ is the Gram matrix associated with vectors $(y_1 x_1, \ldots, y_m x_m)$ and hence is positive semidefinite. We can write the dual SVM optimization problem as

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i A_T(i,j) \alpha_j \tag{3}$$

$$\text{subject to:} \quad \alpha_i \geqslant 0, \text{ for all } i \in [m], \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0.$$

The objective function $G : \alpha \mapsto \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i A_T(i,j) \alpha_j$ is infinitely differentiable, and its Hessian is given by $\nabla^2 G = -A_T \preceq 0$, 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 the equation (2) for the normal to the supporting hyperplane

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

For any $x_i \in V(S)$, we have $y_i = \langle w, x_i \rangle + b$, and hence we can write

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

Remark 6. The hypothesis solution depends only on inner products between vectors and not directly on the vectors themselves.

Since the equation (4) holds for all $x_i \in V(S)$, that is for all i such that $\alpha_i \neq 0$, we can write

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

That is, we can write the margin ρ as

$$\rho^2 = \frac{1}{\|w\|_2^2} = \frac{1}{\|\alpha\|_1}.$$

3.4 Leave-one-out analysis

Now we will look at some results that show us why SVMs work well in practice.

Definition 3.3 (Leave-one-out error). Given a sample S of size m and a hypothesis h_S the **leave-one-out error** is defined as

$$\hat{R}_{LOO}(h_S) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\left\{ h_{S \setminus \{x_i\}}(x_i) \neq y_i \right\}}.$$

Lemma 3.4. *The average leave-one-out error for samples of size m is an unbiased estimate of the average generalization error for samples of size $m-1$. That is,*

$$\mathbb{E}_{S \sim D^m}[\hat{R}_{LOO}(h_S)] = \mathbb{E}_{S' \sim D^{m-1}} R(h_{S'}).$$

Proof. Using the linearity of expectation,

$$\mathbb{E}\hat{R}_{LOO}(h_S) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\mathbf{1}_{\{h_{S \setminus \{x_i\}}(x_i) \neq y_i\}} = \mathbb{E}_{S \sim D^m} \mathbf{1}_{\{h_{S \setminus \{x_1\}}(x_1) \neq y_1\}} = \mathbb{E}_{S' \sim D^{m-1}} \mathbb{E}_{x_1 \sim D^m} \mathbf{1}_{\{h_{S \setminus \{x_1\}}(x_1) \neq y_1\}} = \mathbb{E}_{S' \sim D^{m-1}} R(h_{S'}).$$

Theorem 3.5. *Let S be a linearly separable sample of size $m+1$ and $N_V(S) = |V(S)|$ be the number of support vectors that define the hypothesis h_S returned by the SVM. Then*

$$\mathbb{E}_{S' \sim D^m} R(h_{S'}) = \mathbb{E}_{S \sim D^{m+1}} \frac{|V(S)|}{m+1}$$

Proof. We will first show that $\hat{R}_{LOO}(h_S) \leqslant \frac{|V(S)|}{m+1}$. Let $x \in V(S)$, then $h_{S \setminus \{x\}} = h_S$ and it correctly classifies x. Contrapositively, if $h_{S \setminus \{x\}}$ misclassifies x, then x must be a support vector. Hence,

$$\sum_{i=1}^{m+1} \mathbf{1}_{\{h_{S \setminus \{x_i\}}(x_i) \neq y_i\}} \leqslant |V(S)|.$$

Then by taking expectation on both sides and applying the previous lemma, we get the desired result.

A Convex optimization

Definition A.1 (Gradient). Let $f : X \subset \mathbb{R}^N \to \mathbb{R}$, then

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_N}(x) \end{bmatrix}$$

Definition A.2 (Hessian). Let $f : X \subset \mathbb{R}^N \to \mathbb{R}$, then

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{1 \leqslant i,j \leqslant N}$$

Definition A.3 (Stationary Point). If f attains a local extremum at $x' = x$ then $\nabla f(x') = 0$. x' is called a stationary point.

Definition A.4 (Convex Set). A set X is called convex if $\forall x, y \in X$ and $\alpha \in [0,1]$,

$$\alpha x + (1-\alpha) y \in X$$

Definition A.5 (Convex Hull). A convex hull of a set A is the smallest convex set including A.

$$conv(A) = \left\{ \sum_{x_i \in A} \alpha_i x_i : 0 \leqslant \alpha_i \leqslant 1, \sum \alpha_i = 1 \right\}$$

Definition A.6 (Convex Function). Let $\mathcal{X} \subset \mathbb{R}^N$ be a convex set. Then $f : \mathcal{X} \to \mathbb{R}$ is a convex function if

$$f(\alpha x + (1-\alpha) y) \leqslant \alpha f(x) + (1-\alpha) f(y)$$

If f is differentiable then it is convex if

$$f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle$$

If f is twice differentiable then it is convex if

$$\nabla^2 f \geqslant 0$$

Or in other words, if $\nabla^2 f$ is a positive semi-definite matrix.