Lecture-04: SVMs — non-separable case

1 SVMs — non-separable case

In most practical settings, the given sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ is not linearly separable. For $\mathcal{X} \subseteq \mathbb{R}^N$ and $\mathcal{Y} \subseteq \mathbb{R}$, it would not be possible to draw a hyperplane in $\mathbb{R}^N$ that perfectly separates the two sets of points. More precisely, for any canonical hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$, there exists $i \in [m]$ such that

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) < 1.$$

To minimize the number of such points we can try to find a hyperplane that minimizes the empirical error,

$$\min_{\mathbf{w},b} \sum_{i=1}^{m} \mathbb{1}_{\{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) < 1\}}.$$

This optimization problem is NP-hard in the dimension of the space and cannot be solved efficiently. Moreover we would like to work with a smooth function to optimize. The constraints imposed in the linearly separable case discussed in the linearly spearable case cannot all hold simultaneously. However, a relaxed version of these constraints can indeed hold, where for each example $i \in [m]$, there exists a **slack variable** $\xi_i \geqslant 0$ such that

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geqslant 1 - \xi_i.$$

A slack variable ξ_i measures the distance by which feature vector x_i violates the desired inequality, $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geqslant 1$.

Definition 1.1 (Outliers). For a hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$, a feature vector $\mathbf{x}_i$ with slack variable $\xi_i > 0$ is an **outlier**. The set of outliers O is defined as

$$O \triangleq \{i \in [m] : 1 - \xi_i \leqslant y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) < 1\} = \{i \in [m] : \xi_i > 0\}.$$

Remark 1. Each example $\mathbf{x}_i$ must be positioned on the correct side of the appropriate marginal hyperplane to not be considered an outlier. As a consequence, a feature vector $\mathbf{x}_i$ with $0 < y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) < 1$ is correctly classified by the hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ but is nonetheless considered to be an outlier, that is, $\xi_i > 0$.

Remark 2. If we omit the outliers, the training data is correctly separated by $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ with a margin $\rho = \frac{1}{\|\mathbf{w}\|}$ that we refer to as the **soft margin**, as opposed to the **hard margin** in the separable case.

Remark 3. How should we select the hyperplane in the non-separable case? One idea consists of selecting the hyperplane that minimizes the empirical error. We have already rejected that idea due to the complexity considerations. We have conflicting objectives here. On the one hand, we need to minimize the total slack due to the outliers, measured by $\|\xi\|_p^p = \sum_{i=1}^m \xi_i^p$, for some $p \geqslant 1$. On the other hand, we wish to maximize the margin for non-outliers. Larger margin can lead to more outliers and hence larger slack. Hence, these two are conflicting objectives.

1.1 Primal optimization problem

We define a primal problem by deciding on a trade-off between these two objectives for the non-seperable case, where $C \geqslant 0$ is the trade-off parameter between margin-maximization and the slack penalty. The parameter C is determined by n-fold cross validation for a given dataset. For $\xi \in \mathbb{R}_+^m$, the primal problem is

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\|\xi\|_p^p \\ &\text{subject to } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0, \text{ for all } i \in [m]. \end{aligned} \tag{1}$$

As in the separable case, the objective function is convex and the constraints are affine. Therefore, the primal problem in (1) is a convex optimization problem. In particular, $\xi \mapsto \sum_{i=1}^{m} \xi_i^p = \|\xi\|_p^p$ is convex in view of the convexity of the norm $\|\cdot\|_p$. There are many possible choices for p leading to more or less aggressive penalizations of the slack terms. The choices $p=1$ and $p=2$ lead to the most straightforward solutions and analyses.

Definition 1.2 (Hinge loss). The loss functions associated with $p=1$ and $p=2$ are called the **hinge loss** and the **quadratic hinge loss**, respectively.

Remark 4. Both hinge losses are convex upper bounds on the zero-one loss, thus making them well suited for optimization. We first observe that for all $p \geqslant 1$.

$$\mathbb{1}_{\{x<0\}} \leqslant (1-(x \wedge 1))^p.$$

Recall that a labeled point $(\mathbf{x}, y)$ is incorrectly labeled if $y(\langle \mathbf{w}, \mathbf{x} \rangle + b) < 0$. From the definition of the slack variable ξ, we have $1 - \xi \leqslant y(\langle \mathbf{w}, \mathbf{x} \rangle + b) < 1$. Therefore, we observe that

$$\mathbb{1}_{\{y(\langle \mathbf{w}, \mathbf{x} \rangle + b) < 0\}} \leqslant (1 - (y(\langle \mathbf{w}, \mathbf{x} \rangle + b) \wedge 1))^p \leqslant (1-(1-\xi) \wedge 1)^p = \xi^p.$$

In what follows, the analysis is presented in the case of the hinge loss ($p=1$), which is the most widely used loss function for SVMs.

1.2 Support vectors

In this section, we will show that the normal vector $\mathbf{w}$ to the resulting hyperplane is a linear combination of some feature vectors, referred to as *support vectors*. Consider the dual variable $\alpha, \beta \in \mathbb{R}_+^m$ associated to the m affine relaxed separation constraints and m non negativity constraint on slack variables. Then, we can write the Lagrangian for all canonical pairs $(\mathbf{w}, b) \in \mathbb{R}^{N+1}$ and Lagrange dual variables $\alpha, \beta \in \mathbb{R}_+^m$ as

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\|\xi\|_1 - \sum_{i=1}^{m} \alpha_i (y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \xi_i) - \sum_{i=1}^{m} \beta_i \xi_i. \tag{2}$$

Similar to the separable case, the constraints in the primal problem in (1) are affine and thus qualified. In addition, the objective function as well as the affine constraints are convex and differentiable. It follows that $E_{\mathbf{w}^*, b^*}$ is the optimal separating cannonical hyperplane if and only if there exists $\alpha^*, \beta^* \in \mathbb{R}_+^m$ that satisfies the following three KKT conditions. The first KKT condition is obtained by taking the gradient of Lagrangian with respect to primal variables and equating it to zero, to get

$$\nabla_{\mathbf{w}} \mathcal{L}|_{\mathbf{w}=\mathbf{w}^*} = \mathbf{w}^* - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0, \qquad \nabla_b \mathcal{L}|_{b=b^*} = -\sum_{i=1}^{m} \alpha_i^* y_i = 0, \qquad \nabla_\xi \mathcal{L}|_{\xi=\xi^*} = C - \alpha - \beta = 0.$$

The next KKT condition is obtained by setting the derivative with respect to dual variables, being less than or equal to zero. This is equivalent to constraints being satisfied, i.e. for all $i \in [m]$

$$\nabla_\alpha \mathcal{L} = -y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) + 1 - \xi_i^* \leqslant 0, \qquad \nabla_\beta \mathcal{L} = -\xi^* \leqslant 0.$$

The final KKT conditions looks at the complementary condition, which results in $\sum_{i=1}^{m} \alpha_i^*(y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) - 1 + \xi_i^*) = 0$ and $\sum_{i=1}^{m} \beta_i^* \xi_i^* = 0$. Since $\alpha, \beta \in \mathbb{R}_+^m$, together with second condition of KKT, it follows that the each term of the two summation is positive. Therefore, it means that for all $i \in [m]$

$$\alpha_i^*[y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) - 1 + \xi_i^*] = 0, \qquad \beta_i^* \xi_i^* = 0.$$

Remark 5. The complementary condition implies that $\alpha_i^* = 0$ if $y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) \neq 1 - \xi_i^*$.

Definition 1.3 (Support vectors). An example of feature vector is a **support vector** if the corresponding relaxed constraint Lagrange variable $\alpha_i^* \neq 0$, i.e. $S \triangleq \{i \in [m] : \alpha_i^* \neq 0\} \subseteq \{i \in [m] : y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) = 1 - \xi_i^*\}$.

Remark 6. If for some feature vector $\mathbf{x}_i \in S$ and the corresponding slack variable $\xi_i^* = 0$, then $y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) = 1$ and the example $\mathbf{x}_i$ lies on a marginal hyperplane, as in the separable case. Otherwise, $\xi_i^* \neq 0$ and $\mathbf{x}_i$ is an outlier. In this case, the complementary KKT condition implies that $\beta_i^* = 0$ and hence $\alpha_i^* = C$. Thus, support vectors $\mathbf{x}_i$ are either outliers, in which case $\alpha_i^* = C$, or they lie on the marginal hyperplanes. That is, we can write the support vector as a union of disjoint sets

$$S = \{i \in S : \xi_i^* = 0\} \cup \{i \in S : \xi_i^* > 0\} = \{i \in S : y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) = 1\} \cup \{i \in S : \alpha_i^* = C\}.$$

Remark 7. As in the separable case, note that while the weight vector $\mathbf{w}^*$ solution is unique, the support vectors are not.

1.3 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 $F(\alpha) = \inf_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \alpha)$. The Lagrangian $\mathcal{L}$ is minimized at the optimal primal variables $(\mathbf{w}^*, b^*)$ such that

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^*, b^*, \xi^*) = \nabla_b \mathcal{L}(\mathbf{w}^*, b^*, \xi^*) = \nabla_\xi \mathcal{L}(\mathbf{w}^*, b^*, \xi^*) = 0.$$

Using this condition, we can write the optimal normal vector $\mathbf{w}^* = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$ in terms of the dual variables $\alpha \in \mathbb{R}_+^m$, together with the constraints $\sum_{i=1}^m \alpha_i y_i = 0$ and $C = \alpha_i + \beta_i$ for all $i \in [m]$.

Definition 1.4 (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 $A_{ij} \triangleq \langle y_i \mathbf{x}_i, y_j \mathbf{x}_j \rangle$ for all $i, j \in [m]$.

Remark 8. The matrix A is the Gram matrix associated with vectors $(y_1 \mathbf{x}_1, \dots, y_m \mathbf{x}_m)$ and hence is positive semidefinite.

Substituting $\mathbf{w}^* = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$, the constraints $\sum_{i=1}^m \alpha_i y_i = 0$ and $C = \alpha + \beta$, and the definition of Gram matrix A, in the Lagrangian $\mathcal{L}(\mathbf{w}^*, b^*, \alpha)$, we can write the dual function as $F(\alpha) = \mathcal{L}(\mathbf{w}^*, b^*, \alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i A_{ij} \alpha_j$. The constraints are $\alpha_i \geqslant 0$ together with $\beta_i \geqslant 0$ to get $\alpha_i \leqslant C$, and $\sum_{i=1}^m \alpha_i y_i = 0$. Therefore, we can write the dual SVM optimization problem as

$$\max_{\alpha} \; \|\alpha\|_1 - \frac{1}{2} \alpha^T A \alpha \tag{3}$$

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

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. The solution α^* of the dual problem can be used to return the SVM hypothesis

$$h(x) = \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle + b^*) = \text{sign}\left(\sum_{j=1}^m \alpha_j^* y_j \langle \mathbf{x}_j, \mathbf{x} \rangle + b^* \right).$$

Recall that for all $x_i \in S \cap \{\xi_i = 0\}$, we have $\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^* = 1$. Hence, the constant b^* is given by

$$b^* = y_i - \sum_{j=1}^m \alpha_j^* y_j \langle \mathbf{x}_j, \mathbf{x}_i \rangle, \text{ for any } \mathbf{x}_i \text{ such that } 0 < \alpha_i^* < C.$$