Lecture-03: SVMs — non-separable case

1 SVMs — non-separable case

In most practical settings, the given sample is not linearly separable. That is, 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 $\langle w,x\rangle + b = 0$, there exists $x_i \in S$ such that

$$y_i(\langle w,x_i\rangle + b) \leqslant 1$$

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

$$\min_{w,b} \sum_{i=1}^{m} \mathbb{1}_{\{y_i(\langle w,x_i\rangle + b)\leqslant 1\}}.$$

But 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, that is, for each example $i \in [m]$, t here exist **slack variables** $\xi_i \geqslant 0$ such that

$$y_i(\langle w,x_i\rangle + b) \geqslant 1 - \xi_i.$$

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

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

$$O \triangleq \{x_i \in S : 1 - \xi_i \leqslant y_i(\langle w,x_i\rangle + b) \leqslant 1\} = \{x_i \in S : \xi_i > 0\}..$$

Remark 1. Each example 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 x_i with $0 < y_i(\langle w,x_i\rangle + b) < 1$ is correctly classified by the hyperplane $\langle w,x\rangle + b = 0$ 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 $\langle w,x\rangle + b = 0$ with a margin $\rho = \frac{1}{\|w\|}$ that we refer to as the **soft margin**, as opposed to the **hard margin** in the separable case.

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 $\sum_{i=1}^{m} \xi_i^p$, 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. For $\xi = (\xi_1, \ldots, \xi_m)$, the primal problem is

$$\min_{w,b,\xi} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i^p \tag{1}$$
$$\text{subject to } y_i(\langle w,x_i\rangle + b) \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0, \text{ for all } i \in [m].$$

The parameter C is determined by n-fold cross validation for a given dataset.

As in the separable case, the equation (1) is a convex optimization problem since the constraints are affine and thus convex and since the objective function is convex for any $p \geqslant 1$. 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.

Both hinge losses are convex upper bounds on the zero-one loss, thus making them well suited for optimization. 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

We denote the Lagrange variables associated with the relaxed separation constraint by $\alpha = (\alpha_1, \ldots, \alpha_m)$ and the non-negative constraint of slack variables by $\beta = (\beta_1, \ldots, \beta_m)$. Then, we can write the Lagrangian as

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

Similar to the separable case, the constraints are affine and thus qualified. The objective function as well as the affine constraints are convex and differentiable. Thus, the KKT conditions apply at the optimum. We can write the first KKT conditions as

$$w^* = \sum_{i=1}^{m} \alpha_i y_i x_i, \qquad \sum_{i=1}^{m} \alpha_i y_i = 0,$$

and the next three KKT condition for all $i \in [m]$ as

$$\alpha_i + \beta_i = C, \qquad \alpha_i . (y_i (\langle w, x_i \rangle + b) - 1 + \xi_i) = 0, \qquad \beta_i \xi_i = 0.$$

Definition 1.3 (Support vectors). An example is called a **support vector** if the corresponding relaxed constraint Lagrange variable $\alpha_i \neq 0$. We can write the set of support vectors as

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

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

$$V(S) = \{x_i \in V(S) : \xi_i = 0\} \cup \{x_i \in V(S) : \xi_i > 0\} = \{x_i \in S : y_i (\langle w, x_i \rangle + b) = 1\} \cup \{x_i \in S : \alpha_i = C\}.$$

Remark 3. As in the separable case, note that while the weight vector w solution is unique, the support vectors are not.

1.3 Dual optimization problem

Substituting for the w in terms of the support vectors, we get

$$\mathcal{L} = \|\alpha\|_1 - \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i x_i \right\|_2^2.$$

The constraints are $\alpha_i \geqslant 0$ together with $\beta_i \geqslant 0$ to get $\alpha_i \leqslant C$, and $\sum_{i=1}^{m} \alpha_i y_i = 0$. Hence, the dual problem is

$$\begin{aligned} \min_{\alpha} \quad & \|\alpha\|_1 - \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i x_i \right\|_2^2 \\ \text{subject to } & \sum_{i=1}^{m} \alpha_i y_i = 0, \text{ and } 0 \leqslant \alpha_i \leqslant C, \text{ for all } i \in [m]. \end{aligned} \tag{3}$$

The objective function is concave and infinitely differentiable and is equivalent to a convex quadratic. The pro program. This problem is equivalent to the primal problem. The solution α of the dual problem can be used to return the SVM hypothesis

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

Recall that for all $x_i \in V(S) \cap \{\xi_i = 0\}$, we have $\langle w, x_i \rangle + b = 1$. Hence, the constant b is given by

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

A Review of Linear Algebra

A.1 Vector Space

A vector space over the field $\mathbb{R}$ is a set V equipped with following two operations, each satisfying four axioms.

A.1.1 Vector addition

Vector addition is a mapping $+ : V \times V \to V$ defined by $+(v,w) = v+w$ for any two elements $v,w \in V$, that satisfies the following four axioms.

1. Associativity of addition : $u+(v+w) = (u+v)+w$; for all $u,v,w \in V$
2. Commutativity of addition : $u+v = v+u$; for all $u,v \in V$
3. Existence of Identity: There exists a zero vector ($0 \in V$) s.t, $u+0 = u$; for all $u \in V$
4. Existence of Inverse: For every $u \in V$, there exists an element $-u \in V$; s.t, $u+(-u) = 0$

A.1.2 Scalar Multiplication

Scalar multiplication is a mapping $\cdot : \mathbb{R} \times V \to V$ defined by $\cdot(\alpha, v) = \alpha v \in V$, that satisfies the following four axioms.

1. Compatibility with the field: $a(bu) = (ab)u$; for all $a,b \in \mathbb{R}$ and $u \in V$
2. Existence of Identity : For multiplicative identity element $1 \in \mathbb{R}$, $1u = u$; for all $u \in V$
3. Distributivity over vector addition : $\alpha(vu) = \alpha u + \alpha v$; for all $\alpha \in \mathbb{R}$ and $u,v \in V$
4. Distributivity over field addition : $(\alpha+\beta)u = \alpha u + \beta u$; for all $\alpha,\beta \in \mathbb{R}$ and $u \in V$

Example A.1 (Vector space). zz Following are some common examples of vector spaces.

1. Space of all real numbers $\mathbb{R}$.
2. Euclidean space of N-dimensions, denoted by $\mathbb{R}^N$.
3. Space of continuous functions over a compact subset $[a,b]$ denoted by $C([a,b])$.

A.2 Inner Product Space

A *inner product space* is a vector space equipped with an inner product denoted by $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the following axioms.

1. **Symmetry:** $\langle x,y \rangle = \langle y,x \rangle$
2. **Linearity:** $\langle \alpha x + \beta y, z \rangle = \alpha \langle x,z \rangle + \beta \langle y,z \rangle$
3. **Definiteness:** $\langle x,x \rangle \geqslant 0$; $\langle x,x \rangle = 0$ iff $x = 0$

Example A.2 (inner product spaces). Following are some common examples of inner product spaces.

1. For the vector space $V = \mathbb{R}^N$, we can define the inner product between two N-dimensional vectors as

$$\langle x,y \rangle = \langle \left[x_1,\ldots,x_N\right]^T, \left[y_1,\ldots,y_N\right]^T \rangle_. = x^T y = \sum_i^N x_i y_i.$$

2. For vector space $V = C(\mathbb{R}^N)$, we can define the inner product of two continuous functions over $\mathbb{R}^N$ as

$$\langle f,g \rangle = \int_{\mathbb{R}^N} (f,g)(t)dt.$$

3. For the vector space of random variables, we can define the inner product of two random variables as

$$\langle X,Y \rangle = \mathbb{E}(XY).$$

A.3 Norms

Norm is a mapping $\|\cdot\| : V \to \mathbb{R}_+$ that satisfy the following axioms.

1. **Definiteness:** $\|v\| = 0$ iff $v = 0$
2. **Homogeneity:** $\|\alpha v\| = |\alpha| \|v\|$
3. **Triangle inequality:** $\|v+w\| \leqslant \|v\| + \|w\|$

Example A.3 (Norms). Following are examples of commonly defined norms on some example vector spaces.

1. $V = \mathbb{R}; \|X\| = |X|$
2. $V = \mathbb{R}^N; \|X\|_p = \left(\sum_{i=1}^N |X_i|^p\right)^{\frac{1}{p}}$
3. $V = \mathbb{R}^N; \|X\|_2 = \left(\sum_{i=1}^N |X_i|^2\right)^{\frac{1}{2}}$

Proposition A.4 (Holder's Inequality). *Let $p,q \geqslant 1$ be conjugate, i.e. $\frac{1}{p} + \frac{1}{q} = 1$ Then,*

$$|\langle x,y\rangle| \leqslant \|x\|_p \|y\|_p \text{ for all } x,y \in \mathbb{R}^N.$$

Proof. For any positive $a,b \in \mathbb{R}$ and conjugate pair $p,q \geqslant 1$ such that $1/p + 1/q = 1$, we have from the concavity of log

$$\ln\left(\frac{1}{p}a^p + \frac{1}{q}b^p\right) \geqslant \frac{1}{p}\ln a^p + \frac{1}{q}\ln b^p = \ln ab.$$

Since $\ln(\cdot)$ is an increasing function, the above inequality implies the Young's inequality $\frac{1}{p}a^p + \frac{1}{q}b^p \geqslant ab$.

The Holder's inequality is trivially true if $x = 0$ or $y = 0$. Hence, we assume that $\|x\| \|y\| > 0$, and let $a = \frac{|x_i|}{\|x\|_p}$ and $b = \frac{|y|_i}{\|y\|_p}$. From Young's inequality, we have

$$\frac{|x|_i}{p\|x\|_p^p} + \frac{|y|_i}{q\|y\|_q^q} \geqslant \frac{|x|_i |y|_i}{\|x\|_p \|y\|_q}, \text{ for all } i \in [N].$$

Since $|\langle x,y\rangle| \leqslant \sum_{i=1}^N |x_i||y_i|$, we get the result by summing both sides over $i \in [N]$ in the above inequality.

B Review of Convex Optimization

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a function over N-dimensional reals. Then, we can write its Taylor series expansion around the neighborhood of $x \in \mathbb{R}^N$ as

$$f(y) = f(x) + \sum_{i=1}^N \frac{\partial f}{\partial x_i}(y_i - x_j) + \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \frac{\partial^2 f}{\partial x_i \partial x_j}(y_i - x_i)(y_j - x_j) + o(\|y-x\|_2^2).$$

We can define the gradient vector $\nabla f = \begin{bmatrix}\frac{\partial f}{\partial x_1} & \ldots & \frac{\partial f}{\partial x_N}\end{bmatrix}^T$, and the Hessian $\nabla^2 f \in \mathbb{R}^{N\times N}$ such that $[\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$, to observe

$$f(y) = f(x) + \nabla f^T(y-x) + (y-x)^T \nabla^2 f(y-x) + o(\|y-x\|_2^2).$$

B.1 Convex Function

Let $\mathcal{X} \subseteq \mathbb{R}^N$. For a function $f : \mathcal{X} \to \mathbb{R}$, we define its epigraph as

$$Epi(f) \triangleq \{(x,y) \in \mathbb{R}^N \times \mathbb{R} : y \geqslant f(x)\}.$$

Definition B.1. A function $f : \mathbb{R}^N \to \mathbb{R}$ is convex if it's $\text{dom}(f)$ is convex and $Epi(f)$ is convex.

Note 1. For a convex function $f(\cdot)$; $f(\alpha x + \bar{\alpha} y) \leqslant \alpha f(x) + \bar{\alpha} f(y)$ where $\alpha + \bar{\alpha} = 1$.

- If f is differentiable then f is convex iff

1. $\text{dom}(f)$ is convex
2. $f(y) - f(x) \geqslant \langle \nabla f(x), y - x \rangle$; for all $x, y \in \text{dom}(f)$

Proof : $f(y) - f(x) = \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) \geqslant \langle \nabla f(x), y - x \rangle$.

- If f is twice differentiable then f is convex iff $\text{dom}(f)$ is convex and it's Hessian is positive semi definite : $\nabla^2 f(x) \succeq 0$; for all $x \in \text{dom}(f)$

Example B.2. Convex Function

1. Linear Function: $f(x) = \langle w, x \rangle$; where $f : \mathbb{R}^N \to \mathbb{R}$
2. Quadratic Function: $f(x) = x^T Ax$; where A is positive semi definite
3. Abs Maximum $f(x) = \max |X|_{i \in N} = \|X\|_\infty$

Lemma B.3. *Composition of Functions*
Let, $h(.) : \mathbb{R} \to \mathbb{R}; g(.) : \mathbb{R}^N \to \mathbb{R}$ and $f : \mathbb{R}^N \to \mathbb{R}$; for all $x \in \mathbb{R}^N$ where $f(x)$ is defined by $f(x) = h(g(x))$, then following inequalities are valid

1. *If h is a convex and non decreasing and g is convex, $\Longrightarrow f(.)$ is convex*
   ***Proof**: As $g(.)$ is convex : $g(\alpha x + \bar{\alpha} y) \leqslant \alpha g(x) + \bar{\alpha} g(y)$*
   Now, $h(g(\alpha x + \bar{\alpha} y)) \leqslant h(\alpha g(x) + \bar{\alpha} g(y)) \leqslant \alpha h(g(x)) + \bar{\alpha} h(g(y))$(Proved.)
2. *If h is a convex and non increasing and g is concave, $\Longrightarrow f(.)$ is convex*
3. *If h is a concave and non decreasing and g is concave, $\Longrightarrow f(.)$ is concave*
4. *If h is a concave and non increasing and g is convex, $\Longrightarrow f(.)$ is concave*

Theorem B.4. ***Jensen's Inequality***
Let $X \in C \subset \mathbb{R}^N$ be a r.v with finite mean and $f : C \to \mathbb{R}$ is convex,
Then $\mathbb{E}[X] \in C$, $\mathbb{E}[f(X)] \leqslant \infty$ and $f(\mathbb{E}[X]) \leqslant \mathbb{E}[f(X)]$

Proof: $f(\sum_{i=1}^{m} \alpha_i x_i) \leqslant \sum_{i=1}^{m} \alpha_i f(x_i)$; where α_is could be interpreted as probabilities as $\alpha_i \geqslant 0$ and $\sum_{i=1}^{m} \alpha_i = 1$

B.2 Constrained Optimization

Let $f : \mathbb{R}^N \to \mathbb{R}$ and $g_i : \mathbb{R}^N \to \mathbb{R}, i \in [m]$
Principle Optimization Problem: $\min f(x)$ s.t. $g_i(x) \leqslant 0$; for all $i \in [m]$
Note 2. Let p^* be the optimum value for the above problem.

Definition B.5. Lagrangian
If $x \in \mathbb{R}^N$ and $\alpha \in \mathbb{R}^M_+$, then Lagrangian $\mathcal{L}(x, \alpha) : \mathbb{R}^N \times \mathbb{R}^M_+ \to \mathbb{R}$ associated with the principal problem is defined as, $\mathcal{L}(x, \alpha) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x)$; The variables $\alpha \in \mathbb{R}^M_+$ are called Lagrange or Dual Variables.

Definition B.6. Dual Function

Dual function associated with the Principal Optimization Problem is defined as $F : \mathbb{R}^M_+ \to \mathbb{R}$ defined as
$F(\alpha) = \inf \mathcal{L}(x, \alpha)$ where $x \in \mathbb{R}^N$

Remark 4. Important Properties of Dual Function

1. F is concave in α
2. $F(\alpha) \leqslant \mathcal{L}(x, \alpha) \leqslant f(x)$
3. $F(\alpha) \leqslant \inf_{x \in \mathbb{R}^N} f(x) = p^*$
4. $F(\alpha) \leqslant p^*$ such that $g_i(x) \leqslant 0$

B.2.1 Dual Problem:

Dual Problem associated with Principal Optimization Problem is as follows
Max $F(\alpha)$; such that $\alpha \in \mathbb{R}_{+}^{M}$

Note 3. Let d^* be the optimal value of this dual problem.

Remark 5. Dual Function

1. Dual problem is always convex.
2. $d^* \leqslant p^*$
3. $(p^* - d^*)$ is called duality gap. When $d^* = p^*$, it is known as strong duality. It holds for convex optimization problems where constraints are qualifying.

Definition B.7. Strong Constraint Qualification:
Assume that $\mathbf{int}(\mathcal{X}) \neq \phi$, then the strong constraint qualification or **Slater's Condition** is defined as, there exists $\bar{x} \in int(\mathcal{X})$, such that $g(\bar{x}) < 0$

Definition B.8. Weak Constraint Qualification: Assume that $\mathbf{int}(\mathcal{X}) \neq \phi$, then the strong constraint qualification or **weak Slater's Condition** is defined as there exists $\bar{x} \in int(\mathcal{X})$: for all $i \in [1,m], (g_i(\bar{x}) < 0) \vee (g_i(\bar{x}) = 0 \wedge g_i$ affine)

Theorem B.9. *Saddle Point: Sufficient Condition*
Let P be a constrained optimum problem over $\mathcal{X} = \mathbb{R}^N$ If (x^*, α^*) is a saddle point of the associated Lagrangian, i.e. for all $x \in \mathbb{R}^N$, for all $\alpha \geqslant 0$, $\mathcal{L}(x^*, \alpha)) \leqslant \mathcal{L}(x^*, \alpha^*) \leqslant \mathcal{L}(x, \alpha^*)$ Then, (x^*, α^*) is a saddle point of P.

Theorem B.10. *Saddle point-Necessary Condition*

- *Assume that f and g_i, $i \in [1,m]$ are convex functions and Slater's condition holds, then if x is a solution of the constrained optimization problem, then there exists $\alpha \geqslant 0$ s.t (x, α) is a saddle point of the Lagrangian.*
- *Assume that f and g_i, $i \in [1,m]$ are convex differentiable functions and weak Slater's condition holds, then if x is a solution of the constrained optimization problem, then there exists $\alpha \geqslant 0$ s.t (x, α) is a saddle point of the Lagrangian.*

Theorem B.11. *Karush-Kuhn-Tucker's Theorem*
Let $f, g_i : \mathcal{X} \to \mathbb{R}$, for all $i \in [1,m]$ are convex and differentiable function and that the constrains are qualified. Then $\bar{x}$ is a solution of the constrained problem iff there exists $\bar{\alpha} \geqslant 0$ s.t.

- $\nabla_x \mathcal{L}(\bar{x}, \bar{\alpha}) = \nabla_x f(\bar{x}) + \langle \bar{\alpha}, \nabla_x g(\bar{x}) \rangle = 0$
- $\nabla_\alpha \mathcal{L}(\bar{x}, \bar{\alpha}) = g(\bar{x}) \leqslant 0$
- $\langle \bar{\alpha}, g(\bar{x}) \rangle = 0$