CS 412 — Introduction to Machine Learning (UIC) February 27, 2025

Lecture 12

Instructor: Aadirupa Saha *Scribe(s): (Lokesh Boggavarapu)*

Overview

In the last lecture, we covered the following main topics:

1. Max-Margin Formulation
2. SVM objective

This lecture focuses on:

1. KKT Conditions
2. Dual Optimization
3. SVM with strong duality

1 KKT Conditions

1.1 What are KKT Conditions?

KKT (Karush-Kuhn-Tucker) conditions are used in optimization problems with constraints. They help find the best solution by balancing the goal and the limits.

Let's understand the math behind it.

Consider the general optimization problem (P) shown below, where we have not assumed anything regarding the functions f,g,h,l (like convexity). Here P is called primal and G is called dual of P.

$$\boldsymbol{P}$$

$$\min_{x} f(x)$$

subject to

$$h_i(x) \leq 0,\ i = 1, \ldots, m$$

$$l_j(x) = 0,\ j = 1, \ldots, r$$

$$\boldsymbol{G}$$

$$\max_{u,v} \min_{x} \left(f(x) + \sum_{i=1}^{m} u_i\, h_i(x) + \sum_{j=1}^{r} v_j\, l_j(x) \right)$$

subject to

$$u \geq 0$$

The primal problem P seeks to minimize f(x) under given constraints.
The dual problem G introduces Lagrange multipliers u_i, v_j to incorporate constraints into the objective. The dual problem aims to find the best lower bound for the primal problem.This duality framework is fundamental in optimization, helping to find bounds and sometimes exact solutions for constrained problems.
The KKT conditions associated with problem P are:

1. Stationarity Condition : The gradient of the goal and the constraints must be balanced.

$$\frac{\partial}{\partial(x)}\left(f(x)+\sum_{i=1}^{m} u_i\, h_i(x)+\sum_{j=1}^{r} v_j\, l_j(x)\right)=0$$

2. Primal feasibility : The solution must satisfy the constraints.

$$h_i(x) \leq 0,\, i=1,\dots,m$$

$$l_j(x)=0,\, j=1,\dots,r$$

3. Dual feasibility : The multipliers for constraints must be non-negative.

$$u_i \geq 0,\, i=1,\dots,m$$

4. Complementary slackness : If a constraint is not tight (i.e., $h_i(x) < 0$), its Lagrange multiplier u_i must be zero.

$$u_i h_i(x)=0, i=1,\dots,m$$

Example 1. *Consider the problem*

$$f(x)=x^2$$

$$\min_x f(x)$$

$$\textit{subject to } x \geq 1$$

1. ***Constraint****:* $h(x)=1-x \leq 0$
2. ***Lagrangian****:* $\mathcal{L}(x,u)=x^2+u(1-x)$

Applying KKT Conditions:

1. ***Stationarity:***

$$\frac{\partial \mathcal{L}}{\partial x}=2x-u=0 \quad \Rightarrow \quad u=2x.$$

2. ***Primal Feasibility:***

$$x \geq 1.$$

3. ***Dual Feasibility:***

$$u \geq 0.$$

4. ***Complementary Slackness:***

$$u(1-x) = 0.$$

Solving:

- *If $x > 1$, then $1 - x < 0$. By complementary slackness, $u = 0$. But from stationarity, $u = 2x$, which would require $x = 0$. This contradicts $x \geq 1$.*
- *If $x = 1$, then $u = 2(1) = 2$, which satisfies all conditions.*

Optimal Solution: $x = 1, u = 2$.

This example shows how the KKT conditions work together to find the optimal solution while respecting the constraints.

2 Dual Optimization

The above section introduced the concepts of primal, dual, and lagrangian. In this section, let us understand them in detail. In many optimization problems, directly solving the original (primal) problem can be challenging. Hence, we transform the problem into its **dual form**, which is often easier to solve.
Dual optimization helps in:

- Simplifying complex problems.
- Converting constrained problems into unconstrained ones.
- Deriving strong theoretical properties like **strong duality**, ensuring the primal and dual problems yield the same optimal value under certain conditions.

2.1 Primal and Dual Problems

A general constrained optimization problem (Primal problem) is:

$$\min_x f(x) \quad \text{subject to} \quad h_i(x) \leq 0, \quad l_j(x) = 0$$

where:

- $f(x)$ is the objective function to minimize.
- $h_i(x) \leq 0$ are inequality constraints.
- $l_j(x) = 0$ are equality constraints.

Dual Formulation

We define the **Lagrangian function** as:

$$\mathcal{L}(x, u, v) = f(x) + \sum_i u_i h_i(x) + \sum_j v_j l_j(x)$$

where:

- $u_i \geq 0$ are **Lagrange multipliers** for the inequality constraints.
- v_j are **Lagrange multipliers** for the equality constraints.

The **Lagrange dual function** is:

$$g(u, v) = \min_x \mathcal{L}(x, u, v)$$

The **dual problem** is then:

$$\max_{u \geq 0, v} g(u, v)$$

This means we find the **best lower bound** for the primal problem.

Theorem 2.1: Weak Duality

For any feasible solution x in the primal and any feasible (u, v) in the dual, we have:

$$\min_x f(x) \geq \max_{u \geq 0, v} g(u, v)$$

This means the optimal value of the primal problem(f^*) is always **greater than or equal** to the optimal value of the dual problem(g^*).The difference between the optimal values($f^* - g^*$) is called duality gap.

Theorem 2.2: Strong Duality

If the **Slater's condition** holds (i.e., there exists a strictly feasible solution satisfying $h_i(x) < 0$), then:

$$\min_x f(x) = \max_{u \geq 0, v} g(u, v)$$

This means **primal and dual problems give the same optimal value i.e** $f^* = g^*$.

Example 2. *Consider the same optimization problem shown in Example 1:*

$$f(x) = x^2$$
$$\min_x f(x)$$
$$\textit{subject to } x \geq 1$$

Step 1: Form the Lagrangian

$$\mathcal{L}(x, u) = x^2 + u(1 - x)$$

where $u \geq 0$ is the Lagrange multiplier.

Step 2: Compute the Dual Function

$$g(u) = \min_x \mathcal{L}(x, u)$$

Taking derivative:

$$\frac{d}{dx}\left(x^2 + u(1-x)\right) = 2x - u = 0$$

Solving for x:

$$x = \frac{u}{2}$$

Plugging $x = \frac{u}{2}$ back into $\mathcal{L}(x, u)$:

$$g(u) = u - \frac{u^2}{4}$$

Step 3: Solve the Dual Problem

$$\max_{u \geq 0} \left(u - \frac{u^2}{4}\right)$$

Taking derivative:

$$1 - \frac{u}{2} = 0 \Rightarrow u = 2$$

*Since $u = 2$, we substitute in $x = \frac{u}{2}$, giving **optimal solution**:*

$$x^* = 1, \quad u^* = 2$$

*Substituting optimal values in their respective align*s, we get*

$$f^* = 1, \quad g^* = 1$$

*Thus, the **optimal primal and dual values match**, proving **strong duality**.*

Algorithm 2.1: Lagrange Dual Optimization Algorithm

```
Input: Primal problem $\min f(x)$ with constraints $h_i(x) \leq 0, l_j(x) = 0$.
Formulate the Lagrangian function $\mathcal{L}(x, u, v)$.
Compute the dual function $g(u, v) = \min_x \mathcal{L}(x, u, v)$.
Solve the dual problem $\max_{u \geq 0, v} g(u, v)$.
Check strong duality (if Slater's condition holds).
Recover primal variables from the optimal dual solution.
```

Theorem 2.3:

KKT conditions are satisfied for optimality. That is, if a point (x, u, v) satisfies the four KKT conditions, then it is optimal.

Theorem 2.4:

If strong duality holds, then the KKT conditions are necessary.

Takeaways

- In Support Vector Machines (SVM), strong duality holds.

3 Support Vector Machines(SVM)

As stated above, in SVM **strong duality** holds, which ensures that solving the *dual problem* gives the same optimal solution as the *primal problem*.

3.1 Primal Formulation of SVM

SVM finds a hyperplane that separates two classes with the largest possible margin. The optimization problem is formulated as:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$

subject to:

$$y_i(w^T x_i + b) \geq 1, \quad \forall i = 1, \ldots, N$$

where:

- w is the weight vector of the hyperplane,
- b is the bias term,
- x_i are training samples,
- $y_i \in \{+1, -1\}$ are class labels,
- N is the number of training samples.

This is a **convex optimization problem** with inequality constraints.

3.2 Strong Duality in SVM

To solve this problem efficiently, we use **Lagrange Duality**. Strong duality states that the optimal value of the *primal problem* is equal to the optimal value of its *dual problem*.

We introduce **Lagrange multipliers** $\alpha_i \geq 0$ for each constraint:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(w^T x_i + b) - 1]$$

where α_i are *dual variables*.

3.3 Applying KKT Conditions and dual problem transformation

The **Karush-Kuhn-Tucker (KKT) conditions** state that for optimality:

1. **Stationarity:**

$$\frac{\partial L}{\partial w} = 0 \quad \Rightarrow \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

2. **Primal feasibility:**

$$y_i(w^T x_i + b) \geq 1, \quad \forall i$$

3. **Dual feasibility:**

$$\alpha_i \geq 0, \quad \forall i$$

4. **Complementary slackness:**

$$\alpha_i \left(y_i(w^T x_i + b) - 1 \right) = 0, \quad \forall i$$

   This means that if $\alpha_i > 0$, then the corresponding constraint must be active (i.e., it lies on the margin).

The Lagrangian function is:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \right].$$

Substituting $\mathbf{w}$ into the Lagrangian

First, compute $\|\mathbf{w}\|^2$:

$$\|\mathbf{w}\|^2 = \mathbf{w}^\top \mathbf{w} = \left(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \right)^\top \left(\sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j \right).$$

Expanding:

$$\mathbf{w}^\top \mathbf{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j.$$

Thus,

$$\frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j.$$

Now, expand the constraint term:

$$\sum_{i=1}^{n}\alpha_i\left(y_i\sum_{j=1}^{n}\alpha_j y_j \mathbf{x}_i^\top \mathbf{x}_j + y_i b - 1\right).$$

Splitting the terms:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n}\alpha_i y_i b - \sum_{i=1}^{n}\alpha_i.$$

Since $\sum_{i=1}^{n}\alpha_i y_i = 0$, the second term vanishes. The final dual problem becomes:

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j.$$

subject to:

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad \forall i.$$

This is a **quadratic programming problem** that is easier to solve than the primal form.There are multiple methods to solve these kind of problems.They are Newton's method, Projected Gradient Descent and Barrier Method etc

3.4 SVM Decision Rule

Let α^* is the solution of the above quadratic problem. Now, the optimal weight vector is given by:

$$w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i$$

The bias term b^* can be computed using the complementary slackness condition by substituting w^* obtained above in place of w

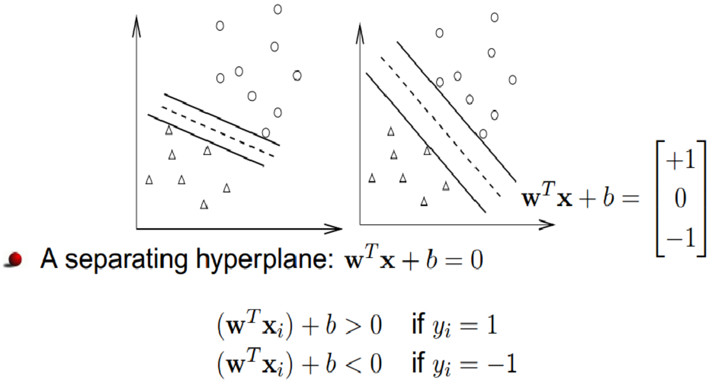
$$y_i(w^T x_i + b) = 1$$

After obtaining w^* and b^*,the final decision function for classification is:

$$f(x) = \text{sign}\left((w^*)^T \cdot x + b^*\right)$$

The hyperplane $\left((w^*)^T \cdot x + b^*\right) = 0$ is called max margin hyperplane.For a given new data point (x_{n+1}), y_{n+1} is calculated below

$$y_{n+1} = sign(w^{*T} \cdot x + b^*)$$


$\mathbf{w}^T\mathbf{x} + b = \begin{bmatrix} +1 \\ 0 \\ -1 \end{bmatrix}$

- A separating hyperplane: $\mathbf{w}^T\mathbf{x} + b = 0$

$$(\mathbf{w}^T\mathbf{x}_i) + b > 0 \quad \text{if } y_i = 1$$
$$(\mathbf{w}^T\mathbf{x}_i) + b < 0 \quad \text{if } y_i = -1$$

Figure 1: Different hyperplanes separate data points.

3.5 What are support vectors in SVM?

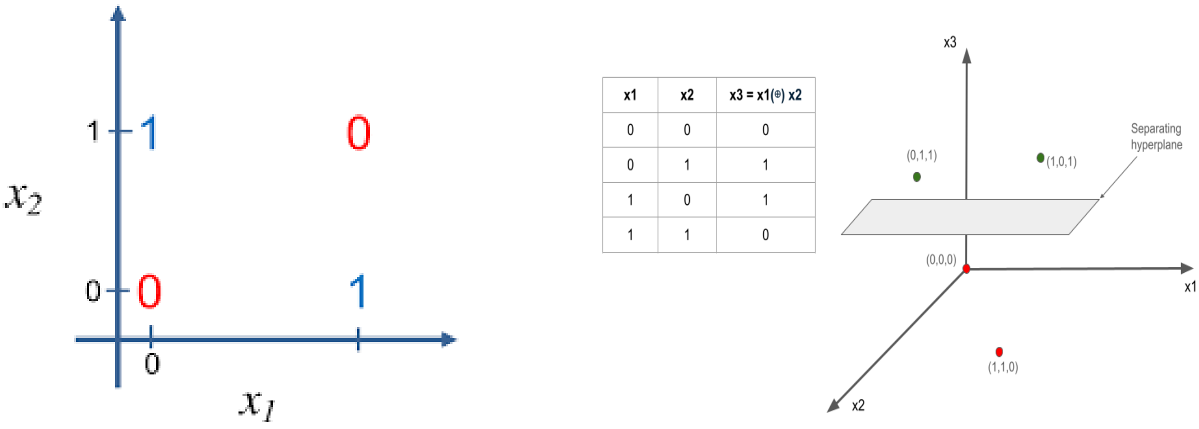
They are the data points that lie closest to the decision boundary (hyperplane) in a Support Vector Machine (SVM). These data points are important because they determine the position and orientation of the hyperplane, and thus have a significant impact on the classification accuracy of the SVM. In fact, SVMs are named after these support vectors because they “support” or define the decision boundary. The support vectors are used to calculate the margin, which is the distance between the hyperplane and the closest data points from each class. The goal of SVMs is to maximize this margin while minimizing classification errors.

3.6 Higher dimension representation

Consider the XOR function takes two binary inputs and outputs a value based on the following rule.If we plot these points in a 2D space where (x_1, x_2) are the coordinates and y is the class label, we see that there is no single straight line that can separate the two classes as shown in Figure 2a.

x_1	x_2	y (XOR Output)
0	0	0
0	1	1
1	0	1
1	1	0

Table 1: Truth table for XOR function



x1	x2	x3 = x1(⊕) x2
0	0	0
0	1	1
1	0	1
1	1	0

(a) XOR representation in 2D

(b) XOR representation in 3D

Figure 2: Illustration of how XOR becomes linearly separable when projected to 3D

To make the XOR data linearly separable, we consider the output y as one of the features(x3).This transformation allows us to apply a linear classifier in the new space, solving the problem.

Next Lecture

The next lecture will cover the following topics:
(i) Kernel SVM
(ii) Kernel Properties
(iii) SVM Examples

Exercise 3.1: KKT Conditions

Consider the following optimization problem:

$$\begin{aligned} \min_{x_1, x_2} \quad & f(x_1, x_2) = x_1^2 + 2x_2^2 \\ \text{subject to} \quad & x_1 + x_2 - 1 \leq 0, \\ & x_1 \geq 0, x_2 \geq 0. \end{aligned}$$

Find the KKT conditions and solve for the optimal values of x_1 and x_2.

References:

1. Support Vector Machines (SVM): An Intuitive Explanation Link
2. Lecture note of Cornell University Link
3. Lecture note by Prof. Adhirupa Saha from course CS412 - Intro to Machine Learning Link
4. Lecture note by Prof. Laurent El Ghaoui Link
5. Generative AI Link