Statistical Computing

May 26, 2021

Introduction

- Quadratic programming (QP).
- The **Lagrangian** formula.
- Support Vector Machine (SVM): Primal and Dual Problems.
- Hard margin and Soft margin.
- Nonlinear SVM: Kernel function.
- Part of materials and all Figures are from Prof. Andrew Zisserman's notes.
- Due to the copyright, the note can not be distributed.

General Optimization Problem: Standard Form and The Lagrangian

$$\begin{aligned} \text{minimize} \quad & J(\theta) \\ \text{subject to} \quad & f_i(\theta) \leq 0, \; i = 1, \ldots, q \\ & h_i(\theta) = 0, \; i = 1, \ldots, p \end{aligned}$$

where

- θ are the **optimization variables** and J is the **objective function**.
- Assume **domain** $\mathcal{D} = \mathbf{dom} J \cap_{i=1}^{q} \mathbf{dom} f_i \cap_{i=1}^{p} \mathbf{dom} h_i$ is nonempty.
- The set of points satisfying the constraints is called the **feasible set**.
- A point θ in the feasible set is called a **feasible point**.
- The **optimal value** p^* of the problem is defined as

$$p^* = \inf \left\{ J(\theta) \mid f_i(\theta) \leq 0, i = 1, \ldots, q, \; h_i(\theta) = 0, i = 1, \ldots, p \right\}.$$

- θ^* is an **optimal point** (or a solution to the problem) if θ^* is feasible and $J(\theta^*) = p^*$.
- **Quadratic programming (QP)**: $J(\theta)$ is a quadratic objective function in θ with linear constraints $f_i(\theta) \leq 0$ and $h_i(\theta) = 0$.

The **Lagrangian** for the general optimization problem is

$$L(\theta, \mu, \lambda) = J(\theta) + \sum_{i=1}^{q} \mu_i f_i(\theta) + \sum_{i=1}^{p} \lambda_i h_i(\theta),$$

- $\mu \geq 0$ and λ are called **Lagrange multipliers**.
- $\mu \geq 0$ and λ also called the **dual variables**.

The Lagrangian

Supremum over Lagrangian gives back objective and constraints:

$$\sup_{\mu \geq 0, \lambda} L(\theta, \mu, \lambda) = \sup_{\mu \geq 0, \lambda} \left(J(\theta) + \sum_{i=1}^{q} \mu_i f_i(\theta) + \sum_{i=1}^{p} \lambda_i h_i(\theta) \right)$$
$$= \begin{cases} J(\theta) & f_i(\theta) \leq 0 \text{ and } h_i(\theta) = 0, \text{ for all } i \\ \infty & \text{otherwise} \end{cases}$$

- **Primal form** of optimization problem:

$$p^* = \inf_{\theta} \sup_{\mu \geq 0, \lambda} L(\theta, \mu, \lambda)$$

- **Dual problem**:

$$d^* = \sup_{\mu \geq 0, \lambda} \inf_{\theta} L(\theta, \mu, \lambda)$$

- **Weak duality**: $p^* \geq d^*$ for any optimization problem.

Proof.

$$p^* = \inf_{\theta} \sup_{\mu \geqslant 0, \lambda} \left[J(\theta) + \sum_{l=1}^{q} \mu_i f_i(\theta) + \sum_{i=1}^{p} \lambda_i h_i(\theta) \right]$$
$$\geqslant \sup_{\mu \geqslant 0, \lambda} \inf_{\theta} \left[J(\theta) + \sum_{l=1}^{q} \mu_i f_i(\theta) + \sum_{i=1}^{p} \lambda_i h_i(\theta) \right] = d^*$$

$\square$

The Lagrange Dual Form

- The difference $p^* - d^*$ is called the **duality gap**.
- **Strong duality**: $p^* = d^*$.
- The **Lagrangian dual problem**:

$$d^* = \sup_{\mu \succeq 0, \lambda} \underbrace{\inf_{\theta} L(\theta, \mu, \lambda)}_{\text{Lagrange dual function}}$$

Definition

The **Lagrange dual function** is

$$g(\mu, \lambda) = \inf_{\theta \in \mathcal{D}} L(\theta, \mu, \lambda) = \inf_{\theta \in \mathcal{D}} \left(J(\theta) + \sum_{i=1}^{q} \mu_i f_i(\theta) + \sum_{i=1}^{p} \lambda_i h_i(\theta) \right)$$

- Weak duality

$$p^* \geq \sup_{\mu \geq 0, \lambda} g(\mu, \lambda) = d^*$$

- **Lagrange dual function gives a lower bound on optimal solution:**

$$g(\mu, \lambda) \leq p^*$$

The Lagrange Dual Problem and Strong Duality

$$\begin{array}{ll} \text{maximize} & g(\mu, \lambda) \\ \text{subject to} & \mu \succeq 0. \end{array}$$

- (μ, λ) **dual feasible** if $\mu \succeq 0$ and $g(\mu, \lambda) > -\infty$.
- (μ^*, λ^*) are **dual optimal** or **optimal Lagrange multipliers** if they are optimal for the Lagrange dual problem.
- Lagrange dual problem often easier to solve (simpler constraints).

If **strong duality** is held, like SVM, we get an interesting relationship between

- the optimal Lagrange multiplier μ_i and the ith constraint at the optimum: $f_i(\theta^*)$
- Relationship is called "**complementary slackness**":

$$\mu_i^* f_i(\theta^*) = 0.$$

Proof: Complementary Slackness

- Assume strong duality: $p^* = d^*$.
- Let θ^* be primal optimal and (μ^*, λ^*) be dual optimal. Then:

$$\begin{aligned}
J(\theta^*) &= g(\mu^*, \lambda^*) \\
&= \inf_\theta \left(J(\theta) + \sum_{l=1}^{q} \mu_i^* f_i(\theta) + \sum_{i=1}^{p} \lambda_i^* h_i(\theta) \right) \\
&\leqslant J(\theta^*) + \sum_{i=1}^{q} \underbrace{\mu_i^* f_i(\theta^*)}_{\leqslant 0} + \sum_{i=1}^{p} \underbrace{\lambda_i^* h_i(\theta^*)}_{=0} \\
&\leqslant J(\theta^*).
\end{aligned}$$

Each term in $\sum_{i=1} \mu_i^* f_i(\theta^*)$ must actually be 0. That is

$$\mu_i^* f_i(\theta^*) = 0, i = 1, \ldots, q.$$

This condition is known as **complementary slackness**.

Motivation of SVM

Given training data $(\mathbf{x}_i, y_i)$ for $i = 1, ..., n$, with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$, learn a classifier $f(\mathbf{x})$ such that

$$f(\mathbf{x}_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$

- i.e. $y_i f(\mathbf{x}_i) > 0$ or $y_i = \mathrm{sign}\{\mathrm{f}(\mathrm{x_i})\}$ for a correct classification.
- Try to find $f(X)$, such that

$$\min_f E_{testing\ data}[Y \neq sign\{f(X)\}].$$

Considering a linear classifier: How to define the best w ?

A linear classifier has the form: $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$.

- Maximum margin solution: most stable under perturbations of the inputs

Intuition to find w

- Since $\mathbf{w}^T\mathbf{x} + b = 0$ and $c(\mathbf{w}^T\mathbf{x} + b) = 0$ define the same plane, choose normalization such that $\mathbf{w}^T\mathbf{x}_+ + b = +1$ and $\mathbf{w}^T\mathbf{x}_- + b = -1$ for the positive and negative support vectors respectively.
- The fact: the distance between two parallel planes $\mathbf{w}^T\mathbf{x} + b_1 = 0$ and $\mathbf{w}^T\mathbf{x} + b_2 = 0$ is

$$\frac{|b_1 - b_2|}{\|w\|},$$

  where $\|w\| = \|w\|_2$.
- Thus, the margin (the distance between two parallel planes $\mathbf{w}^T\mathbf{x} + b = +1$ and $\mathbf{w}^T\mathbf{x} + b = -1$) is given by $\frac{2}{\|\mathbf{w}\|}$.

Optimization: the Primal and Dual Problem

- Maximizing the margin under the constraint can be formulated as an optimization:

$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|} \text{ subject to } \mathbf{w}^T\mathbf{x}_i + b \begin{array}{ll} \geq 1 & \text{if } y_i = +1 \\ \leq -1 & \text{if } y_i = -1 \end{array} \quad \text{for } i = 1, \ldots, n$$

- Or equivalently

$$\min_{\mathbf{w},b} 2^{-1}\|\mathbf{w}\|^2 \text{ subject to } y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \text{ for } i = 1, \ldots, n$$

- This is a quadratic optimization problem subject to linear constraints and there is a unique minimum.
- The Lagrange function is

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= 2^{-1}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left\{ y_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 \right\} \\ &= 2^{-1}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i \left\{ y_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 \right\}, \end{aligned}$$

  where $\alpha_i, i = 1 \ldots, n$, are lagrange multipliers and $\alpha_i \geq 0$.
- Lagrange dual function is the inf over primal variables of L:

$$\begin{aligned} L_d(\alpha) &= \inf_{\mathbf{w},b} L(\mathbf{w}, b, \alpha) \\ &= \inf_{\mathbf{w},b} \left[2^{-1}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} \alpha_i \left\{ y_i \left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 \right\} \right], \end{aligned}$$

Optimization: the Primal and Dual Problem, Conti,

- Setting the derivatives with respect to $\mathbf{w}$ and b to zero, we have

$$\begin{aligned} \partial L(\mathbf{w}, b, \alpha)/\partial \mathbf{w} &= \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i x_i = 0, \\ \partial L(\mathbf{w}, b, \alpha)/\partial b &= -\sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}$$

which implies

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{n} \alpha_i y_i x_i, \\ 0 &= \sum_{i=1}^{n} \alpha_i y_i. \end{aligned}$$

- If we know $\mathbf{w}$, we know all $\alpha_i, i = 1, \ldots, n$; if we know all α_i, we know $\mathbf{w}$.

SVM: Dual Function

- Substituting these results back into $L(\mathbf{w}, b, \alpha)$, we have,

$$\begin{aligned} \frac{1}{2}w^T w &= \frac{1}{2}\sum_{i=1}^{n}\alpha_i y_i x_i^T \sum_{j=1}^{n}\alpha_j y_j x_j = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j, \\ \sum_{i=1}^{n}\alpha_i\{1 - y_i(w^T x_i + b)\} &= \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j - b\underbrace{\sum_{i=1}^{n}\alpha_i y_i}_{=0}. \end{aligned}$$

- Putting it together, the dual function is

$$L_d(\alpha) = \begin{cases} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j, & \sum_{i=1}^{n}\alpha_i y_i = 0, \alpha_i \geq 0, \text{ all } i \\ -\infty, & \text{otherwise.} \end{cases}$$

Karush, Kuhn and Tucker (KKT) Conditions

Considering the optimization problem with contraints

$$\text{problem } \mathcal{P} = \begin{cases} \min_\theta J(\theta) \\ \text{with } h_j(\theta) = 0,\ j = 1, \ldots, p \\ \text{and } g_i(\theta) \leq 0\ i = 1, \ldots, q \end{cases}$$

Definition: Karush, Kuhn and Tucker (KKT) conditions

- stationarity: $\nabla J(\theta^*) + \sum_{j=1}^{p} \lambda_j \nabla h_j(\theta^*) + \sum_{i=1}^{q} \mu_i \nabla g_i(\theta^*) = 0$
- primal admissibility: $h_j(\theta^*) = 0,\ j = 1, \ldots, p$
  $g_i(\theta^*) \leq 0,\ i = 1, \ldots, q$
- dual admissibility: $\mu_i \geq 0,\ i = 1, \ldots, q$
- complementarity: $\mu_i g_i(\theta^*) = 0, i = 1, \ldots, q.$

λ_j and μ_i are called the Lagrange multipliers of problem $\mathcal{P}$.

KKT conditions for SVM:

- stationarity: $\mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = \mathbf{0}$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$
- primal admissibility: $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1,\ i = 1, \ldots, n.$
- dual admissibility: $\alpha_i \geq 0,\ i = 1, \ldots, n.$
- complementarity: $\alpha_i \big(y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 \big) = 0,\ i = 1, \ldots, n.$

Optimization: the Dual Problem

Substituting the above results into the lagrange function (L_p), we obtain the so-called dual problem by maximizing L_d with the constraints:

- $$L_d = \sum_{i=1}^{n} \alpha_i - 2^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j, \text{ subject to } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0.$$

- The solution is obtained by maximizing L_d subject to these constraints and the solution is a very large quadratic programming (QP) optimization problem.
- Sequential minimal optimization (SMO, Platt, 1998) breaks this large QP problem into a series of smallest possible QP problems (coordinate descent approach).
- In addition, by KKT conditions, we have

$$\alpha_i \left\{ y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) - 1 \right\} = 0, \forall i.$$

- From the above result, we can find that if $\alpha_i > 0$, then $y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) = 1$. So x_i is on the boundary of the slab.
- On the other hand, if $y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) > 1$, then x_i is not on the boundary of the slab and $\alpha_i = 0$.

Optimization: the Dual Problem, Cont.

- From this view, **w** is obtained only from these data points with $\alpha_i > 0$ or these data points on the boundary of the slab.
- b can be estimated from

$$\alpha_i \left\{ y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) - 1 \right\} = 0, \forall\, i.$$

  after **w** and α_i are estimated which is equivalent to estimate b by using the fact that $b = y_i - \mathbf{w}^T \mathbf{x}_i$ for $\alpha_i > 0$.
-

$$\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i y_i x_i^T x + \widehat{b}.$$

- $\widehat{f}(x)$ only depends on $\widehat{\alpha}_i > 0$ and can be calculated based on inputs x_i through their inner products (similarities) with other inputs.

Extension the concept of linear separability but allowing error: What is the best w?

Original approach: the points can be linearly separated but there is a very narrow margin.
Allowing error: possibly the large margin solution is better, even though one constraint is violated.
In general there is a trade off between the margin and the number of mistakes on the training data.

Slack Variable

- Define the slack variables $\xi = (\xi_1, \ldots, \xi_n)$ and $\xi_i \geq 0$, $\forall i$.
- Modify the original constraint as $y_i \left(\mathbf{w}^T \mathbf{x}_i + b\right) \geq 1 - \xi_i$ with an additional constraint $\sum_{i=1}^{n} \xi \leq$ constant .
- For $\xi_i = 0$, the constraint is defined as usual.
- For $0 < \xi_i \leq 1$, the point i is between margin and correct side of hyperplane if

$$\mathbf{w}^T \mathbf{x}_i + b \begin{array}{ll} \geq 1 - \xi_i & \text{and } y_i = +1 \\ \leq -1 + \xi_i & \text{and } y_i = -1 \end{array} \quad \text{for } i = 1, \ldots, n$$

- For $\xi_i > 1$, point i is misclassified if

$$\mathbf{w}^T \mathbf{x}_i + b \begin{array}{ll} \geq 1 - \xi_i & \text{and } y_i = +1 \\ \leq -1 + \xi_i & \text{and } y_i = -1 \end{array} \quad \text{for } i = 1, \ldots, n$$

.

Soft Margin

Minimizing the margin under the constraint with errors can be formulated as an optimization:

$$\min_{\mathbf{w},b,\xi} 2^{-1}\|\mathbf{w}\|^2 \text{ subject to } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \sum_{i=1}^{n} \xi_i \leq \text{constant}\,, \text{ for } i = 1, \ldots, n.$$

Or equivalently can be formulated as

$$\min_{\mathbf{w}\in\mathbb{R}^p, b} 2^{-1}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

subject to

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \text{ for } i = 1, \ldots, n,$$

where

- C is a regularization (cost) parameter (the value of C related to magnitude of margin):
  - small C to allow $\sum_{i=1}^{n} \xi_i$ larger to reach the minimum of $\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \to$ large margin.
  - large C to force $\sum_{i=1}^{n} \xi_i$ smaller to reach the minimum of $\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \to$ narrow margin.
  - $C = \infty$ to make $\sum_{i=1}^{n} \xi_i = 0 \to$ hard margin.
- This is still a quadratic optimization problem and there is a unique minimum.

Optimization for SVM: Hinge Loss

Learning an SVM has been formulated as a constrained optimization problem over $\mathbf{w}$, b, and ξ

$$\min_{\mathbf{w}\in\mathbb{R}^p, b, \xi} 2^{-1}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \text{ subject to } y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \xi \geq 0, \text{ for } i = 1, \ldots, n.$$

- The constraint $y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i$, can be written more concisely as

$$y_i f(\mathbf{x}_i) \geq 1 - \xi_i,$$

  or

$$\xi_i \geq 1 - y_i f(\mathbf{x}_i).$$

- together with $\xi_i \geq 0$, is equivalent to

$$\xi_i = \max\left(0, 1 - y_i f(\mathbf{x}_i)\right).$$

Hence the optimization problem is equivalent to the unconstrained optimization problem over $\mathbf{w}$ and b

$$\min_{\mathbf{w}\in\mathbb{R}^d, b} 2^{-1}\underbrace{\|\mathbf{w}\|^2}_{\text{regularization}} + C\sum_{i=1}^{n}\underbrace{\max\left(0, 1 - y_i f(\mathbf{x}_i)\right)}_{\text{loss function}}$$

or

$$\min_{\mathbf{w}\in\mathbb{R}^d, b} \sum_{i=1}^{n}\underbrace{\{1 - y_i f(\mathbf{x}_i)\}_+}_{\text{loss function}} + 2^{-1}\lambda\underbrace{\|\mathbf{w}\|^2}_{\text{regularization}},$$

where $\{1 - y_i f(\mathbf{x}_i)\}_+ = \max\left(0, 1 - y_i f(\mathbf{x}_i)\right)$ and $\lambda = 1/C$.

Compare with 0/1 Loss Function

Comparing with 0/1 loss function which allows errors: points on the wrong side of the decision boundary.

- $$\min_{w,b}\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\mathbb{I}\left\{y_i\left(w^\top x_i + b\right) < 0\right\}\right],$$

  where C controls the tradeoff between maximum margin and loss.
- Replace 0/1 loss function with other convex function $h(\cdot)$:

  $$\min_{w,b}\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}h\left\{y_i\left(w^\top x_i + b\right)\right\}\right].$$
- With hinge loss,

  $$h(s) = (1-s)_+ = \max(0, 1-s) = \begin{cases} 1-s, & 1-s > 0 \\ 0, & \text{otherwise.} \end{cases}$$
- Subgradient of hinge loss:

  $$\nabla h(s) = \begin{cases} -1, & 1-s > 0 \\ 0, & s > 1 \end{cases}$$

  where the hinge loss function at $s = 1$ is not differentiable but with subgradient $[-1, 0]$ at $s = 1$.
- Under the form of loss function + penalty, w and b can be estimated via stochastic gradient (subgradient) descent approach (Solving the primal problem, Shalev-Shwartz et al, 2007).

Loss functions

- SVM uses ”hinge” loss $\max(0, 1 - y_i f(\mathbf{x}_i))$.
- The 0-1 loss: $I\{yf(x) \leq 0\}$.
- Square loss: $\{y - f(x)\}^2 = \{1 - yf(x)\}^2$.
- logit loss: $\log[1 + \exp\{-yf(x)\}]$.

Optimization: the Primal and Dual Problem

- The Lagrange function is

$$
\begin{aligned}
L(\mathbf{w}, b, \xi, \alpha, \mu) &= 2^{-1}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left\{y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1 + \xi_i\right\} - \sum_{i=1}^{n}\mu_i\xi_i, \\
&= 2^{-1}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n}\xi_i\left(C - \alpha_i - \mu_i\right) + \sum_{i=1}^{n}\alpha_i\left\{1 - y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right)\right\},
\end{aligned}
$$

  where $\alpha_i \geq 0, \mu_i \geq 0, \xi_i \geq 0, i = 1\ldots, n$.
- Lagrange dual function is the inf over primal variables of L:

$$
\begin{aligned}
L_d(\alpha, \mu) &= \inf_{\mathbf{w},b,\xi} L(\mathbf{w}, b, \xi, \alpha, \mu) \\
&= \inf_{\mathbf{w},b,\xi}\left[2^{-1}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n}\xi_i\left(C - \alpha_i - \mu_i\right) + \sum_{i=1}^{n}\alpha_i\left\{1 - y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right)\right\}\right],
\end{aligned}
$$

- Setting the derivatives with respect to $\mathbf{w}$, b and ξ_i to zero, we have

$$
\begin{aligned}
\partial L(\mathbf{w}, b, \xi, \alpha, \mu)/\partial\mathbf{w} &= \mathbf{w} - \sum_{i=1}^{n}\alpha_i y_i x_i = 0, \\
\partial L(\mathbf{w}, b, \xi, \alpha, \mu)/\partial b &= -\sum_{i=1}^{n}\alpha_i y_i = 0 \\
\partial L(\mathbf{w}, b, \xi, \alpha, \mu)/\partial\xi &= C - \alpha_i - \mu_i = 0.
\end{aligned}
$$

Optimization: the Primal and Dual Problem, Cont.

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{n} \alpha_i y_i x_i, \\ 0 &= \sum_{i=1}^{n} \alpha_i y_i \\ \alpha_i &= C - \mu_i. \end{aligned}$$

- Substituting these results back into $L(\mathbf{w}, b, \xi, \alpha, \mu)$, we have,

$$\begin{aligned} \frac{1}{2} w^T w &= \frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i^T \sum_{j=1}^{n} \alpha_j y_j x_j = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j, \\ \sum_{i=1}^{n} \alpha_i \{1 - y_i(w^T x_i + b)\} &= \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{=0}. \end{aligned}$$

- Putting it together, the dual function is

$$L_d(\alpha, \mu) = \begin{cases} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j, & \sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \in [0, C], \text{ all } i \\ -\infty, & \text{otherwise.} \end{cases}$$

Optimization: the Dual Problem

The so-called dual problem is maximizing $L_d(\alpha, \mu)$ with the constraints:

- $$L_d(\alpha,\mu) \quad = \quad \sum_{i=1}^{n} \alpha_i - 2^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j x_i^T x_j, \text{ subject to } 0 \le \alpha_i \le C \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0.$$
- The solution is obtained by maximizing L_d subject to these constraints (QP problem).
- By complementary slackness, we have

$$\begin{aligned} \alpha_i \left\{ y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) - (1 - \xi_i) \right\} &= 0, \\ (C - \alpha_i)\xi_i &= 0. \end{aligned}$$

$\alpha_i = C > 0$:

- We immediately have $y_i \left(w^\top x_i + b \right) = 1 - \xi_i$
- Also, from condition $\alpha_i = C - \mu_i$, we have $\mu_i = 0$, so $\xi_i \ge 0$.
- Under this case $\alpha_i = C > 0$, the data point could either violate the constraint ($\xi_i > 0$) or locate on the margin ($\xi_i = 0$).

$0 < \alpha_i < C$:

- We again have $y_i \left(w^\top x_i + b \right) = 1 - \xi_i$
- This time, from $\alpha_i = C - \mu_i$, we have $\mu_i > 0$, hence $\xi_i = 0$.
- Under this case $0 < \alpha_i < C$, the data point locates on the margin.

Optimization: the Dual Problem, Cont.

$\alpha_i = 0$:

- From $\alpha_i = C - \mu_i$, we have $\mu_i > 0$, hence $\xi_i = 0$.
- Thus, $y_i \left(w^\top x_i + b\right) \geq 1$
- Under this case $\alpha_i = 0$, the data point could either on the right side of the margin or locate on the margin.

We observe that

- only those points on the decision boundary, or which are margin errors, contribute the support vectors ($\alpha_i \neq 0$) to estimate w;
- only for those satisfying $0 < \alpha_i < C$ can be used to estimate b since $b = y_i - w^\top x_i$, $0 < \alpha_i < C$.

SVM: Nonlinear classifiers

Suppose the basis function $\Phi(\mathbf{x}) = (\Phi(\mathbf{x})_1, \ldots, \Phi(\mathbf{x})_D)$. We extend the original SVM classifier by using the basis function $\Phi(\mathbf{x})$ as

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b$$

where Φ is a function that transforms $\mathbf{x}$ from $\mathbb{R}^d$ to $\mathbb{R}^D$ and considering $D \gg d$.

SVM: Nonlinear classifiers as penalized method

Classifier, with $\mathbf{w} \in \mathbb{R}^D$:

$$f(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + b.$$

Objective function for $\mathbf{w} \in \mathbb{R}^D$

$$\min_{\mathbf{w} \in \mathbb{R}^D} 2^{-1} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \max\{0, 1 - y_i f(\mathbf{x}_i)\}$$

or

$$\min_{\mathbf{w} \in \mathbb{R}^D} \sum_{i=1}^{n} \{1 - y_i f(\mathbf{x}_i)\}_+ + 2^{-1} \lambda \|\mathbf{w}\|^2.$$

- Simply map $\mathbf{x}$ to $\Phi(\mathbf{x})$ to expect that data are separable under the transformed basis function
- Solve for $\mathbf{w}$ in high dimensional space $\mathbb{R}^D$.

Nonlinear classifiers in Dual problem

- Primal problem to find $\mathbf{w}$: $f(\mathbf{x}) = \mathbf{w}^T\Phi(\mathbf{x}) + b$.
- Dual problem to find α: $f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}) + b$, where $\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \Phi(x_i)$.

Thus, in dual problem, the objective function becomes as

$$\max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{j=1}^n\sum_{k=1}^n \alpha_j\alpha_k y_j y_k \mathbf{x}_j^T\mathbf{x}_k$$

$$\to \max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{j=1}^n\sum_{k=1}^n \alpha_j\alpha_k y_j y_k \Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_k)$$

subject to

$$0 \leq \alpha_i \leq C \; \forall i \text{ and } \sum_{i=1}^n \alpha_i y_i = 0.$$

Note that

- only n dimensional vector $\boldsymbol{\alpha}$ needs to be estimated in the dual problem; it is not necessary to learn in the D dimensional space, as it is for the primal problem.
- Write kernel function $k(\mathbf{x}_j, \mathbf{x}_i) = \Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i)$.
- Linear kernels $k(\mathbf{s}, \mathbf{z}) = \mathbf{s}^T\mathbf{z}$
- Polynomial kernels $k(\mathbf{s}, \mathbf{z}) = (1 + \mathbf{s}^T\mathbf{z})^d$ for any $d > 0$
  - Contains all polynomials terms up to degree d
- Gaussian kernels $k(\mathbf{s}, \mathbf{z}) = \exp(-\|\mathbf{s} - \mathbf{z}\|^2/2\sigma^2)$ for $\sigma > 0$
  - Infinite dimensional feature space.

Kernel SVM

- Consider the dual soft SVM with explicit non-linear transformation $x \mapsto \Phi(x)$:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_i \Phi\left(x_i\right)^\top \Phi\left(x_j\right) \quad \text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \preceq \alpha \preceq C \end{cases}$$

- Introduce quadratic non-linearities, $s = (s_1, s_2)$, and $z = (z_1, z_2)$, we have

$$\Phi(s) = \left(1, \sqrt{2}s_1, \sqrt{2}s_2, \sqrt{2}s_1 s_2, s_1^2, s_2^2\right)^\top .$$

  Then

$$\begin{aligned} \Phi\left(s\right)^\top \Phi\left(z\right) &= 1 + 2s_1 z_1 + 2s_2 z_2 + 2s_1 s_2 z_1 z_2 \\ &\quad + \left(s_1\right)^2 \left(z_1\right)^2 + \left(s_2\right)^2 \left(z_2\right)^2 = \left(1 + s^\top z\right)^2 \end{aligned}$$

- Since only inner products are needed, non-linear transform need not be computed explicitly - inner product between features can be a simple function (**kernel**) of s and z : $k\left(s, z\right) = \Phi\left(s\right)^\top \Phi\left(z\right) = \left(1 + s^\top z\right)^2$
- d -order interactions can be implemented by $k\left(s, z\right) = \left(1 + s^\top z\right)^d$ (**polynomial kernel**).

Kernel SVM: Kernel trick

- Kernel SVM with $k(x_i, x_j)$. Non-linear transformation $x \mapsto \Phi(x)$ still present, but implicit (coordinates of the vector $\Phi(x)$ are never computed).

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \quad \text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \preceq \alpha \preceq C \end{cases}$$

- Prediction? $\hat{y}(x) = \text{sign}\left(w^\top \Phi(x) + b\right)$, where $w = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i)$ and b obtained from a margin support vector x_j with $\alpha_j \in (0, C)$.
  - No need to compute w either! Just need

$$w^\top \Phi(x) = \sum_{i=1}^{n} \alpha_i y_i \Phi(x_i)^\top \Phi(x) = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x).$$

  - Get b from

$$b = y_j - w^\top \Phi(x_j) = y_j - \sum_{i=1}^{n} \alpha_i y_i k(x_i, x_j)$$

  for any margin support-vector x_j $(\alpha_j \in (0, C))$.

- Fitted a separating hyperplane in a high-dimensional feature space without ever mapping explicitly to that space.