EECS 545: Machine Learning

Lecture 10. Kernel methods: Kernelizing Support Vector Machines

Honglak Lee

02/14/2024

UNIVERSITY OF MICHIGAN

Overview

- Support Vector Machine (SVM)
- Dual optimization
  - General recipe for constrained optimization
  - Hard-margin SVM
  - Soft-margin SVM

Maximum Margin Classifier

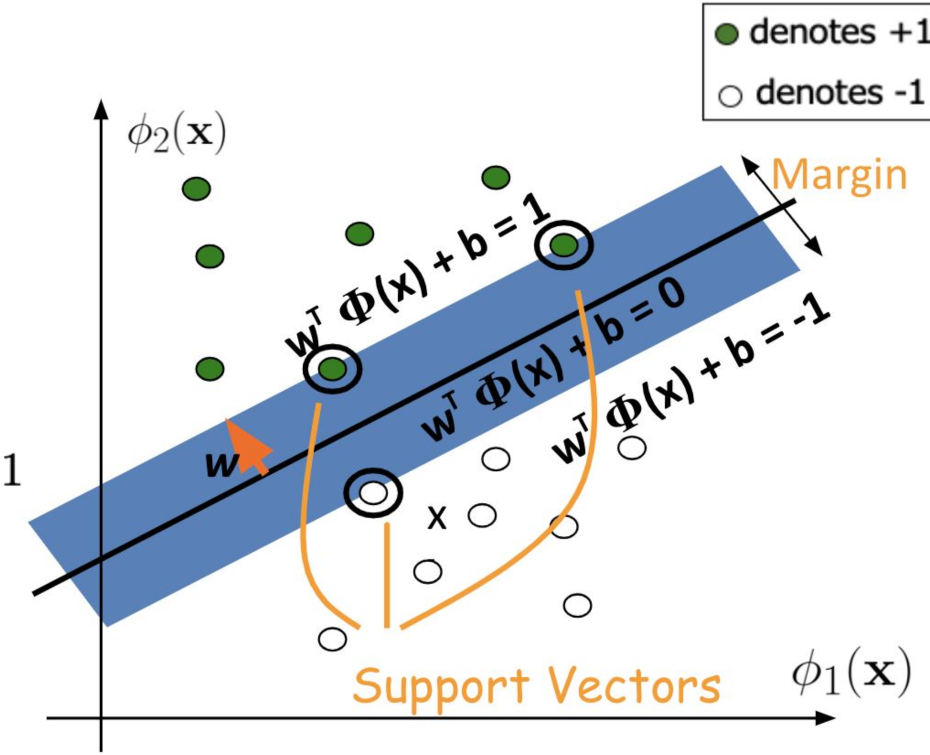
- Optimization problem:

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

subject to

$$\text{For } \ y^{(n)} = 1, \qquad \mathbf{w}^\top \phi\left(\mathbf{x}^{(n)}\right) + b \geq 1$$

$$\text{For } \ y^{(n)} = -1, \quad \mathbf{w}^\top \phi\left(\mathbf{x}^{(n)}\right) + b \leq -1$$



Dual optimization

- So far, we have considered primal optimization which requires a direct access to the feature vectors $\phi\left(\mathbf{x}^{(n)}\right)$
- It is also possible to "kernelize" SVM
  - This formulation is called "Dual" formulation.
  - In this case, you can use any kernel function (such as polynomial, RBF, etc.)

With dual variables $\alpha^{(n)}$,
we have the following relations
(without proofs)

$$\mathbf{w} = \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} \phi\left(\mathbf{x}^{(n)}\right)$$

$$h(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b = \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} k\left(\mathbf{x}, \mathbf{x}^{(n)}\right) + b$$

Kernelizing SVM: back to hard-margin case

- Optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to} \quad y^{(n)}\left(\mathbf{w}^\top \phi\left(\mathbf{x}^{(n)}\right) + b\right) \geq 1, n = 1, ..., N$$

- This is a constrained optimization problem.
  - We solve this using Lagrange multipliers (convex optimization)
  - Solving dual optimization problem naturally leads to kernalization

Solving Constrained Optimization: General Overview and Recipe

(This section is just a recap, see the supplementary lecture slides for more details)

General (Constrained) Optimization

- General optimization problem:

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & f(\mathbf{x}) \\
\text{subject to} \quad & g_i(\mathbf{x}) \leq 0, i = 1, ..., m \\
& h_i(\mathbf{x}) = 0, i = 1, ..., p
\end{aligned}
$$

objective (cost) function

inequality constraint functions

equality constraint functions

- If **x** satisfies all the constraints, **x** is called <u>feasible</u> (a feasible solution).
- In general, this is a nontrivial problem to solve, so we use techniques for convex optimization.

Recap: General Recipe

- Given an original optimization

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0,\ i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ i = 1, ..., p \end{aligned}$$

- Solve dual optimization with Lagrangian function:

$$\begin{aligned} \max_{\lambda,\nu} \min_{\mathbf{x}} \quad & \mathcal{L}(\mathbf{x}, \lambda, \nu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) \\ \text{subject to} \quad & \lambda_i \geq 0,\ \forall i \end{aligned}$$

Add constraint terms with Lagrange multipliers

- Alternatively, solve the dual optimization with Lagrange dual:

$$\begin{aligned} \max_{\lambda,\nu} \quad & \tilde{\mathcal{L}}(\lambda, \nu) \\ \text{subject to} \quad & \lambda_i \geq 0,\ \forall i \end{aligned} \qquad \text{where } \tilde{\mathcal{L}}(\lambda, \nu) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \nu)$$

A Big Picture

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0,\ i = 1, \ldots, p \end{aligned}$$

Constrained Optimization Problem

$$\mathcal{L}(\mathbf{x}, \lambda, \nu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

Lagrangian

e.g. convex optimizations, KKT conditions

strong duality (if some conditions are met)

$$p^* = d^*$$

Primal Optimization Problem (min-max)

$$\min_{\mathbf{x}} \max_{\nu, \lambda : \lambda_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \lambda, \nu)$$

weak duality

$$p^* \geq d^*$$

Dual Optimization Problem (max-min)

$$\max_{\nu, \lambda : \lambda_i \geq 0, \forall i} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \nu)$$

Lagrangian Formulation

$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$
$$\text{subject to} \quad g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m$$
$$h_i(\mathbf{x}) = 0, i = 1, \ldots, p$$

- The **Lagrangian function** is

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

  - Here, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]$ $(\lambda_i \geq 0, \forall\, i)$ and $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_p]$ are called Lagrange multipliers (or dual variables)

- This leads to **primal optimization problem**

$$\min_{\mathbf{x}} \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}:\, \lambda_i \geq 0\, \forall i} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

  - Difficult to solve directly!

Primal and Feasibility

- Primal optimization problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0,\ i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ i = 1, ..., p \end{aligned}$$

$$p^* = \min_{\mathbf{x}} \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}:\, \lambda_i \geq 0\, \forall i} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

where $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$

- Notice that:

$$\mathcal{L}_p(\mathbf{x}) = \max_{\nu, \lambda: \lambda_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \lambda, \nu) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{x} \text{ is feasible} \\ \infty & \text{otherwise} \end{cases}$$

This eliminates the constraints on $\mathbf{x}$, yielding an equivalent optimization problem.

Lagrange Dual

primal vs dual: switching the order of min / max

Note: these are different problems!

- Dual optimization problem:

$$d^* = \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0, \forall i} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

cf) primal optimization problem

$$p^* = \min_{\mathbf{x}} \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

- We can also write as:

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} \min_{\mathbf{x}} \quad & \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} \quad & \lambda_i \geq 0, \forall i \end{aligned}$$

or

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} \quad & \tilde{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} \quad & \lambda_i \geq 0, \forall i \\ \text{where} \quad & \tilde{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \end{aligned}$$

Lagrange Dual function

Weak Duality

strong duality (if some conditions are met)

$p^* = d^*$

Primal Optimization Problem (min-max)

$$\min_{\mathbf{x}} \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

weak duality

$p^* \geq d^*$

Dual Optimization Problem (max-min)

$$\max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0, \forall i} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

- Claim:

$$\begin{aligned} d^* &= \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &\leq \min_{\mathbf{x}} \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= p^* \end{aligned}$$

- Difference between p^* and d^* is called the **<u>duality gap</u>**.
- In other words, the dual maximization problem (usually easier) gives a "**lower bound**" for the primal minimization problem (usually more difficult).

Weak Duality

Also see Convex Optimization Review Session

$$d^* = \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \min_{\mathbf{x}} \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = p^*$$

- Proof: Let $\tilde{\mathbf{x}}$ be feasible. Then for any $\boldsymbol{\lambda}, \boldsymbol{\nu}$ with $\lambda_i \geq 0$,

$$\mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \lambda_i g_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{\mathbf{x}}) \leq f(\tilde{\mathbf{x}})$$

Thus, $\tilde{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\tilde{\mathbf{x}})$
for any $\boldsymbol{\lambda}, \boldsymbol{\nu}$ with $\lambda_i \geq 0$, any feasible $\tilde{\mathbf{x}}$

Then, maximize LHS (w.r.t. dual variables)

$$d^* = \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0} \tilde{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f(\tilde{\mathbf{x}}) \text{ for any feasible } \tilde{\mathbf{x}}$$

Finally, minimize RHS (w.r.t. primal variable)

$$d^* = \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0} \tilde{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \min_{\tilde{\mathbf{x}}:\text{feasible}} f(\tilde{\mathbf{x}}) = p^*$$

Strong Duality

- If $p^* = d^*$, we say <u>strong duality</u> holds.
- What are the conditions for strong duality?
  - does not hold in general
  - holds for convex problems (under mild conditions)
  - conditions that guarantee strong duality in convex problems are called constraint qualification.
- Two well-known conditions (in convex problems)
  - Slater's constraint qualification (review session)
  - Karush-Kuhn-Tucker (KKT) condition (main focus)

Convex Optimization

Also see Convex Optimization Review Session

- Standard form of **convex problem** has the form:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, i = 1, ..., m \\ & h_i(\mathbf{x}) = 0, i = 1, ..., p \end{aligned}$$

(where *f*, g_i are convex, *and* h_i are affine)

- If **x** satisfies all the constraints, **x** is called feasible.
  - In general, this is a nontrivial problem to solve, so we use techniques for convex optimization.

(Sufficient) Conditions for strong duality: **Slater's constraint qualification**

- Strong duality holds for a **convex** problem

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & f(\mathbf{x}) \\
\text{subject to} \quad & g_i(\mathbf{x}) \leq 0, i = 1, ..., m \\
& h_i(\mathbf{x}) = 0, i = 1, ..., p
\end{aligned}
$$

(where f, g_i are **convex**, *and* h_i are **affine**)

if the constraint is <u>strictly</u> feasible (by any solution), i.e.,

$$
\begin{aligned}
\exists \mathbf{x} : \quad & g_i(\mathbf{x}) < 0, \forall i = 1, ..., m \\
& h_i(\mathbf{x}) = 0, \forall i = 1, ..., p
\end{aligned}
$$

(Not necessarily an optimal solution)

Slater's condition is a **sufficient** condition for strong duality to hold for a convex problem

Karush-Kuhn-Tucker (KKT) condition

Let $\mathbf{x}^*$ be a primal optimal and $\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ be a dual optimal solution. If the strong duality holds, then we have the following:

$$\nabla_{\mathbf{x}} f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_{\mathbf{x}} g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \nu_i^* \nabla_{\mathbf{x}} h_i(\mathbf{x}^*) = 0,$$ Stationarity (1)

$$g_i(\mathbf{x}^*) \le 0, \quad i = 1, \dots, m,$$ Primal feasibility (2)

$$h_i(\mathbf{x}^*) = 0, \quad i = 1, \dots, p,$$ Primal feasibility (3)

$$\lambda_i^* \ge 0, \quad i = 1, \dots, m,$$ Dual feasibility (4)

$$\boxed{\lambda_i^* g_i(\mathbf{x}^*) = 0, \quad i = 1, \dots, m}$$ Complementary slackness (5)

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \le 0,\ i = 1, ..., m \\ & h_i(\mathbf{x}) = 0,\ i = 1, ..., p \end{aligned}$$

$$\begin{aligned} \max_{\lambda, \nu} \min_{\mathbf{x}} \quad & \mathcal{L}(\mathbf{x}, \lambda, \nu) \\ \text{subject to} \quad & \lambda_i \ge 0,\ \forall i \end{aligned}$$

Dual problem

Note: we do **not** assume the optimization problem is necessarily convex for describing KKT condition. However, when the problem is convex (and differentiable), KKT condition ensures strong duality.

(Sufficient) Conditions for strong duality: KKT Conditions

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0, i = 1, \ldots, p \end{aligned}$$

- Assume f, g_i, h_i are differentiable
- If the original problem is **convex** (where f, g_i are **convex** and h_i are affine), and $\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ satisfy the KKT conditions, then:
    - $\mathbf{x}^*$ is primal optimal
    - $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is dual optimal, and
    - the duality gap is zero (i.e., strong duality holds)

For convex optimization problems (+ differentiable objectives/constraints)**,**
KKT is a sufficient condition for strong duality.

Proof for sufficiency (KKT => Strong duality)

Claim: When KKT (1)-(5) holds, the strong duality holds.

- From (2) and (3), $\mathbf{x}^*$ is primal feasible.
- From (4), $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ is dual feasible.
- $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is a convex differentiable function.
  Thus, from (1), $\mathbf{x}^*$ is a minimizer of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$.
- Then,

$$\begin{aligned} \tilde{\mathcal{L}}(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) &= \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\ &= \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\ &= f(\mathbf{x}^*) + \underbrace{\sum_i \lambda_i^* g_i(\mathbf{x}^*)}_{=0} + \sum_i \nu_i^* h_i(\mathbf{x}^*) \\ &= f(\mathbf{x}^*) \end{aligned}$$

(See also: derivation of complementary slackness)

$= 0$ $\because$ (5) complementary slackness

- But, $\tilde{\mathcal{L}}(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \le \underbrace{\max_{\boldsymbol{\lambda}, \boldsymbol{\nu}: \lambda_i \ge 0} \tilde{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \le \min_{\mathbf{x}:\mathbf{x} \text{ is feasible}} f(\mathbf{x})}_{\text{weak duality}} \le f(\mathbf{x}^*) = \tilde{\mathcal{L}}(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$

- Then, $\max_{\boldsymbol{\lambda}, \boldsymbol{\nu}: \lambda_i \ge 0} \tilde{\mathcal{L}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}:\mathbf{x} \text{ is feasible}} f(\mathbf{x})$

which proves that the strong duality holds (i.e., duality gap is zero).

KKT conditions: Conclusion

- If a constrained optimization if **differentiable** and has **convex** objective function and constraint sets, then the KKT conditions are **(necessary and) sufficient conditions** for **strong duality** (zero duality gap).

- Thus, the KKT conditions can be used to solve such problems.

Applying Constrained Optimization Techniques for solving SVM

Kernelizing SVM: back to hard-margin case

- Optimization problem:

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

label is either -1 or +1

$$\text{subject to} \quad y^{(n)}\left(\mathbf{w}^\top \phi\left(\mathbf{x}^{(n)}\right) + b\right) \geq 1, n = 1, ..., N$$

- This is a constrained optimization problem.
  - We solve this using Lagrange multipliers (convex optimization)

Back to hard-margin SVM

- Use Lagrange multipliers to enforce constraints while optimizing

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{n=1}^{N} \alpha^{(n)} \left\{1 - y^{(n)} \left(\mathbf{w}^\top \phi\left(\mathbf{x}^{(n)}\right) + b\right)\right\}$$

- Here, $\alpha^{(n)} \geq 0$ is the Lagrange multiplier (or dual variable) for each constraint (one per data point)

$$y^{(n)} \left(\mathbf{w}^\top \phi\left(\mathbf{x}^{(n)}\right) + b\right) \geq 1 \qquad n = 1, \ldots, N$$

Lagrangian and Lagrange Dual

- Optimizing the Lagrange dual problem :

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{n=1}^{N} \alpha^{(n)} \left\{1 - y^{(n)} \left(\mathbf{w}^\top \phi\left(\mathbf{x}^{(n)}\right) + b\right)\right\}$$

$$\text{subject to} \quad \alpha^{(n)} \geq 0, \forall n$$

- We first minimize w.r.t. primal variables **w** and b, and get a Lagrange dual problem:

$$\max_{\boldsymbol{\alpha}} \; \tilde{\mathcal{L}}(\boldsymbol{\alpha})$$

$$\text{subject to} \quad \alpha^{(n)} \geq 0, \forall n$$

$$\text{where} \quad \tilde{\mathcal{L}}(\boldsymbol{\alpha}) = \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$$

(a.k.a. Lagrange dual function)

(Please see the supplementary material for more explanation about Lagrange Dual)

Maximize the Margin

- Lagrangian function:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{n=1}^{N} \alpha^{(n)} \left\{1 - y^{(n)} \left(\mathbf{w}^\top \phi\left(\mathbf{x}^{(n)}\right) + b\right)\right\}$$

- Set the derivatives of $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha})$ to zero, to get

$$\mathbf{w} = \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} \phi\left(\mathbf{x}^{(n)}\right) \qquad 0 = \sum_{n=1}^{N} \alpha^{(n)} y^{(n)}$$

c.f. KKT (1) Stationarity
$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = 0$
$\nabla_{b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = 0$

- Substitute in, to eliminate **w** and *b*,

$$\max_{\boldsymbol{\alpha}} \tilde{\mathcal{L}}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha^{(n)} - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha^{(n)} \alpha^{(m)} y^{(n)} y^{(m)} \phi\left(\mathbf{x}^{(n)}\right)^\top \phi\left(\mathbf{x}^{(m)}\right)$$

$$\text{subject to} \quad \alpha^{(n)} \geq 0, \quad \forall n$$

Dual Representation (with kernel)

- Define a kernel $k\left(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}\right) = \phi\left(\mathbf{x}^{(n)}\right)^\top \phi\left(\mathbf{x}^{(m)}\right)$
- Dual optimization is to maximize

$$\max_{\boldsymbol{\alpha}} \tilde{\mathcal{L}}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha^{(n)} - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha^{(n)} \alpha^{(m)} y^{(n)} y^{(m)} \underbrace{\phi\left(\mathbf{x}^{(n)}\right)^\top \phi\left(\mathbf{x}^{(m)}\right)}_{=k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)})}$$

$$\text{subject to} \quad \alpha^{(n)} \geq 0, \quad \forall n$$

- Once we have $\boldsymbol{\alpha}$, we don't need **w**.
- Predict classification for arbitrary input **x** using:

$$h(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b = \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} k\left(\mathbf{x}, \mathbf{x}^{(n)}\right) + b$$

$$\mathbf{w} = \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} \phi\left(\mathbf{x}^{(n)}\right)$$

Support Vectors

- The KKT conditions are:

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = 0$$
$$\nabla_{b}\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = 0$$
$$\alpha^{(n)} \geq 0$$
$$1 - y^{(n)} h\left(\mathbf{x}^{(n)}\right) \leq 0$$
$$\alpha^{(n)} \left\{1 - y^{(n)} h\left(\mathbf{x}^{(n)}\right)\right\} = 0$$

- The last condition (complementary slackness) means:
  - either $\alpha^{(n)} = 0$ or $y^{(n)} h\left(\mathbf{x}^{(n)}\right) = 1$.

support vectors

- That is, only the support vectors matter!
  - To compute $h(\mathbf{x})$ (prediction), sum only over support vectors $h(\mathbf{x}) = \sum_{m\text{:support vectors}} \alpha^{(m)} y^{(m)} k\left(\mathbf{x}, \mathbf{x}^{(m)}\right) + b$

Recovering b

- For any support vector $\mathbf{x}^{(n)} : y^{(n)} h\left(\mathbf{x}^{(n)}\right) = 1$

- Replacing with $h(\mathbf{x}) = \sum_{m \in S} \alpha^{(m)} y^{(m)} k\left(\mathbf{x}, \mathbf{x}^{(m)}\right) + b$

$$y^{(n)} \left(\sum_{m \in S} \alpha^{(m)} y^{(m)} k\left(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}\right) + b \right) = 1$$

(index) set of support vectors

- Multiply $y^{(n)}$, and sum over n:

$$b = \frac{1}{N_S} \sum_{n \in S} \left(y^{(n)} - \sum_{m \in S} \alpha^{(m)} y^{(m)} k\left(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}\right) \right)$$

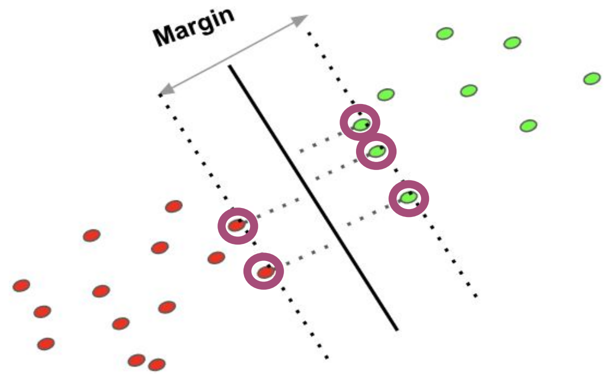


Image adapted from: https://www.vebuso.com/2020/02/a-top-machine-learning-algorithm-explained-support-vector-machines-svms/

Formulation of soft-margin SVM

- Maximize the margin, and also penalize for the slack variables

$$C \sum_{n=1}^{N} \xi^{(n)} + \frac{1}{2} \|\mathbf{w}\|^2$$

- The support vectors are now those with

$$y^{(n)} h\left(\mathbf{x}^{(n)}\right) = 1 - \xi^{(n)}$$

Dual formulation of soft-margin SVM

- Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\| + C\sum_{n=1}^{N}\xi^{(n)} + \sum_{n=1}^{N}\alpha^{(n)}\left\{1 - y^{(n)}h(\mathbf{x}^{(n)}) - \xi^{(n)}\right\} + \sum_{n=1}^{N}\mu^{(n)}\left(-\xi^{(n)}\right)$$

$$\text{where} \quad \alpha^{(n)} \geq 0, \quad \mu^{(n)} \geq 0, \quad \xi^{(n)} \geq 0, \quad \forall n$$

Dual formulation of soft-margin SVM

- Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\|\mathbf{w}\| + C\sum_{n=1}^{N}\xi^{(n)} + \sum_{n=1}^{N}\alpha^{(n)}\left\{1 - y^{(n)}h(\mathbf{x}^{(n)}) - \xi^{(n)}\right\} + \sum_{n=1}^{N}\mu^{(n)}\left(-\xi^{(n)}\right)$$

where $\alpha^{(n)} \geq 0, \quad \mu^{(n)} \geq 0, \quad \xi^{(n)} \geq 0, \quad \forall n$

- KKT conditions for the constraints

$$\left.\begin{aligned} 1 - y^{(n)}h\left(\mathbf{x}^{(n)}\right) - \xi^{(n)} &\leq 0 \\ -\xi^{(n)} &\leq 0 \end{aligned}\right\} \text{Primal variables satisfy the inequality constraints}$$

$$\left.\begin{aligned} \alpha^{(n)} &\geq 0 \\ \mu^{(n)} &\geq 0 \end{aligned}\right\} \text{Dual variables (for above inequalities) are feasible}$$

$$\left.\begin{aligned} \alpha^{(n)}\left(1 - y^{(n)}h\left(\mathbf{x}^{(n)}\right) - \xi^{(n)}\right) &= 0 \\ \mu^{(n)}\xi^{(n)} &= 0 \end{aligned}\right\} \text{Complementary slackness condition}$$

Dual formulation of soft-margin SVM

- Taking derivatives

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} \phi\left(\mathbf{x}^{(n)}\right)$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \quad \Rightarrow \quad \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \xi^{(n)}} = 0 \quad \Rightarrow \quad \alpha^{(n)} = C - \mu^{(n)}$$

Dual formulation of soft-margin SVM

$$\mathbf{w} = \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} \phi\left(\mathbf{x}^{(n)}\right) \qquad \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} = 0 \qquad \alpha^{(n)} = C - \mu^{(n)}$$

- Plug these back into the Lagrangian:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) &= \frac{1}{2}\mathbf{w}^\top\mathbf{w} + \sum_{n=1}^{N} \underbrace{\left(C - \mu^{(n)}\right)}_{\alpha^{(n)}} \xi^{(n)} + \sum_{n=1}^{N} \alpha^{(n)} \{1 - y^{(n)}(\mathbf{w}^\top \phi(\mathbf{x}^{(n)}) + b)) - \xi^{(n)}\} \\
&= \frac{1}{2}\mathbf{w}^\top\mathbf{w} - \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} \mathbf{w}^\top \phi(\mathbf{x}^{(n)}) - b \underbrace{\sum_{n=1}^{N} \alpha^{(n)} y^{(n)}}_{0} + \sum_{n=1}^{N} \alpha^{(n)} \\
&= \frac{1}{2}\mathbf{w}^\top\mathbf{w} - \mathbf{w}^\top \underbrace{\left(\sum_{n=1}^{N} \alpha^{(n)} y^{(n)} \phi(\mathbf{x}^{(n)})\right)}_{\mathbf{w}} + \sum_{n=1}^{N} \alpha^{(n)} \\
&= \sum_{n=1}^{N} \alpha^{(n)} - \frac{1}{2}\mathbf{w}^\top\mathbf{w} \\
&= \sum_{n=1}^{N} \alpha^{(n)} - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha^{(n)} \alpha^{(m)} y^{(n)} y^{(m)} \phi(\mathbf{x}^{(n)})^\top \phi(\mathbf{x}^{(m)})
\end{aligned}$$

Dual formulation of soft-margin SVM

- Dual optimization (via Lagrange dual)

$$\max_{\boldsymbol{\alpha}} \quad \sum_{n=1}^{N} \alpha^{(n)} - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \alpha^{(n)} \alpha^{(m)} y^{(n)} y^{(m)} k\left(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}\right)$$

Inner product of features replaced with kernel

$$\text{subject to} \quad 0 \leq \alpha^{(n)} \leq C \longleftarrow \mu^{(n)} = C - \alpha^{(n)} \geq 0$$

$$\sum_{n=1}^{N} \alpha^{(n)} y^{(n)} = 0$$

- Solve quadratic problem (convex optimization)

SVM: practical issues

Support Vector Machine: Algorithm

1. Choose a kernel function

2. Choose a value for *C*
   (i.e., smaller C $\rightarrow$ larger regularization)

3. Solve the optimization problem (many software packages available) – primal or dual

4. Construct the discriminant function from the support vectors

Some Issues

- Linear kernels work fairly well, but can be suboptimal.
- Choice of (nonlinear) kernels
  – Gaussian or polynomial kernel is default
  – If the simple kernels are ineffective, more elaborate kernels are needed
  – Domain experts can give assistance in formulating appropriate similarity measures
- Choice of kernel parameters
  – E.g., Gaussian kernel: $K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}\right)$
    - σ is the distance between neighboring points whose labels will likely to affect the prediction of the query point.

  – In the absence of reliable criteria, applications rely on the use of a validation set or cross-validation to set such parameters.

Summary: Support Vector Machine

- Max margin classifier: improved robustness & less over-fitting
- Solved by convex optimization techniques
- Kernel trick can learn complex decision boundaries

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0, i = 1, \ldots, p \end{aligned}$$

Primal Optimization Problem

strong duality

(e.g., KKT condition)

$$p^* = d^*$$

weak duality

$$p^* \geq d^*$$

$$\max_{\nu, \lambda : \lambda_i \geq 0, \forall i} \; \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \nu)$$

Dual Optimization Problem (max-min)

$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad & C \sum_{n=1}^{N} \xi^{(n)} + \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subject to} \quad & y^{(n)} h\left(\mathbf{x}^{(n)}\right) \geq 1 - \xi^{(n)}, \forall n \\ & \xi^{(n)} \geq 0, \forall n \end{aligned}$$

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{n=1}^{N} \alpha^{(n)} - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \alpha^{(n)} \alpha^{(m)} y^{(n)} y^{(m)} k\left(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}\right) \\ \text{subject to} \quad & 0 \leq \alpha^{(n)} \leq C \\ & \sum_{n=1}^{N} \alpha^{(n)} y^{(n)} = 0 \end{aligned}$$

Additional Resource

- Kernel Methods
  - http://www.kernel-machines.org/

- Convex Optimization
  - http://www.stanford.edu/~boyd/cvxbook/
  - http://www.stanford.edu/class/ee364a/
  - see Chapter 5 (and earlier chapters)

SVM Implementation

- LIBSVM
  - http://www.csie.ntu.edu.tw/~cjlin/libsvm/
  - One of the most popular generic SVM solver (supports nonlinear kernels)
- Liblinear
  - http://www.csie.ntu.edu.tw/~cjlin/liblinear/
  - One of the fastest <u>linear</u> SVM solver (linear kernel)
- SVMlight
  - http://www.cs.cornell.edu/people/tj/svm_light/
  - Structured outputs, various objective measure (e.g., F1, ROC area), Ranking, etc.
- Scikit-learn
  - https://scikit-learn.org/stable/modules/svm.html

SVM demo code

- http://www.mathworks.com/matlabcentral/fileexchange/28302-svm-demo

- http://www.alivelearn.net/?p=912

Any feedback (about lecture, slide, homework, project, etc.)?

(via **anonymous** google form: https://forms.gle/99jeftYTaozJvCEF8)

Change Log of lecture slides:
https://docs.google.com/document/d/e/2PACX-1vRKx40eOJKACqrKWraio0AmlFS1_xBMINuWcc-jzpfo-ySj_gBuqTVdfHy8v4HDmqDJ3b3TvAW1FVuH/pub