Support vector machines

CS 446 / ECE 449

2022-02-16 19:35:44 -0600 (0670b06)

Another algorithm for linear prediction. **Why?!**

“`pytorch` meta-algorithm”.

$\vdots$

5. Tweak 1-4 until training error is small.
6. Tweak 1-5, possibly reducing model complexity, until testing error is small.

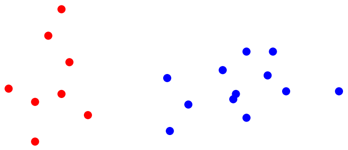
Support vector machines (SVMs) have three purposes for us.

1. Demonstrate **maximum margin predictors**, an example of “low complexity models”, which appear throughout machine learning (not just linear predictors).
2. Demonstrate **nonlinear kernels**, also pervasive.
3. Exercise convex optimization and duality.

Plan for SVM

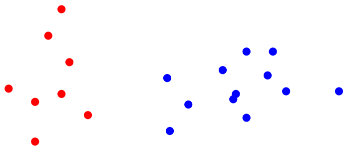
- Hard-margin SVM.
- Soft-margin SVM.
- SVM duality.
- Nonlinear SVM: kernels

Maximum margin linear separators

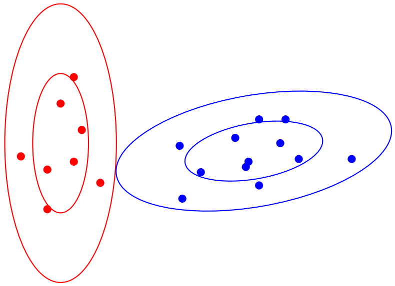


Which linear separator is best?

Maximum margin linear separators

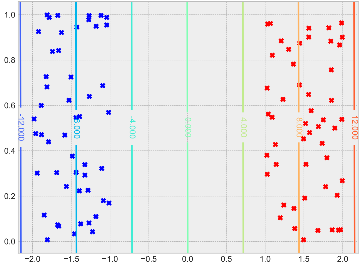


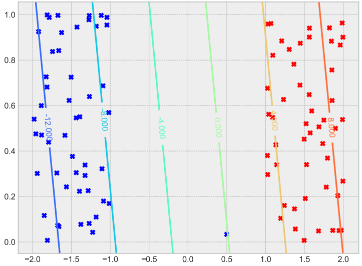
Which linear separator is best?



Which linear separator is best?

The max margin separator is **one** choice for a good predictor.
It is not always the best idea. Recall from lecture 3:

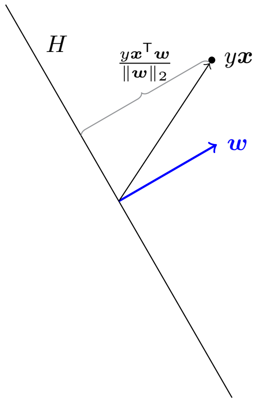




Even so, the maximum margin concept is pervasive in machine learning.

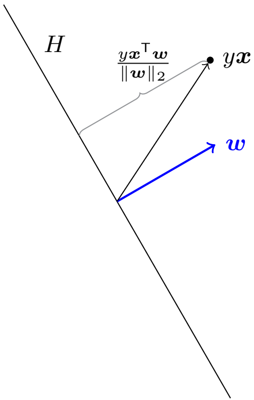
How to write "maximum margin classifier"?

Input $\boldsymbol{x} \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $\boldsymbol{w} \in \mathbb{R}^d$ with $H := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{w}^\mathsf{T}\boldsymbol{z} = 0\}$.



How to write "maximum margin classifier"?

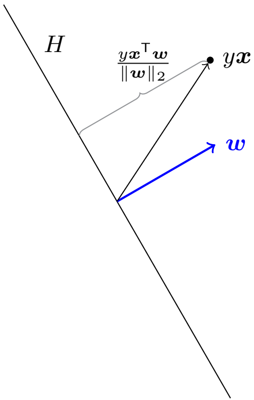
Input $\boldsymbol{x} \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $\boldsymbol{w} \in \mathbb{R}^d$ with $H := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{w}^\mathsf{T}\boldsymbol{z} = 0\}$.



Formulation #1: constrain the norm.

How to write "maximum margin classifier"?

Input $\boldsymbol{x} \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $\boldsymbol{w} \in \mathbb{R}^d$ with $H := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{w}^\top \boldsymbol{z} = 0\}$.



Formulation #1: constrain the norm.

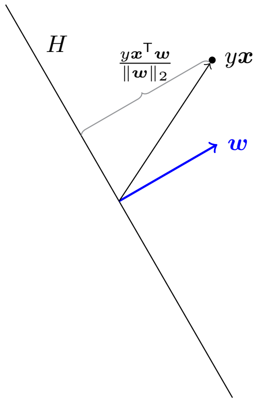
Single margin $\frac{y\boldsymbol{x}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}$.

Overall margin $\min_i \frac{y_i \boldsymbol{x}_i^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}$.

Max margin $\max_{\|\boldsymbol{u}\|=1} \min_i y_i \boldsymbol{x}_i^\top \boldsymbol{u}$.

How to write "maximum margin classifier"?

Input $\boldsymbol{x} \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $\boldsymbol{w} \in \mathbb{R}^d$ with $H := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{w}^\top \boldsymbol{z} = 0\}$.



Formulation #1: constrain the norm.

Single margin $\frac{y\boldsymbol{x}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}$.

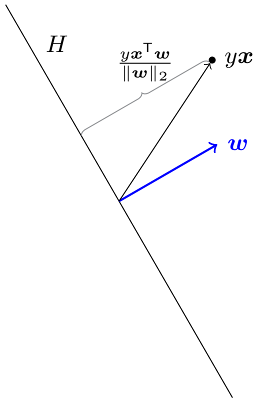
Overall margin $\min_i \frac{y_i \boldsymbol{x}_i^\top \boldsymbol{w}}{\|\boldsymbol{w}\|}$.

Max margin $\max_{\|\boldsymbol{u}\|=1} \min_i y_i \boldsymbol{x}_i^\top \boldsymbol{u}$.

Formulation #2: constrain the margins.

How to write "maximum margin classifier"?

Input $\boldsymbol{x} \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $\boldsymbol{w} \in \mathbb{R}^d$ with $H := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{w}^\mathsf{T}\boldsymbol{z} = 0\}$.



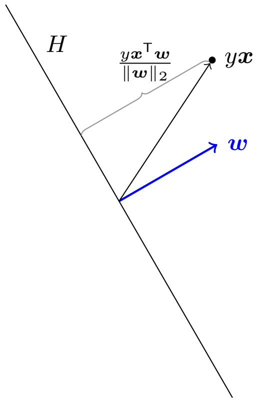
Formulation #1: constrain the norm.
Single margin $\frac{y\boldsymbol{x}^\mathsf{T}\boldsymbol{w}}{\|\boldsymbol{w}\|}$.
Overall margin $\min_i \frac{y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}}{\|\boldsymbol{w}\|}$.
Max margin $\max_{\|\boldsymbol{u}\|=1} \min_i y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{u}$.

Formulation #2: constrain the margins.
Consider any $\boldsymbol{v}$ with $\min_i y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{v} \geq 1$.
Can suppose $y_k\boldsymbol{x}_k^\mathsf{T}\boldsymbol{v} = 1$ for some k (why?).
Since margin scales with $\frac{1}{\|\boldsymbol{v}\|}$, choose

$$\begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{v}\|^2 \\ \text{subject to} \quad & \boldsymbol{v} \in \mathbb{R}^d, \\ & y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{v} \geq 1 \quad \forall i. \end{aligned}$$

How to write “maximum margin classifier”?

Input $\boldsymbol{x} \in \mathbb{R}^d$, label $y \in \{\pm 1\}$, predictor $\boldsymbol{w} \in \mathbb{R}^d$ with $H := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{w}^\mathsf{T}\boldsymbol{z} = 0\}$.



Formulation #1: constrain the norm.
Single margin $\frac{y\boldsymbol{x}^\mathsf{T}\boldsymbol{w}}{\|\boldsymbol{w}\|}$.
Overall margin $\min_i \frac{y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}}{\|\boldsymbol{w}\|}$.
Max margin $\max_{\|\boldsymbol{u}\|=1} \min_i y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{u}$.

Formulation #2: constrain the margins.
Consider any $\boldsymbol{v}$ with $\min_i y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{v} \geq 1$.
Can suppose $y_k\boldsymbol{x}_k^\mathsf{T}\boldsymbol{v} = 1$ for some k (why?).
Since margin scales with $\frac{1}{\|\boldsymbol{v}\|}$, choose

$$\begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{v}\|^2 \\ \text{subject to} \quad & \boldsymbol{v} \in \mathbb{R}^d, \\ & y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{v} \geq 1 \quad \forall i. \end{aligned}$$

These two are equivalent
(up to scaling, when solutions exist).

Hard-margin SVM.

Take the solution to either optimization problem:

$$\begin{aligned} \max \quad & \min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{u}, \\ \text{subject to} \quad & \boldsymbol{u} \in \mathbb{R}^d, \\ & \|\boldsymbol{u}\| = 1; \end{aligned} \qquad\qquad \begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{v}\|^2 \\ \text{subject to} \quad & \boldsymbol{v} \in \mathbb{R}^d, \\ & y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v} \geq 1 \quad \forall i. \end{aligned}$$

Hard-margin SVM.

Take the solution to either optimization problem:

$$\begin{aligned} \max \quad & \min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{u}, \\ \text{subject to} \quad & \boldsymbol{u} \in \mathbb{R}^d, \\ & \|\boldsymbol{u}\| = 1; \end{aligned} \qquad\qquad \begin{aligned} \min \quad & \frac{1}{2}\|\boldsymbol{v}\|^2 \\ \text{subject to} \quad & \boldsymbol{v} \in \mathbb{R}^d, \\ & y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v} \geq 1 \quad \forall i. \end{aligned}$$

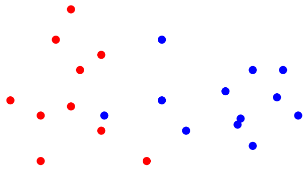
Remarks.

- Since the second is a convex program, many approaches exist. Homework will investigate a simple SGD-based strategy.
- What happens if the second formulation is infeasible? (That is, what if no vector $\boldsymbol{v}$ satisfies $\min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v} \geq 1$?)

Soft-margin SVM

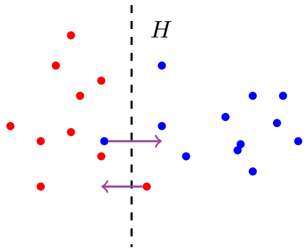
Soft-margin SVM

What is the max margin predictor for the following data?



Soft-margin SVM

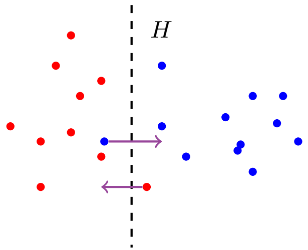
What is the max margin predictor for the following data?



Idea. pay a price for each $y_i \boldsymbol{x}_i^\top \boldsymbol{v} < 1$ with slack variables $(\xi_i)_{i=1}^n$:

Soft-margin SVM

What is the max margin predictor for the following data?



Idea. pay a price for each $y_i \boldsymbol{x}_i^\top \boldsymbol{v} < 1$ with slack variables $(\xi_i)_{i=1}^n$:

$$
\begin{aligned}
\min_{\boldsymbol{w}\in\mathbb{R}^d, \xi_1,\ldots,\xi_n\in\mathbb{R}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i \\
\text{s.t.} \quad & y_i \boldsymbol{x}_i^\top \boldsymbol{w} \geq 1 - \xi_i && \text{for all } i = 1, 2, \ldots, n, \\
& \xi_i \geq 0 && \text{for all } i = 1, 2, \ldots, n.
\end{aligned}
$$

Re-formulation as **regularized ERM**.

Formulation with slack variables $(\xi_i)_{i=1}^n$.

$$
\begin{aligned}
\min_{\boldsymbol{w}\in\mathbb{R}^d,\xi_1,\dots,\xi_n\in\mathbb{R}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i \\
\text{s.t.} \quad & y_i \boldsymbol{x}_i^\top \boldsymbol{w} \geq 1 - \xi_i && \text{for all } i = 1, 2, \dots, n, \\
& \xi_i \geq 0 && \text{for all } i = 1, 2, \dots, n.
\end{aligned}
$$

Re-formulation as regularized ERM.

Formulation with slack variables $(\xi_i)_{i=1}^n$.

$$\begin{aligned}
\min_{\boldsymbol{w}\in\mathbb{R}^d,\xi_1,\ldots,\xi_n\in\mathbb{R}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i \\
\text{s.t.} \quad & y_i \boldsymbol{x}_i^\mathsf{T}\boldsymbol{w} \geq 1-\xi_i && \text{for all } i=1,2,\ldots,n, \\
& \xi_i \geq 0 && \text{for all } i=1,2,\ldots,n.
\end{aligned}$$

Regularized ERM formulation.

Given any $\boldsymbol{w}$, choose $\xi_i := \max\{0, 1-y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}\}$, whereby

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \max\left\{0, 1-y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}\right\},$$

where $\ell_{\text{hinge}}(y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}) := \max\{0, 1-y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w})$ is the hinge loss.

Re-formulation as regularized ERM.

Formulation with slack variables $(\xi_i)_{i=1}^n$.

$$\begin{aligned} \min_{\boldsymbol{w}\in\mathbb{R}^d,\xi_1,\ldots,\xi_n\in\mathbb{R}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w} \geq 1 - \xi_i \qquad \text{for all } i = 1,2,\ldots,n, \\ & \xi_i \geq 0 \qquad \text{for all } i = 1,2,\ldots,n. \end{aligned}$$

Regularized ERM formulation.
Given any $\boldsymbol{w}$, choose $\xi_i := \max\{0, 1 - y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}\}$, whereby

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \max\left\{0, 1 - y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}\right\},$$

where $\ell_{\text{hinge}}(y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}) := \max\{0, 1 - y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w})$ is the hinge loss.

Remarks.

- Normally we'd write $\frac{1}{n}\sum_{i=1}^n \ell_{\text{hinge}}(y_i\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$.
- C (or λ) is a hyper-parameter; it has no good search procedure.

SVM duality

SVM duality

A convex program is an optimization problem (minimization or maximization) where a convex objective is minimized over a convex constraint (feasible) set.

SVM duality

A convex program is an optimization problem (minimization or maximization) where a convex objective is minimized over a convex constraint (feasible) set.

Every convex program has a corresponding dual program.
For the SVM, the dual has many nice properties:

- Clarifies the role of support vectors.
- Leads to a nice nonlinear approach via kernels.
- Gives another choice for optimization algorithms.

SVM hard-margin duality.
Define the two optimization problems

$$\min \left\{ \frac{1}{2}\|\boldsymbol{w}\|^2 \ : \ \boldsymbol{w} \in \mathbb{R}^d, \forall i \,.\, 1 - y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w} \le 0 \right\} \qquad \text{(primal)},$$

$$\max \left\{ \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\mathsf{T} \boldsymbol{x}_j \ : \ \boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\alpha} \ge 0 \right\} \qquad \text{(dual)}.$$

If the primal is feasible,
then the primal optimal value equals the dual optimal value.
Given a primal optimum $\bar{\boldsymbol{w}}$ and a dual optimum $\bar{\boldsymbol{\alpha}}$, they satisfy

$$\bar{\boldsymbol{w}} = \sum_{i=1}^n \bar{\alpha}_i y_i \boldsymbol{x}_i.$$

SVM hard-margin duality.
Define the two optimization problems

$$\min\left\{\frac{1}{2}\|\boldsymbol{w}\|^2 \;:\; \boldsymbol{w}\in\mathbb{R}^d, \forall i\,\boldsymbol{.}\, 1 - y_i\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{w} \le 0\right\} \qquad \text{(primal)},$$

$$\max\left\{\sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_j \;:\; \boldsymbol{\alpha}\in\mathbb{R}^n, \boldsymbol{\alpha}\ge 0\right\} \qquad \text{(dual)}.$$

If the primal is feasible,
then the primal optimal value equals the dual optimal value.
Given a primal optimum $\bar{\boldsymbol{w}}$ and a dual optimum $\bar{\boldsymbol{\alpha}}$, they satisfy

$$\bar{\boldsymbol{w}} = \sum_{i=1}^n \bar{\alpha}_i y_i \boldsymbol{x}_i.$$

- The dual variables $\boldsymbol{\alpha}$ have dimension n, same as examples.
- We can write the primal optimum as a linear combination of examples.
- The dual objective is a concave quadratic.
- We will derive this duality using Lagrange multipliers.

Lagrange multipliers

Move constraints to objective using Lagrange multipliers.

$$\text{Original problem:} \quad \min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2$$
$$\text{s.t.} \quad 1 - y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w} \leq 0 \quad \text{for all } i = 1, \dots, n.$$

- For each constraint $1 - y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w} \leq 0$, associate a dual variable (Lagrange multiplier) $\alpha_i \geq 0$.
- Move constraints to objective by adding $\sum_{i=1}^n \alpha_i (1 - y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w})$ and maximizing over $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ s.t. $\boldsymbol{\alpha} \geq \mathbf{0}$ (i.e., $\alpha_i \geq 0$ for all i).

Lagrange multipliers

Move constraints to objective using Lagrange multipliers.

$$\text{Original problem:} \quad \min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2$$
$$\text{s.t.} \quad 1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w} \le 0 \quad \text{for all } i = 1, \dots, n.$$

- For each constraint $1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w} \le 0$, associate a dual variable (Lagrange multiplier) $\alpha_i \ge 0$.
- Move constraints to objective by adding $\sum_{i=1}^n \alpha_i(1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w})$ and maximizing over $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ s.t. $\boldsymbol{\alpha} \ge \mathbf{0}$ (i.e., $\alpha_i \ge 0$ for all i).

Lagrangian $L(\boldsymbol{w}, \boldsymbol{\alpha})$:

$$L(\boldsymbol{w}, \boldsymbol{\alpha}) := \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^n \alpha_i(1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w}).$$

Maximizing over $\boldsymbol{\alpha} \ge 0$ recovers primal problem: for any $\boldsymbol{w} \in \mathbb{R}^d$,

$$P(\boldsymbol{w}) := \sup_{\boldsymbol{\alpha}\ge 0} L(\boldsymbol{w}, \boldsymbol{\alpha}) = \begin{cases} \frac{1}{2}\|\boldsymbol{w}\|_2^2 & \text{if } \min_i y_i \boldsymbol{x}_i^\top \boldsymbol{w} \ge 1, \\ \infty & \text{otherwise.} \end{cases}$$

Lagrange multipliers

Move constraints to objective using Lagrange multipliers.

$$\text{Original problem:} \quad \min_{\boldsymbol{w} \in \mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2$$
$$\text{s.t.} \quad 1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w} \le 0 \quad \text{for all } i = 1, \dots, n.$$

- For each constraint $1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w} \le 0$, associate a dual variable (Lagrange multiplier) $\alpha_i \ge 0$.
- Move constraints to objective by adding $\sum_{i=1}^n \alpha_i (1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w})$ and maximizing over $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ s.t. $\boldsymbol{\alpha} \ge \boldsymbol{0}$ (i.e., $\alpha_i \ge 0$ for all i).

Lagrangian $L(\boldsymbol{w}, \boldsymbol{\alpha})$:

$$L(\boldsymbol{w}, \boldsymbol{\alpha}) := \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^n \alpha_i (1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w}).$$

Maximizing over $\boldsymbol{\alpha} \ge 0$ recovers primal problem: for any $\boldsymbol{w} \in \mathbb{R}^d$,

$$P(\boldsymbol{w}) := \sup_{\boldsymbol{\alpha} \ge 0} L(\boldsymbol{w}, \boldsymbol{\alpha}) = \begin{cases} \frac{1}{2}\|\boldsymbol{w}\|_2^2 & \text{if } \min_i y_i \boldsymbol{x}_i^\top \boldsymbol{w} \ge 1, \\ \infty & \text{otherwise.} \end{cases}$$

What if we leave $\boldsymbol{\alpha}$ fixed, and minimize $\boldsymbol{w}$?

Dual problem

Lagrangian

$$L(\boldsymbol{w}, \boldsymbol{\alpha}) := \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w}).$$

Primal hard-margin SVM

$$P(\boldsymbol{w}) = \sup_{\boldsymbol{\alpha} \geq \boldsymbol{0}} L(\boldsymbol{w}, \boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha} \geq \boldsymbol{0}} \left[\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w}) \right].$$

Dual problem

Lagrangian

$$L(\boldsymbol{w}, \boldsymbol{\alpha}) := \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^\intercal \boldsymbol{w}).$$

Primal hard-margin SVM

$$P(\boldsymbol{w}) = \sup_{\boldsymbol{\alpha} \geq \boldsymbol{0}} L(\boldsymbol{w}, \boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha} \geq \boldsymbol{0}} \left[\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^\intercal \boldsymbol{w}) \right].$$

Dual problem $D(\boldsymbol{\alpha}) = \min_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{\alpha})$:

Dual problem

Lagrangian

$$L(\boldsymbol{w}, \boldsymbol{\alpha}) := \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w}).$$

Primal hard-margin SVM

$$P(\boldsymbol{w}) = \sup_{\boldsymbol{\alpha} \geq \boldsymbol{0}} L(\boldsymbol{w}, \boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha} \geq \boldsymbol{0}} \left[\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^\top \boldsymbol{w}) \right].$$

Dual problem $D(\boldsymbol{\alpha}) = \min_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{\alpha})$: given $\boldsymbol{\alpha} \geq 0$, then $\boldsymbol{w} \mapsto L(\boldsymbol{w}, \boldsymbol{\alpha})$ is a convex quadratic with minimum $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$, giving

$$\begin{aligned} D(\boldsymbol{\alpha}) = \min_{\boldsymbol{w} \in \mathbb{R}^d} L(\boldsymbol{w}, \boldsymbol{\alpha}) = L\left(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i, \boldsymbol{\alpha} \right) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \right\|_2^2 \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j. \end{aligned}$$

Summarizing,

$$L(\boldsymbol{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w}) \qquad \text{Lagrangian,}$$

$$P(\boldsymbol{w}) = \max_{\boldsymbol{\alpha} \geq 0} L(\boldsymbol{w}, \boldsymbol{\alpha}) \qquad \text{primal problem,}$$

$$D(\boldsymbol{\alpha}) = \min_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{\alpha}) \qquad \text{dual problem.}$$

Summarizing,

$$
\begin{aligned}
L(\boldsymbol{w}, \boldsymbol{\alpha}) &= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w}) && \text{Lagrangian,} \\
P(\boldsymbol{w}) &= \max_{\boldsymbol{\alpha} \geq 0} L(\boldsymbol{w}, \boldsymbol{\alpha}) && \text{primal problem,} \\
D(\boldsymbol{\alpha}) &= \min_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{\alpha}) && \text{dual problem.}
\end{aligned}
$$

- For general Lagrangians, have weak duality

$$
P(\boldsymbol{w}) \geq D(\boldsymbol{\alpha}),
$$

since $P(\boldsymbol{w}) = \max_{\boldsymbol{\alpha}' \geq 0} L(\boldsymbol{w}, \boldsymbol{\alpha}') \geq L(\boldsymbol{w}, \boldsymbol{\alpha}) \geq \min_{\boldsymbol{w}'} L(\boldsymbol{w}', \boldsymbol{\alpha}) = D(\boldsymbol{\alpha})$.

Summarizing,

$$\begin{aligned} L(\boldsymbol{w}, \boldsymbol{\alpha}) &= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}) && \text{Lagrangian,} \\ P(\boldsymbol{w}) &= \max_{\boldsymbol{\alpha} \geq 0} L(\boldsymbol{w}, \boldsymbol{\alpha}) && \text{primal problem,} \\ D(\boldsymbol{\alpha}) &= \min_{\boldsymbol{w}} L(\boldsymbol{w}, \boldsymbol{\alpha}) && \text{dual problem.} \end{aligned}$$

- For general Lagrangians, have weak duality

$$P(\boldsymbol{w}) \geq D(\boldsymbol{\alpha}),$$

  since $P(\boldsymbol{w}) = \max_{\boldsymbol{\alpha}' \geq 0} L(\boldsymbol{w}, \boldsymbol{\alpha}') \geq L(\boldsymbol{w}, \boldsymbol{\alpha}) \geq \min_{\boldsymbol{w}'} L(\boldsymbol{w}', \boldsymbol{\alpha}) = D(\boldsymbol{\alpha})$.

- By convexity, have strong duality $\min_{\boldsymbol{w}} P(\boldsymbol{w}) = \max_{\boldsymbol{\alpha} \geq 0} D(\boldsymbol{\alpha})$, and an optimum $\bar{\boldsymbol{\alpha}}$ for D gives an optimum $\bar{\boldsymbol{w}}$ for P via

$$\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{x}_i = \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}, \bar{\boldsymbol{\alpha}}).$$

Support vectors

Optimal solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_n)$ satisfy

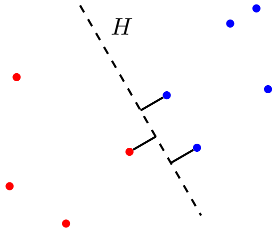
- $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{x}_i = \sum_{i:\bar{\alpha}_i > 0} \bar{\alpha}_i y_i \boldsymbol{x}_i$,
- $\bar{\alpha}_i > 0 \quad \Rightarrow \quad y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}} = 1$ for all $i = 1, \ldots, n$ (complementary slackness).

Support vectors

Optimal solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_n)$ satisfy

- $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{x}_i = \sum_{i:\bar{\alpha}_i > 0} \bar{\alpha}_i y_i \boldsymbol{x}_i$,
- $\bar{\alpha}_i > 0 \quad \Rightarrow \quad y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}} = 1$ for all $i = 1, \ldots, n$ (complementary slackness).

The $y_i \boldsymbol{x}_i$ where $\bar{\alpha}_i > 0$ are called support vectors.

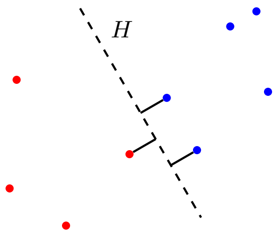


Support vectors

Optimal solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \dots, \bar{\alpha}_n)$ satisfy

- $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{x}_i = \sum_{i:\bar{\alpha}_i > 0} \bar{\alpha}_i y_i \boldsymbol{x}_i$,
- $\bar{\alpha}_i > 0 \quad \Rightarrow \quad y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}} = 1$ for all $i = 1, \dots, n$ (complementary slackness).

The $y_i \boldsymbol{x}_i$ where $\bar{\alpha}_i > 0$ are called support vectors.



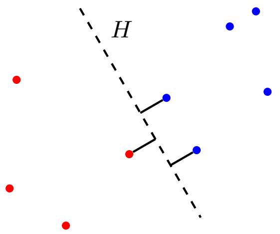
- Support vector examples satisfy "margin" constraints with equality.

Support vectors

Optimal solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \dots, \bar{\alpha}_n)$ satisfy

- $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{x}_i = \sum_{i:\bar{\alpha}_i > 0} \bar{\alpha}_i y_i \boldsymbol{x}_i$,
- $\bar{\alpha}_i > 0 \quad \Rightarrow \quad y_i \boldsymbol{x}_i^\top \bar{\boldsymbol{w}} = 1$ for all $i = 1, \dots, n$ (complementary slackness).

The $y_i \boldsymbol{x}_i$ where $\bar{\alpha}_i > 0$ are called support vectors.



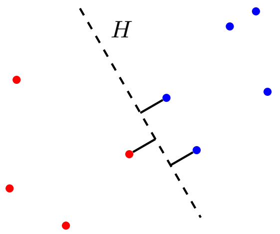
- Support vector examples satisfy "margin" constraints with equality.
- Get same solution if non-support vectors deleted.

Support vectors

Optimal solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \dots, \bar{\alpha}_n)$ satisfy

- $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{x}_i = \sum_{i:\bar{\alpha}_i > 0} \bar{\alpha}_i y_i \boldsymbol{x}_i,$
- $\bar{\alpha}_i > 0 \quad \Rightarrow \quad y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}} = 1$ for all $i = 1, \dots, n$ (complementary slackness).

The $y_i \boldsymbol{x}_i$ where $\bar{\alpha}_i > 0$ are called support vectors.



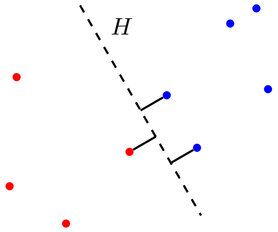
- Support vector examples satisfy "margin" constraints with equality.
- Get same solution if non-support vectors deleted.
- Primal optimum is a linear combination of support vectors.

Support vectors

Optimal solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \dots, \bar{\alpha}_n)$ satisfy

- $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{x}_i = \sum_{i:\bar{\alpha}_i > 0} \bar{\alpha}_i y_i \boldsymbol{x}_i$,
- $\bar{\alpha}_i > 0 \quad \Rightarrow \quad y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}} = 1$ for all $i = 1, \dots, n$ (complementary slackness).

The $y_i \boldsymbol{x}_i$ where $\bar{\alpha}_i > 0$ are called support vectors.



- Support vector examples satisfy "margin" constraints with equality.
- Get same solution if non-support vectors deleted.
- Primal optimum is a linear combination of support vectors.
- Dual solution and support vectors not necessarily unique! (Why not?)

Proof of complementary slackness

For the optimal (feasible) solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}}$, we have

$$P(\bar{\boldsymbol{w}}) = D(\bar{\boldsymbol{\alpha}}) = \min_{\boldsymbol{w} \in \mathbb{R}^d} L(\boldsymbol{w}, \bar{\boldsymbol{\alpha}}) \qquad \text{(by strong duality)}$$

Proof of complementary slackness

For the optimal (feasible) solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}}$, we have

$$\begin{aligned} P(\bar{\boldsymbol{w}}) = D(\bar{\boldsymbol{\alpha}}) &= \min_{\boldsymbol{w}\in\mathbb{R}^d} L(\boldsymbol{w}, \bar{\boldsymbol{\alpha}}) \qquad \text{(by strong duality)} \\ &\leq L(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\alpha}}) \end{aligned}$$

Proof of complementary slackness

For the optimal (feasible) solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}}$, we have

$$
\begin{aligned}
P(\bar{\boldsymbol{w}}) = D(\bar{\boldsymbol{\alpha}}) &= \min_{\boldsymbol{w} \in \mathbb{R}^d} L(\boldsymbol{w}, \bar{\boldsymbol{\alpha}}) \qquad \text{(by strong duality)} \\
&\leq L(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\alpha}}) \\
&= \frac{1}{2}\|\bar{\boldsymbol{w}}\|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i \boldsymbol{x}_i^\mathsf{T} \bar{\boldsymbol{w}})
\end{aligned}
$$

Proof of complementary slackness

For the optimal (feasible) solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}}$, we have

$$\begin{aligned}
P(\bar{\boldsymbol{w}}) = D(\bar{\boldsymbol{\alpha}}) &= \min_{\boldsymbol{w}\in\mathbb{R}^d} L(\boldsymbol{w}, \bar{\boldsymbol{\alpha}}) \qquad \text{(by strong duality)} \\
&\leq L(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\alpha}}) \\
&= \frac{1}{2}\|\bar{\boldsymbol{w}}\|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i \boldsymbol{x}_i^\mathsf{T} \bar{\boldsymbol{w}}) \\
&\leq \frac{1}{2}\|\bar{\boldsymbol{w}}\|_2^2 \qquad \text{(constraints are satisfied)} \\
&= P(\bar{\boldsymbol{w}}).
\end{aligned}$$

Proof of complementary slackness

For the optimal (feasible) solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}}$, we have

$$
\begin{aligned}
P(\bar{\boldsymbol{w}}) \;=\; D(\bar{\boldsymbol{\alpha}}) \;&=\; \min_{\boldsymbol{w}\in\mathbb{R}^d} L(\boldsymbol{w}, \bar{\boldsymbol{\alpha}}) \qquad \text{(by strong duality)} \\
&\leq\; L(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\alpha}}) \\
&=\; \frac{1}{2}\|\bar{\boldsymbol{w}}\|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}}) \\
&\leq\; \frac{1}{2}\|\bar{\boldsymbol{w}}\|_2^2 \qquad \text{(constraints are satisfied)} \\
&=\; P(\bar{\boldsymbol{w}}).
\end{aligned}
$$

Therefore, every term in sum $\sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}})$ must be zero:

$$
\bar{\alpha}_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}}) = 0 \quad \text{for all } i = 1, \ldots, n.
$$

Proof of complementary slackness

For the optimal (feasible) solutions $\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\alpha}}$, we have

$$\begin{aligned}
P(\bar{\boldsymbol{w}}) = D(\bar{\boldsymbol{\alpha}}) &= \min_{\boldsymbol{w} \in \mathbb{R}^d} L(\boldsymbol{w}, \bar{\boldsymbol{\alpha}}) && \text{(by strong duality)} \\
&\leq L(\bar{\boldsymbol{w}}, \bar{\boldsymbol{\alpha}}) \\
&= \frac{1}{2}\|\bar{\boldsymbol{w}}\|_2^2 + \sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}}) \\
&\leq \frac{1}{2}\|\bar{\boldsymbol{w}}\|_2^2 && \text{(constraints are satisfied)} \\
&= P(\bar{\boldsymbol{w}}).
\end{aligned}$$

Therefore, every term in sum $\sum_{i=1}^{n} \bar{\alpha}_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}})$ must be zero:

$$\bar{\alpha}_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}}) = 0 \quad \text{for all } i = 1, \ldots, n.$$

If $\bar{\alpha}_i > 0$, then must have $1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}} = 0$. (Not iff!)

SVM (hard-margin) duality summary

Lagrangian

$$L(\boldsymbol{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w}).$$

Primal maximum margin problem was

$$P(\boldsymbol{w}) = \sup_{\boldsymbol{\alpha} \geq \mathbf{0}} L(\boldsymbol{w}, \boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha} \geq \mathbf{0}} \left[\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{w}) \right].$$

Dual problem

$$D(\boldsymbol{\alpha}) = \min_{\boldsymbol{w} \in \mathbb{R}^d} L(\boldsymbol{w}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \right\|_2^2.$$

Given dual optimum $\bar{\boldsymbol{\alpha}}$,

- Corresponding primal optimum $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$;
- Strong duality $P(\bar{\boldsymbol{w}}) = D(\bar{\boldsymbol{\alpha}})$;
- $\bar{\alpha}_i > 0$ implies $y_i \boldsymbol{x}_i^{\mathsf{T}} \bar{\boldsymbol{w}} = 1$,
  and these $y_i \boldsymbol{x}_i$ are support vectors.

SVM soft-margin dual

SVM soft-margin dual

Similarly,

$$
\begin{aligned}
L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) &= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i(1 - \xi_i - y_i \boldsymbol{x}_i^\top \boldsymbol{w}) && \text{(Lagrangian)},\\
P(\boldsymbol{w}, \boldsymbol{\xi}) &= \sup_{\boldsymbol{\alpha} \geq 0} L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) && \text{(Primal)},\\
&= \begin{cases} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^n \xi_i & \forall i \,.\, 1 - \xi_i - y_i \boldsymbol{x}_i^\top \boldsymbol{w} \leq 0,\\ \infty & \text{otherwise}, \end{cases}\\
D(\boldsymbol{\alpha}) &= \min_{\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{\xi} \in \mathbb{R}^n_{\geq 0}} L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) && \text{(Dual)},\\
&= \max_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^n \\ 0 \leq \alpha_i \leq C}} \left[\sum_{i=1}^n \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^n \alpha_i y_i \boldsymbol{x}_i\right\|^2\right].
\end{aligned}
$$

SVM soft-margin dual

Similarly,

$$
\begin{aligned}
L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) &= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1-\xi_i - y_i\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{w}) && \text{(Lagrangian)},\\
P(\boldsymbol{w}, \boldsymbol{\xi}) &= \sup_{\boldsymbol{\alpha}\geq 0} L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) && \text{(Primal)},\\
&= \begin{cases} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n}\xi_i & \forall i \,.\, 1-\xi_i - y_i\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{w} \leq 0,\\ \infty & \text{otherwise}, \end{cases}\\
D(\boldsymbol{\alpha}) &= \min_{\boldsymbol{w}\in\mathbb{R}^d, \boldsymbol{\xi}\in\mathbb{R}^n_{\geq 0}} L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) && \text{(Dual)},\\
&= \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n\\ 0\leq\alpha_i\leq C}} \left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right\|^2\right].
\end{aligned}
$$

Remarks.

- Dual solution $\bar{\boldsymbol{\alpha}}$ still gives primal solution $\bar{\boldsymbol{w}} = \sum_{i=1}^{n}\bar{\alpha}_i y_i \boldsymbol{x}_i$.
- Can take $C \to \infty$ to recover hard-margin case.
- Dual is still a constrained concave quadratic (used in many solvers).
- Some presentations include bias in primal $(\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{w} + b)$; this introduces a constraint $\sum_{i=1}^{n} y_i\alpha_i = 0$ in dual.

Nonlinear SVM: feature mapping annoying in the primal?

Nonlinear SVM: feature mapping annoying in the primal?

SVM hard-margin primal, with a feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$:

$$\min \left\{ \frac{1}{2} \|\boldsymbol{w}\|^2 \;:\; \boldsymbol{w} \in \mathbb{R}^p, \forall i \,.\, \phi(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{w} \geq 1 \right\}.$$

Now the search space has p dimensions; potentially $p \gg d$.

Can we do better?

Feature mapping in the dual

Feature mapping in the dual

SVM hard-margin dual, with a feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$:

$$\max_{\alpha_1, \alpha_2, \ldots, \alpha_n \geq 0} \qquad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\mathsf{T}} \phi(\boldsymbol{x}_j).$$

Given dual optimum $\bar{\boldsymbol{\alpha}}$, since $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(\boldsymbol{x}_i)$, we can predict on future $\boldsymbol{x}$ with

$$\boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})^{\mathsf{T}} \bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{\phi}(\boldsymbol{x})^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_i).$$

Feature mapping in the dual

SVM hard-margin dual, with a feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$:

$$\max_{\alpha_1,\alpha_2,\ldots,\alpha_n \geq 0} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i)^\mathsf{T} \phi(\boldsymbol{x}_j).$$

Given dual optimum $\bar{\boldsymbol{\alpha}}$, since $\bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \phi(\boldsymbol{x}_i)$, we can predict on future $\boldsymbol{x}$ with

$$\boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})^\mathsf{T} \bar{\boldsymbol{w}} = \sum_{i=1}^{n} \bar{\alpha}_i y_i \boldsymbol{\phi}(\boldsymbol{x})^\mathsf{T} \boldsymbol{\phi}(\boldsymbol{x}_i).$$

- Dual form never needs $\phi(\boldsymbol{x}) \in \mathbb{R}^p$, only $\phi(\boldsymbol{x})^\mathsf{T} \phi(\boldsymbol{x}_i) \in \mathbb{R}$.
- **Kernel trick:** replace every $\phi(\boldsymbol{x})^\mathsf{T} \phi(\boldsymbol{x}')$ with **kernel evaluation** $k(\boldsymbol{x}, \boldsymbol{x}')$. Sometimes $k(\cdot, \cdot)$ is much cheaper than $\phi(\boldsymbol{x})^\mathsf{T} \phi(\boldsymbol{x}')$.
- This idea started with SVM, but appears in many other places.
- **Downside:** implementations usually store Gram matrix $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ where $G_{ij} := k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Kernel example: affine features

Affine features: $\phi : \mathbb{R}^d \to \mathbb{R}^{1+d}$, where

$$\phi(\boldsymbol{x}) = (1, x_1, \ldots, x_d).$$

Kernel form:

$$\boldsymbol{\phi}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\phi}(\boldsymbol{x}') = 1 + \boldsymbol{x}^\mathsf{T}\boldsymbol{x}'.$$

Kernel example: quadratic features

Consider re-normalized quadratic features
$\phi\colon \mathbb{R}^d \to \mathbb{R}^{1+2d+\binom{d}{2}}$, where

$$\phi(\boldsymbol{x}) = \Big(1,\ \sqrt{2}x_1,\ \ldots, \sqrt{2}x_d,\ x_1^2,\ \ldots,\ x_d^2,$$
$$\sqrt{2}x_1x_2,\ \ldots,\ \sqrt{2}x_1x_d,\ \ldots,\ \sqrt{2}x_{d-1}x_d\Big).$$

Just writing this down takes time $\mathcal{O}(d^2)$.
Meanwhile,

$$\phi(\boldsymbol{x})^\mathsf{T}\phi(\boldsymbol{x}') = (1 + \boldsymbol{x}^\mathsf{T}\boldsymbol{x}')^2,$$

time $\mathcal{O}(d)$.

Kernel example: quadratic features

Consider re-normalized quadratic features
$\phi \colon \mathbb{R}^d \to \mathbb{R}^{1+2d+\binom{d}{2}}$, where

$$\phi(\boldsymbol{x}) = \Big(1,\ \sqrt{2}x_1,\ \dots, \sqrt{2}x_d,\ x_1^2,\ \dots,\ x_d^2,$$
$$\sqrt{2}x_1x_2,\ \dots,\ \sqrt{2}x_1x_d,\ \dots,\ \sqrt{2}x_{d-1}x_d\Big).$$

Just writing this down takes time $\mathcal{O}(d^2)$.
Meanwhile,

$$\phi(\boldsymbol{x})^{\mathsf{T}}\phi(\boldsymbol{x}') = (1 + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}')^2,$$

time $\mathcal{O}(d)$.

Tweaks:

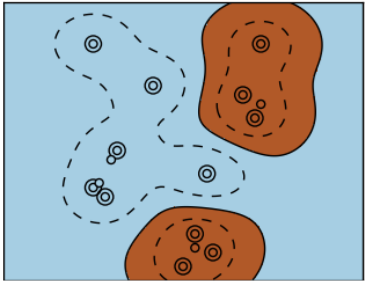
- What if we change exponent "2"?
- What if we replace additive "1" with 0?

RBF kernel

For any $\sigma > 0$, there is an infinite feature expansion $\boldsymbol{\phi}\colon \mathbb{R}^d \to \mathbb{R}^\infty$ such that

$$\boldsymbol{\phi}(\boldsymbol{x})^\intercal \boldsymbol{\phi}(\boldsymbol{x}') = \exp\left(-\frac{\left\|\boldsymbol{x}-\boldsymbol{x}'\right\|_2^2}{2\sigma^2}\right),$$

which can be computed in $O(d)$ time.



This is called a Gaussian kernel or RBF kernel.
It has some similarities to nearest neighbor methods (later lecture).

ϕ maps to an infinite-dimensional space, but there's no reason to know that.

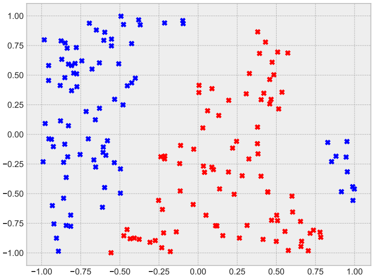
Defining kernels without ϕ

A (positive definite) kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric function so that for any n and any data examples $(\boldsymbol{x}_i)_{i=1}^n$, the corresponding Gram matrix $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ with $G_{ij} := k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive semi-definite.

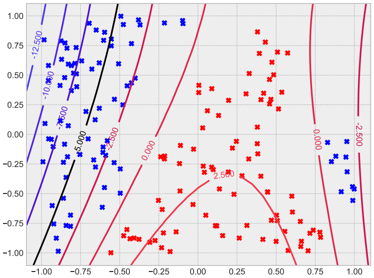
Defining kernels without ϕ

A (positive definite) kernel function $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric function so that for any n and any data examples $(\boldsymbol{x}_i)_{i=1}^n$, the corresponding Gram matrix $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ with $G_{ij} := k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive semi-definite.

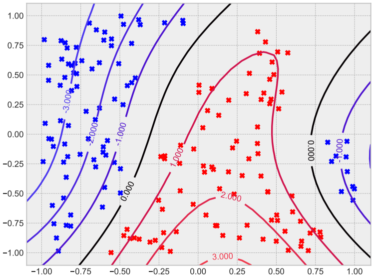
- There is a ton of theory about this formalism;
  e.g., keywords RKHS, representer theorem, Mercer's theorem.
- Given any such k, there always exists a corresponding ϕ.
- This definition ensures the SVM dual is still concave.



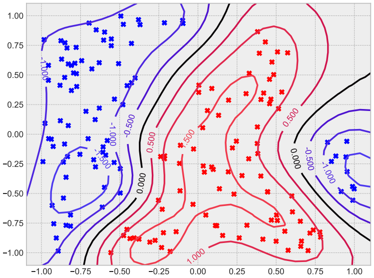
Source data.



Quadratic SVM.



RBF SVM ($\sigma = 1$).



RBF SVM ($\sigma = 0.1$).

Summary for SVM

- Hard-margin SVM.
- Soft-margin SVM.
- SVM duality.
- Nonlinear SVM: kernels

(Appendix.)

Equivalence of two hard-margin formulations.

Proof. First note that both have unique solutions (when feasible). For the first formulation, suppose we have two solutions $\boldsymbol{u}$ and $\boldsymbol{u}'$, and define another vector $\boldsymbol{u}'' := (\boldsymbol{u} + \boldsymbol{u}')/2$. Then $\boldsymbol{u}''$ achieves the same margin value as $\boldsymbol{u}$ and $\boldsymbol{u}'$, but if $\boldsymbol{u} \neq \boldsymbol{u}'$, then $\|\boldsymbol{u}''\| < 1$, which means $\boldsymbol{u}''/\|\boldsymbol{u}''\|$ achieves a larger margin value than the purported optima $\boldsymbol{u}$ and $\boldsymbol{u}'$, a contradiction. For the second formulation, it suffices to note that the objective is strictly convex.

Now consider solution $\boldsymbol{u}$ to the first, with margin γ. Then $\boldsymbol{v} := \boldsymbol{u}/\gamma$ is feasible for second, with optimal value $1/(2\gamma^2)$. So the optimal value is at most this; if it is exactly the optimal value, we are done, otherwise suppose the optimum $\bar{\boldsymbol{v}}$ has $\|\bar{\boldsymbol{v}}\|^2/2 = 1/(2\rho^2) < 1/(2\gamma^2)$. Then $\bar{\boldsymbol{u}} := \rho\bar{\boldsymbol{v}}$ is a unit vector, and moreover has $\min_i y_i \boldsymbol{x}_i^\top \bar{\boldsymbol{u}} = \rho > \gamma$, a contradiction since this is better than the supposed optimum $\boldsymbol{u}$.

$\square$

Soft-margin dual derivation

Let's derive the final dual form:

$$L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w})$$

$$D(\boldsymbol{\alpha}) = \min_{\boldsymbol{w}\in\mathbb{R}^d, \boldsymbol{\xi}\in\mathbb{R}^n_{\geq 0}} L(\boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n \\ 0\leq\alpha_i\leq C}} \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i\right\|^2\right].$$

Given $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$, the minimizing $\boldsymbol{w}$ is still $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$; plugging in,

$$D(\boldsymbol{\alpha}) = \min_{\boldsymbol{\xi}\in\mathbb{R}^n_{\geq 0}} L\left(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i, \boldsymbol{\xi}, \boldsymbol{\alpha}\right) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i\right\|^2 + \sum_{i=1}^{n} \xi_i (C - \alpha_i).$$

The goal is to maximize D; if $\alpha_i > C$, then $\xi_i \uparrow \infty$ gives $D(\boldsymbol{\alpha}) = -\infty$. Otherwise, minimized at $\xi_i = 0$. Therefore the dual problem is

$$\max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n \\ 0\leq\alpha_i\leq C}} \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i\right\|^2\right].$$

Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi \colon \mathbb{R} \to \mathbb{R}^\infty$.

What ϕ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi\colon \mathbb{R} \to \mathbb{R}^\infty$.

What ϕ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$e^{-(x-y)^2/(2\sigma^2)} = e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2}$$

Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi\colon \mathbb{R} \to \mathbb{R}^\infty$.

What ϕ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$\begin{aligned} e^{-(x-y)^2/(2\sigma^2)} &= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2} \\ &= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{xy}{\sigma^2}\right)^k \end{aligned}$$

Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi\colon \mathbb{R} \to \mathbb{R}^\infty$.

What ϕ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$\begin{aligned} e^{-(x-y)^2/(2\sigma^2)} &= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2} \\ &= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{xy}{\sigma^2}\right)^k \end{aligned}$$

So let

$$\phi(x) := e^{-x^2/(2\sigma^2)} \left(1, \frac{x}{\sigma}, \frac{1}{\sqrt{2!}} \left(\frac{x}{\sigma}\right)^2, \frac{1}{\sqrt{3!}} \left(\frac{x}{\sigma}\right)^3, \ldots\right).$$

Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi\colon \mathbb{R} \to \mathbb{R}^\infty$.

What ϕ has $\phi(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$\begin{aligned} e^{-(x-y)^2/(2\sigma^2)} &= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2} \\ &= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{xy}{\sigma^2}\right)^k \end{aligned}$$

So let

$$\phi(x) := e^{-x^2/(2\sigma^2)} \left(1, \frac{x}{\sigma}, \frac{1}{\sqrt{2!}}\left(\frac{x}{\sigma}\right)^2, \frac{1}{\sqrt{3!}}\left(\frac{x}{\sigma}\right)^3, \ldots\right).$$

How to handle $d > 1$?

Gaussian kernel feature expansion

First consider $d = 1$, meaning $\phi\colon \mathbb{R} \to \mathbb{R}^\infty$.

What ϕ has $\boldsymbol{\phi}(x)\phi(y) = e^{-(x-y)^2/(2\sigma^2)}$?

Reverse engineer using Taylor expansion:

$$\begin{aligned} e^{-(x-y)^2/(2\sigma^2)} &= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot e^{xy/\sigma^2} \\ &= e^{-x^2/(2\sigma^2)} \cdot e^{-y^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{xy}{\sigma^2} \right)^k \end{aligned}$$

So let

$$\boldsymbol{\phi}(x) := e^{-x^2/(2\sigma^2)} \left(1, \frac{x}{\sigma}, \frac{1}{\sqrt{2!}} \left(\frac{x}{\sigma} \right)^2, \frac{1}{\sqrt{3!}} \left(\frac{x}{\sigma} \right)^3, \ldots \right).$$

How to handle $d > 1$?

$$\begin{aligned} e^{-\|\boldsymbol{x}-\boldsymbol{y}\|^2/(2\sigma^2)} &= e^{-\|\boldsymbol{x}\|^2/(2\sigma^2)} \cdot e^{-\|\boldsymbol{y}\|^2/(2\sigma^2)} \cdot e^{\boldsymbol{x}^\mathsf{T}\boldsymbol{y}/\sigma^2} \\ &= e^{-\|\boldsymbol{x}\|^2/(2\sigma^2)} \cdot e^{-\|\boldsymbol{y}\|^2/(2\sigma^2)} \cdot \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\boldsymbol{x}^\mathsf{T}\boldsymbol{y}}{\sigma^2} \right)^k . \end{aligned}$$

Kernel example: products of all subsets of coordinates

Consider $\phi\colon \mathbb{R}^d \to \mathbb{R}^{2^d}$, where

$$\phi(\boldsymbol{x}) \;=\; \left(\prod_{i \in S} x_i\right)_{S \subseteq \{1,2,\ldots,d\}}$$

Time $\mathcal{O}(2^d)$ just to write down.
Kernel evaluation takes time $\mathcal{O}(d)$:

$$\phi(\boldsymbol{x})^{\mathsf{T}}\phi(\boldsymbol{x}') = \prod_{i=1}^{d}(1 + x_i x_i').$$

Other kernels

Suppose k_1 and k_2 are positive definite kernel functions.

Other kernels

Suppose k_1 and k_2 are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?

Other kernels

Suppose k_1 and k_2 are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?
2. $k(x, y) := c \cdot k_1(x, y)$ (for $c \geq 0$) define a positive definite kernel?

Other kernels

Suppose k_1 and k_2 are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?
2. $k(x, y) := c \cdot k_1(x, y)$ (for $c \geq 0$) define a positive definite kernel?
3. $k(x, y) := k_1(x, y) \cdot k_2(x, y)$ define a positive definite kernel?

Other kernels

Suppose k_1 and k_2 are positive definite kernel functions.

1. $k(x, y) := k_1(x, y) + k_2(x, y)$ define a positive definite kernel?
2. $k(x, y) := c \cdot k_1(x, y)$ (for $c \geq 0$) define a positive definite kernel?
3. $k(x, y) := k_1(x, y) \cdot k_2(x, y)$ define a positive definite kernel?

Another approach: random features $k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w}} F(\boldsymbol{w}, \boldsymbol{x}')^\top F(\boldsymbol{w}, \boldsymbol{x}')$ for some F; we will revisit this with deep networks and the neural tangent kernel (NTK).

Kernel ridge regression

Kernel ridge regression:

$$\min_{\boldsymbol{w}} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2.$$

Solution:

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda n\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

Linear algebra fact:

$$\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda n\boldsymbol{I})^{-1}\boldsymbol{y}.$$

Therefore predict with

$$\boldsymbol{x} \mapsto \boldsymbol{x}^\top\hat{\boldsymbol{w}} = (\boldsymbol{X}\boldsymbol{x})^\top(\boldsymbol{X}\boldsymbol{X}^\top + \lambda n\boldsymbol{I})^{-1}\boldsymbol{y} = \sum_{i=1}^n (\boldsymbol{x}_i^\top\boldsymbol{x})^\top\left[(\boldsymbol{X}\boldsymbol{X}^\top + \lambda n\boldsymbol{I})^{-1}\boldsymbol{y}\right]_i.$$

Kernel approach:

- Compute $\boldsymbol{\alpha} := (G + \lambda n\boldsymbol{I})^{-1}$, where $\boldsymbol{G} \in \mathbb{R}^n$ is the Gram matrix: $G_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
- Predict with $\boldsymbol{x} \mapsto \sum_{i=1}^n \alpha_i\boldsymbol{x}_i^\top\boldsymbol{x}$.

Multiclass SVM

There are a few ways;
one is to use one-against-all as in lecture.

Many researchers have proposed various multiclass SVM methods,
but some of them can be shown to fail in trivial cases.

New two-forms explanation idea **INCOMPLETE**

- All that matters is the direction; so for each direction which is a strict separator (meaning $\min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v} > 0$), we will pick one vector, and then compare the various directions. The final step is to write this as a convex (or concave) program.
- First approach: in each strict separation direction, normalize it so that $\|\boldsymbol{v}\| = 1$. If $\|\boldsymbol{v}\| = 1$, then the margin on the training set is $\min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v}$. We want the maximum amongst all these directions, which is $\max_{\|\boldsymbol{v}\|=1, \min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v} > 0} \min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v}$. Firstly we can drop the strict separation constraint since the objective enforces it. The first constraint however is not a convex set, so we don't have a concave program yet, but note that we have the same solutions if we instead write $\max_{\|\boldsymbol{v}\| \leq 1} \min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v}$. To see this, note that a vector $\boldsymbol{u}$ with $\|\boldsymbol{u}\| < 1$ can not be a maximizer, since $\boldsymbol{u}/\|\boldsymbol{u}\|$ achieves a $1/\|\boldsymbol{u}\| > 1$ factor larger marginm thus the solution to both problems lies on the boundary.
- Second approach: in each strict separation direction, normalize it so that $\min_i y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v} = 1$. Since the distance to the separator is $y \boldsymbol{x}^\mathsf{T} \boldsymbol{v} / \|\boldsymbol{v}\|$, then for a minimizing example k with $y_k x_k^\mathsf{T} \boldsymbol{v} = 1$, it follows that the margin is $y_k x_k^\mathsf{T} \boldsymbol{v} / \|\boldsymbol{v}\| = 1/\|\boldsymbol{v}\|$, and therefore the overall margin after this normalization is $1/\|\boldsymbol{v}\|$. We want to maximize the margin, so we could maximize $1/\|\boldsymbol{v}\|$, but this is a little awkward for optimization algorithms, so we instead minimize the reciprocal, $\|\boldsymbol{v}\|$. This too is a little awkward, so we square it, which preserves the optima, and minimize $\|\boldsymbol{v}\|^2/2$.
- Lastly, we can't write the program as $\min\{\|\boldsymbol{v}\|^2/2 : \forall i,\, y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{v} > 0\}$, because the infimal value is 0, which is not attained despite the set being bounded, so this is not a well-posed formulation for the maximum margin direction.

Supplemental reading

- Shalev-Shwartz/Ben-David: chapter 15.
- Murphy: chapter 14.