

EE6550 Machine Learning

Lecture Three – Support Vector Machines

Chung-Chin Lu

Department of Electrical Engineering

National Tsing Hua University

March 2, 2017

Binary Classification Problem

- $\mathcal{X} \subseteq \mathbb{R}^N$: the input space.
- $\mathcal{Y}' = \mathcal{Y} = \{-1, +1\}$: the output, label space with loss function $L(y', y) = 1_{y' \neq y}$.
- c : a fixed but unknown target concept in the concept class \mathcal{C} .
- \mathcal{H} : the hypothesis set.
- $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$: a sample of m items, drawn i.i.d. from the input space according to P , with labels $(c(\mathbf{x}_1), \dots, c(\mathbf{x}_m))$.
- **Problem:** find a hypothesis (binary classifier)
 $h : \mathcal{X} \rightarrow \{-1, +1\}$ in \mathcal{H} with small generalization error

$$R(h) = E[1_{h(\mathbf{x}) \neq c(\mathbf{x})}] = P(h(\mathbf{x}) \neq c(\mathbf{x})).$$

Linear Binary Classifiers

- Occam's razor principle: hypothesis sets with smaller complexity – e.g., smaller VC-dimension or Rademacher complexity– provide better learning guarantees, when everything else being equal.
- A natural hypothesis set with relatively small complexity is that of linear classifiers, or halfspaces (represented by their boundary hyperplanes), which can be defined as follows:

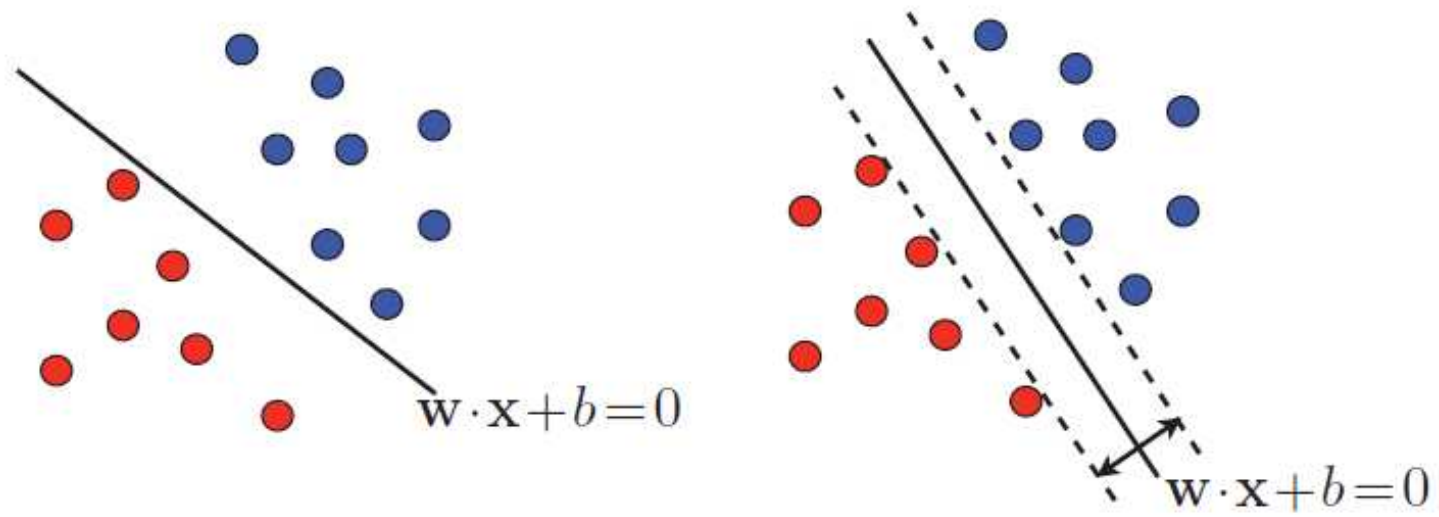
$$\mathcal{H} = \{\mathbf{x} \mapsto \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b) \mid \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}.$$

The Contents of This Lecture

- Support vector machines - separable case.
- Support vector machines - general case.
- Margin guarantees.

Linearly Separable Labeled Training Samples

- $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$: a labeled training sample of m items, drawn i.i.d. from the input space according to P , with labels $(c(\mathbf{x}_1), \dots, c(\mathbf{x}_m))$.
- **Assumption:** there is a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ which perfectly separates the training sample into two populations of positively and negatively labeled points.
- Existence of one perfectly separating hyperplane implies that of infinitely many such separating hyperplanes.
- Which hyperplane should a learning algorithm select?



Two possible separating hyperplanes.

SVM - Maximum-Margin Hyperplane

- $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$: a **linearly separable** labeled training sample of m items, drawn i.i.d. from the input space according to P , with labels $(c(\mathbf{x}_1), \dots, c(\mathbf{x}_m))$.
- $H : \mathbf{w} \cdot \mathbf{x} + b = 0$ with $c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) > 0, 1 \leq i \leq m$: a perfectly separating hyperplane for S .
- Geometric margin of a perfectly separating hyperplane (\mathbf{w}, b) w.r.t. S :

$$\rho = \min_{1 \leq i \leq m} \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|}.$$

- The SVM algorithm will return a hyperplane with the maximum margin, or distance to the closest points, which is

known as the maximum-margin hyperplane,

$$(\mathbf{w}, b)^{SVM} = \arg \max_{\substack{(\mathbf{w}, b): c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) > 0, 1 \leq i \leq m \\ \mathbf{w} \neq \mathbf{0}}} \min_{1 \leq i \leq m} \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|}.$$

Canonical Representation

The canonical representation of a perfectly separating hyperplane to a linearly separable labeled training sample $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$ is an affine equation for the hyperplane

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

such that

$$\min_{1 \leq i \leq m} |\mathbf{w} \cdot \mathbf{x}_i + b| = 1.$$

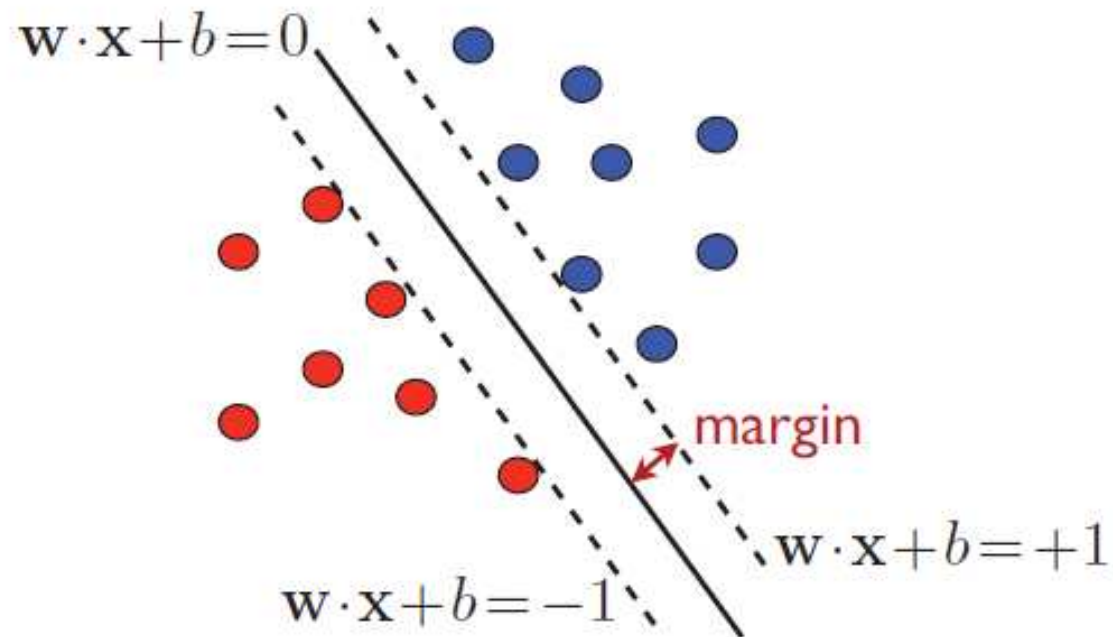
The geometric margin of a canonically represented perfectly separating hyperplane to S is

$$\rho = \min_{1 \leq i \leq m} \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}.$$

- If a maximum-margin hyperplane is canonically represented as $\mathbf{w} \cdot \mathbf{x} + b = 0$, then the two hyperplanes

$$\mathbf{w} \cdot \mathbf{x} + b = \pm 1$$

are called marginal hyperplanes.



Margin Maximization Problem

$$\begin{aligned}
 & \rho_{\max} \\
 = & \max_{\substack{(\mathbf{w}, b): c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) > 0, 1 \leq i \leq m \\ \mathbf{w} \neq \mathbf{0}}} \min_{1 \leq i \leq m} \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|} \\
 = & \max_{\substack{(\mathbf{w}, b): c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) > 0, 1 \leq i \leq m \\ \mathbf{w} \neq \mathbf{0}, \min_{1 \leq i \leq m} |\mathbf{w} \cdot \mathbf{x}_i + b| = 1}} \min_{1 \leq i \leq m} \frac{\mathbf{w} \cdot \mathbf{x}_i + b}{\|\mathbf{w}\|} \quad (\text{scaling invariance}) \\
 = & \max_{\substack{(\mathbf{w}, b): c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) > 0, 1 \leq i \leq m \\ \mathbf{w} \neq \mathbf{0}, \min_{1 \leq i \leq m} |\mathbf{w} \cdot \mathbf{x}_i + b| = 1}} \frac{1}{\|\mathbf{w}\|} \\
 = & \max_{\substack{(\mathbf{w}, b): c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, 1 \leq i \leq m \\ \mathbf{w} \neq \mathbf{0}}} \frac{1}{\|\mathbf{w}\|} \quad (\text{minimum reached}).
 \end{aligned}$$

- **Assumption:** the sample S is not trivially labeled, i.e., the points in the sample S are neither all positively labeled nor all negatively labeled.

In this case, we have

$$\rho_{\max} = \max_{(\mathbf{w}, b): c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, 1 \leq i \leq m} \frac{1}{\|\mathbf{w}\|}.$$

The Primal Problem for SVM - Separable Case

$$\text{Minimize} \quad F(\mathbf{w}, b) = \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{Subject to} \quad 1 - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) \leq 0, i = 1, \dots, m$$

$$(\mathbf{w}, b) \in \mathbb{R}^N \times \mathbb{R}.$$

- A quadratic programming (QP) problem.

Kuhn-Tucker Necessary Conditions for Local Minimal Solutions

Consider a nonlinear programming problem with equality constraints as well as inequality constraints, defined as

$$\begin{array}{ll}\text{Minimize} & f(\mathbf{x}) \\ \text{Subject to} & g_i(\mathbf{x}) \leq 0, i = 1, \dots, m \\ & h_j(\mathbf{x}) = 0, j = 1, \dots, l \\ & \mathbf{x} \in X,\end{array}$$

where X is a nonempty set in \mathbb{R}^N . Assume that

- $\bar{\mathbf{x}}$: a feasible solution, i.e., a point in X satisfying all equality constraints as well as inequality constraints;
- $I = \{i | g_i(\bar{\mathbf{x}}) = 0\}$;
- f and $g_i, i \in I$: differentiable at $\bar{\mathbf{x}}$;

- $g_i, i \notin I$: continuous at $\bar{\mathbf{x}}$;
- $h_j, j = 1, \dots, l$: continuously differentiable at $\bar{\mathbf{x}}$;
- $\nabla g_i(\bar{\mathbf{x}})$ for $i \in I$ and $\nabla h_j(\bar{\mathbf{x}})$ for $j = 1, \dots, l$ are linearly independent.

If $\bar{\mathbf{x}}$ is a **local minimal solution**, then there exist scalars λ_i for all $i \in I$ and μ_j for $j = 1, \dots, l$ such that

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{j=1}^l \mu_j \nabla h_j(\bar{\mathbf{x}}) = \mathbf{0}$$

$$\lambda_i \geq 0 \text{ for all } i \in I$$

In addition, if $g_i, i \notin I$, are also differentiable at $\bar{\mathbf{x}}$, then an equivalent form can be written as

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i=1}^m \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{j=1}^l \mu_j \nabla h_j(\bar{\mathbf{x}}) = \mathbf{0}$$

$$\lambda_i g_i(\bar{\mathbf{x}}) = 0 \text{ for all } i = 1, \dots, m$$

$$\lambda_i \geq 0 \text{ for all } i = 1, \dots, m.$$

- The scalars $\lambda_1, \dots, \lambda_m$ and μ_1, \dots, μ_l are called Lagrangian multipliers.
- The conditions $\lambda_i g_i(\bar{\mathbf{x}}) = 0, i = 1, \dots, m$, are called complementary slackness conditions.

Various Convexity and Concavity Concepts

- $S \subseteq X$: the feasible region of the nonlinear programming problem, defined as

$$S \triangleq \{\mathbf{x} \in X \mid g_i(\mathbf{x}) \leq 0, i = 1, \dots, m, \text{ and } h_j(\mathbf{x}) = 0, j = 1, \dots, l\}.$$

- A function u is said to be **pseudoconvex** at a feasible solution $\hat{\mathbf{x}} \in S$ if it is differentiable at $\hat{\mathbf{x}}$ and $\nabla u(\hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}}) \geq 0$ for $\mathbf{x} \in S$ implies that $u(\mathbf{x}) \geq u(\hat{\mathbf{x}})$.
- A function u is said to be **pseudoconcave** if $-u$ is pseudoconvex.
- A function u is said to be **quasiconvex** at a feasible solution $\hat{\mathbf{x}} \in S$ if u is defined in a convex set containing S and

$$u(\lambda \mathbf{x} + (1 - \lambda)\hat{\mathbf{x}}) \leq \max\{u(\mathbf{x}), u(\hat{\mathbf{x}})\}$$

for all $\lambda \in (0, 1)$ and all $\mathbf{x} \in S$.

- A function u is said to be **quasiconcave** if $-u$ is quasiconvex.
- A real-valued function u is said to be **convex** at a feasible solution $\hat{\mathbf{x}} \in S$ if u is defined in a convex set containing S and

$$u(\lambda \mathbf{x} + (1 - \lambda)\hat{\mathbf{x}}) \leq \lambda u(\mathbf{x}) + (1 - \lambda)u(\hat{\mathbf{x}})$$

for all $\lambda \in (0, 1)$ and all $\mathbf{x} \in S$.

- A function u is said to be **concave** if $-u$ is convex.
- If a real-valued function u is both convex and differentiable at a feasible solution $\hat{\mathbf{x}} \in S$, then it is pseudoconvex at $\hat{\mathbf{x}}$.
- If a real-valued function u is convex at a feasible solution $\hat{\mathbf{x}} \in S$, then it is quasiconvex at $\hat{\mathbf{x}}$.

Kuhn-Tucker Sufficient Conditions for Global Minimum Solutions

Consider a nonlinear programming problem with inequality as well as equality constraints, defined as

$$\begin{array}{ll}\text{Minimize} & f(\mathbf{x}) \\ \text{Subject to} & g_i(\mathbf{x}) \leq 0, i = 1, \dots, m \\ & h_j(\mathbf{x}) = 0, j = 1, \dots, l \\ & \mathbf{x} \in X,\end{array}$$

where X is a nonempty set in \mathbb{R}^N . Assume that

- $\bar{\mathbf{x}}$: a feasible solution;
- $I = \{i | g_i(\bar{\mathbf{x}}) = 0\}$.

Assume that $f, g_i, i \in I$ and $h_j, j = 1, \dots, k$ are differentiable at $\bar{\mathbf{x}}$ such that the Kuhn-Tucker necessary conditions hold true at $\bar{\mathbf{x}}$, i.e.,

there exist scalars $\lambda_i \geq 0, i \in I$, and $\mu_j \in \mathbb{R}, j = 1, 2, \dots, l$, such that

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{j=1}^l \mu_j \nabla h_j(\bar{\mathbf{x}}) = \mathbf{0}.$$

In addition, if $g_i, i \notin I$, are also differentiable at $\bar{\mathbf{x}}$, then an equivalent form of the Kuhn-Tucker necessary conditions can be written as

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i=1}^m \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{j=1}^l \mu_j \nabla h_j(\bar{\mathbf{x}}) = \mathbf{0}$$

$$\lambda_i g_i(\bar{\mathbf{x}}) = 0 \text{ for all } i = 1, \dots, m$$

$$\lambda_i \geq 0 \text{ for all } i = 1, \dots, m.$$

Also assume that

- $J = \{j | \mu_j > 0\}$ and $K = \{j | \mu_j < 0\}$;
- f : pseudoconvex at $\bar{\mathbf{x}}$;
- $g_i, i \in I$: quasiconvex at $\bar{\mathbf{x}}$;

- $h_j, j \in J$: quasiconvex at $\bar{\mathbf{x}}$;
- $h_j, j \in K$: quasiconcave at $\bar{\mathbf{x}}$.

Then $\bar{\mathbf{x}}$ is a **global minimum solution**.

Convex Function

Let

- X : a nonempty open convex subset of \mathbb{R}^n ;
- $f : X \rightarrow \mathbb{R}$: a twice differentiable function.

Then $f(\mathbf{x})$ is convex if and only if its Hessian matrix $\mathbf{H}(\mathbf{x})$ is positive semi-definite for $\mathbf{x} \in X$.

Qualification of the Primal Problem

- The object function $F(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2$ is infinitely differentiable and convex so that it is pseudoconvex at any feasible point.
- The inequality constraint functions $g_i(\mathbf{w}, b) = 1 - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b)$, $1 \leq i \leq m$, are affine functions so that they are infinitely differentiable and convex and then quasiconvex at any feasible point.
- $\nabla F = [\mathbf{w}^T 0]^T$, $\nabla g_i = -c(\mathbf{x}_i)[\mathbf{x}_i^T 1]^T$.
- The Kuhn-Tucker necessary conditions are:

$$\begin{aligned} \nabla F + \sum_{i=1}^m \lambda_i \nabla g_i &= \mathbf{0} \Leftrightarrow \mathbf{w} = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \mathbf{x}_i, 0 = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \\ \lambda_i g_i(\mathbf{w}, b) &= 0, \quad i = 1, 2, \dots, m \\ \lambda_i &\geq 0, \quad i = 1, 2, \dots, m. \end{aligned}$$

- Any feasible point (\mathbf{w}, b) which satisfies the Kuhn-Tucker necessary conditions in above is a global minimum solution.
- The weight vector \mathbf{w} solution of the SVM problem is a linear combination of the training set vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$.

Support Vectors

- Support vectors: any vector \mathbf{x}_i which appears in the linear combination $\mathbf{w} = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \mathbf{x}_i$, i.e., $\lambda_i \neq 0$.
- If $\lambda_i \neq 0$, we must have $c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$ by the complementary slackness conditions.
- Support vectors lie in the two marginal hyperplanes $\mathbf{w} \cdot \mathbf{x} + b = \pm 1$.

Remarks

- Support vectors fully define the maximum-margin hyperplane or SVM solution.
- Vectors in the sample not lying on the marginal hyperplanes do not affect the solution to the SVM problem.
- While the solution \mathbf{w} of the SVM problem is usually unique, the support vectors are not.

How to Determine Optimal Lagrangian Variables λ_i^{SVM} ?

- Once optimal Lagrangian variables λ_i^{SVM} are determined, we can compute

$$\mathbf{w}^{SVM} = \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i) \mathbf{x}_i$$

and for any support vector \mathbf{x}_j , we have

$$b^{SVM} = c(\mathbf{x}_j) - \mathbf{w}^{SVM} \cdot \mathbf{x}_j = c(\mathbf{x}_j) - \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i) (\mathbf{x}_i \cdot \mathbf{x}_j).$$

- We will use the Lagrangian dual problem to determine optimal λ_i^{SVM} .

Lagrangian Dual Function

- Primal problem:

$$\begin{aligned} \text{Minimize} \quad & f(\mathbf{x}) \\ \text{Subject to} \quad & g_i(\mathbf{x}) \leq 0, i = 1, \dots, m \\ & h_j(\mathbf{x}) = 0, j = 1, \dots, l \\ & \mathbf{x} \in X, \end{aligned}$$

where X is a nonempty set in \mathbb{R}^n .

- Lagrangian function: for all $\mathbf{x} \in X$, $\lambda \in \mathbb{R}^m$, and $\nu \in \mathbb{R}^k$,

$$L(\mathbf{x}, \lambda, \nu) \triangleq f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^k \nu_j h_j(\mathbf{x}).$$

- Lagrangian dual function: for all $\lambda \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^k$,

$$\theta(\lambda, \nu) \triangleq \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) = \inf_{\mathbf{x} \in X} \left(f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^k \nu_j h_j(\mathbf{x}) \right).$$

Global Minimum of a Convex Function

Let

- X : a nonempty open convex subset of \mathbb{R}^n ;
- $f : X \rightarrow \mathbb{R}$: a differentiable convex function.

Then $\bar{\mathbf{x}}$ is an optimal solution to the minimization of $f(\mathbf{x})$ subject to $\mathbf{x} \in X$ if and only if $\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$.

Lagrangian Dual Function for SVM - Separable Case

- $X = \mathbb{R}^N \times \mathbb{R}$: a nonempty open convex set.
- Lagrangian function: for all $\mathbf{w} \in \mathbb{R}^N$, $b \in \mathbb{R}$, and $\lambda \in \mathbb{R}^m$,

$$\begin{aligned}
 L(\mathbf{w}, b, \lambda) &= F(\mathbf{w}, b) + \sum_{i=1}^m \lambda_i g_i(\mathbf{w}, b) \\
 &= \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^m \lambda_i (1 - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b))
 \end{aligned}$$

- For any fixed $\lambda \in \mathbb{R}^m$, the gradient ∇L of the Lagrangian

function w.r.t. (\mathbf{w}, b) is

$$\begin{aligned}\nabla L &= \nabla F + \sum_{i=1}^m \lambda_i \nabla g_i \\ &= \begin{bmatrix} \mathbf{w} \\ 0 \end{bmatrix} - \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}\end{aligned}$$

and the Hessian matrix is

$$\mathbf{H} = \begin{bmatrix} I_{N \times N} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix}$$

which is positive semi-definite.

- For any fixed $\lambda \in \mathbb{R}^m$, the Lagrangian function is differentiable and convex over a non-empty open convex set X so that $(\hat{\mathbf{w}}, \hat{b})$ is an optimal solution to the minimization of $L(\mathbf{w}, b, \lambda)$ subject

to $(\mathbf{w}, b) \in X$ if and only if $\nabla L(\hat{\mathbf{w}}, \hat{b}, \lambda) = \mathbf{0}$ if and only if

$$\hat{\mathbf{w}} = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \mathbf{x}_i \quad \text{and} \quad 0 = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i).$$

- Note that for a fixed $\lambda \in \mathbb{R}^m$, $\sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \neq 0$ if and only if the infimum of the Lagrangian function $L(\mathbf{w}, b, \lambda)$ is $-\infty$.

- Lagrangian dual function: for all $\lambda \in \mathbb{R}^m$,

$$\begin{aligned}
 \theta(\lambda) &= \inf_{(\mathbf{w}, b) \in X} L(\mathbf{w}, b, \lambda) \\
 &= \begin{cases} \frac{1}{2} \|\hat{\mathbf{w}}\|^2 + \sum_{i=1}^m \lambda_i (1 - c(\mathbf{x}_i)(\hat{\mathbf{w}} \cdot \mathbf{x}_i + \hat{b})), \\ \hspace{15em} \text{if } \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) = 0, \\ -\infty, \text{ if } \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \neq 0 \end{cases} \\
 &= \begin{cases} \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j c(\mathbf{x}_i) c(\mathbf{x}_j) (\mathbf{x}_i \cdot \mathbf{x}_j), \\ \hspace{15em} \text{if } \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) = 0, \\ -\infty, \text{ if } \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \neq 0 \end{cases}
 \end{aligned}$$

Lagrangian Dual Problem

Maximize $\theta(\mathbf{u}, \mathbf{v})$

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

$\mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^k$

- Also referred to as the max-min dual problem.
- Given a primal problem, several Lagrangian dual problems can be devised, depending on which constraints are handled as $g_i(\mathbf{x}) \leq 0$ and $h_j(\mathbf{x}) = 0$ and which constraints are treated by the set X .

Lagrangian Dual Problem for SVM - Separable Case

$$\text{Maximize} \quad \theta(\lambda) = \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j c(\mathbf{x}_i) c(\mathbf{x}_j) (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{Subject to} \quad \lambda_i \geq 0, i = 1, \dots, m$$

$$\sum_{i=1}^m \lambda_i c(\mathbf{x}_i) = 0$$

$$\lambda \in \mathbb{R}^m$$

- A quadratic programming (QP) problem.

Kuhn-Tucker Necessary Conditions for Local Maximal Solutions

Consider a nonlinear programming problem with equality constraints as well as inequality constraints, defined as

$$\begin{array}{ll}\text{Maximize} & f(\mathbf{x}) \\ \text{Subject to} & g_i(\mathbf{x}) \geq 0, i = 1, \dots, m \\ & h_j(\mathbf{x}) = 0, j = 1, \dots, l \\ & \mathbf{x} \in X,\end{array}$$

where X is a nonempty set in \mathbb{R}^N . Let

- $\bar{\mathbf{x}}$: a feasible solution, i.e., a point in X satisfying all equality constraints as well as inequality constraints;
- $I = \{i | g_i(\bar{\mathbf{x}}) = 0\}$;
- f and $g_i, i \in I$: differentiable at $\bar{\mathbf{x}}$;

- $g_i, i \notin I$: continuous at $\bar{\mathbf{x}}$;
- $h_j, j = 1, \dots, l$: continuously differentiable at $\bar{\mathbf{x}}$;
- $\nabla g_i(\bar{\mathbf{x}})$ for $i \in I$ and $\nabla h_j(\bar{\mathbf{x}})$ for $j = 1, \dots, l$ are linearly independent.

If $\bar{\mathbf{x}}$ is a **local optimal solution**, then there exist scalars λ_i for all $i \in I$ and μ_j for $j = 1, \dots, l$ such that

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{j=1}^l \mu_j \nabla h_j(\bar{\mathbf{x}}) = \mathbf{0}$$

$$\lambda_i \geq 0 \text{ for all } i \in I$$

In addition, if $g_i, i \notin I$, are also differentiable at $\bar{\mathbf{x}}$, then an equivalent form can be written as

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i=1}^m \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{j=1}^l \mu_j \nabla h_j(\bar{\mathbf{x}}) = \mathbf{0}$$

$$\lambda_i g_i(\bar{\mathbf{x}}) = 0 \text{ for all } i = 1, \dots, m$$

$$\lambda_i \geq 0 \text{ for all } i = 1, \dots, m.$$

Kuhn-Tucker Sufficient Conditions for Global Maximum Solutions

Consider a nonlinear programming problem with inequality as well as equality constraints, defined as

$$\begin{aligned} &\text{Maximize} && f(\mathbf{x}) \\ &\text{Subject to} && g_i(\mathbf{x}) \geq 0, i = 1, \dots, m \\ & && h_j(\mathbf{x}) = 0, j = 1, \dots, l \\ & && \mathbf{x} \in X, \end{aligned}$$

where X is a nonempty set in \mathbb{R}^N . Let

- $\bar{\mathbf{x}}$: a feasible solution;
- $I = \{i | g_i(\bar{\mathbf{x}}) = 0\}$.

Assume that $f, g_i, i \in I$ and $h_j, j = 1, \dots, k$ are differentiable at $\bar{\mathbf{x}}$ such that the Kuhn-Tucker necessary conditions hold true at $\bar{\mathbf{x}}$, i.e.,

there exist scalars $\lambda_i \geq 0, i \in I$, and $\mu_j \in \mathbb{R}, j = 1, 2, \dots, l$, such that

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i \in I} \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{j=1}^l \mu_j \nabla h_j(\bar{\mathbf{x}}) = \mathbf{0}.$$

In addition, if $g_i, i \notin I$, are also differentiable at $\bar{\mathbf{x}}$, then an equivalent form of the Kuhn-Tucker necessary conditions can be written as

$$\nabla f(\bar{\mathbf{x}}) + \sum_{i=1}^m \lambda_i \nabla g_i(\bar{\mathbf{x}}) + \sum_{j=1}^l \mu_j \nabla h_j(\bar{\mathbf{x}}) = \mathbf{0}$$

$$\lambda_i g_i(\bar{\mathbf{x}}) = 0 \text{ for all } i = 1, \dots, m$$

$$\lambda_i \geq 0 \text{ for all } i = 1, \dots, m.$$

Also assume that

- $J = \{j | \mu_j > 0\}$ and $K = \{j | \mu_j < 0\}$;
- f : pseudoconcave at $\bar{\mathbf{x}}$;
- $g_i, i \in I$: quasiconcave at $\bar{\mathbf{x}}$;

- $h_j, j \in J$: quasiconcave at $\bar{\mathbf{x}}$;
- $h_j, j \in K$: quasiconvex at $\bar{\mathbf{x}}$.

Then $\bar{\mathbf{x}}$ is a **global maximum solution**.

Concave Function

Let

- X : a nonempty open convex subset of \mathbb{R}^n ;
- $f : X \rightarrow \mathbb{R}$: a twice differentiable function.

Then $f(\mathbf{x})$ is concave if and only if its Hessian matrix $\mathbf{H}(\mathbf{x})$ is negative semi-definite for $\mathbf{x} \in X$.

Qualification of the Dual Problem

- The object function

$$\theta(\lambda) = \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j c(\mathbf{x}_i) c(\mathbf{x}_j) (\mathbf{x}_i \cdot \mathbf{x}_j)$$

is infinitely differentiable and concave so that it is pseudoconcave at any feasible point.

- The inequality constraint functions $g_i(\lambda) = \lambda_i, 1 \leq i \leq m$, and the equality constraint function $h(\lambda) = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i)$ are affine functions so that they are infinitely differentiable, concave and convex and then quasiconcave and quasiconvex at any feasible point.

- $\nabla\theta(\lambda) = \mathbf{1} - \mathbf{A}\lambda$, where \mathbf{A} is the Gram matrix of the vectors $c(\mathbf{x}_i)\mathbf{x}_i, i = 1, 2, \dots, m$,

$$\begin{aligned} \mathbf{A} &= [c(\mathbf{x}_i)\mathbf{x}_i \cdot c(\mathbf{x}_j)\mathbf{x}_j] \\ &= \begin{bmatrix} c(\mathbf{x}_1)\mathbf{x}_1 \cdot c(\mathbf{x}_1)\mathbf{x}_1 & \cdots & c(\mathbf{x}_1)\mathbf{x}_1 \cdot c(\mathbf{x}_m)\mathbf{x}_m \\ c(\mathbf{x}_2)\mathbf{x}_2 \cdot c(\mathbf{x}_1)\mathbf{x}_1 & \cdots & c(\mathbf{x}_2)\mathbf{x}_2 \cdot c(\mathbf{x}_m)\mathbf{x}_m \\ \vdots & \ddots & \vdots \\ c(\mathbf{x}_m)\mathbf{x}_m \cdot c(\mathbf{x}_1)\mathbf{x}_1 & \cdots & c(\mathbf{x}_m)\mathbf{x}_m \cdot c(\mathbf{x}_m)\mathbf{x}_m \end{bmatrix} \end{aligned}$$

- $\nabla g_i(\lambda) = \mathbf{e}_i, i = 1, 2, \dots, m$, and $\nabla h(\lambda) = [c(\mathbf{x}_1), \dots, c(\mathbf{x}_m)]^T$.

- The Kuhn-Tucker necessary conditions are:

$$\nabla\theta + \sum_{i=1}^m u_i \nabla g_i + v \nabla h = \mathbf{0} \Leftrightarrow \mathbf{A}\lambda = \mathbf{1} + \mathbf{u} + v \begin{bmatrix} c(\mathbf{x}_1) \\ \vdots \\ c(\mathbf{x}_m) \end{bmatrix}$$

$$u_i \lambda_i = 0, \quad i = 1, 2, \dots, m$$

$$u_i \geq 0, \quad i = 1, 2, \dots, m.$$

- Any feasible point λ which satisfies the Kuhn-Tucker necessary conditions in above is a global maximum solution.

Weak Duality Theorem

Assume that

- \mathbf{x} : a feasible solution to the primal problem P, i.e.,
 $\mathbf{x} \in X, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}$;
- (\mathbf{u}, \mathbf{v}) : a feasible solution to the dual problem D, i.e., $\mathbf{u} \geq \mathbf{0}$.

Then we have

$$f(\mathbf{x}) \geq \theta(\mathbf{u}, \mathbf{v}).$$

Proof. Since $\mathbf{x} \in X$,

$$\begin{aligned}\theta(\mathbf{u}, \mathbf{v}) &= \inf_{\mathbf{y} \in X} (f(\mathbf{y}) + \mathbf{u}^T \mathbf{g}(\mathbf{y}) + \mathbf{v}^T \mathbf{h}(\mathbf{y})) \\ &\leq f(\mathbf{x}) + \mathbf{u}^T \mathbf{g}(\mathbf{x}) + \mathbf{v}^T \mathbf{h}(\mathbf{x}) \\ &= f(\mathbf{x}),\end{aligned}$$

since $\mathbf{u} \geq \mathbf{0}, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ and $\mathbf{h}(\mathbf{x}) = \mathbf{0}$.

□

Corollaries of the Weak Duality Theorem

- $\inf\{f(\mathbf{x}) \mid \mathbf{x} \in X, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}\} \geq \sup\{\theta(\mathbf{u}, \mathbf{v}) \mid \mathbf{u} \geq \mathbf{0}\}.$
- If $f(\bar{\mathbf{x}}) \leq \theta(\bar{\mathbf{u}}, \bar{\mathbf{v}})$, where $\bar{\mathbf{u}} \geq \mathbf{0}$ and $\bar{\mathbf{x}} \in \{\mathbf{x} \in X, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$, then $\bar{\mathbf{x}}$ and $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$ solve the primal and dual problems respectively.
- If $\inf\{f(\mathbf{x}) \mid \mathbf{x} \in X, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}\} = -\infty$, then $\theta(\mathbf{u}, \mathbf{v}) = -\infty$ for all $\mathbf{u} \geq \mathbf{0}, \mathbf{v} \in \mathbb{R}^k$.
- If $\sup\{\theta(\mathbf{u}, \mathbf{v}) \mid \mathbf{u} \geq \mathbf{0}\} = +\infty$, then the primal problem has no feasible solution.

Strong Duality Theorem

Assume that

- X : a nonempty convex set in \mathbb{R}^n ;
- $f : X \rightarrow \mathbb{R}$ and $\mathbf{g} : X \rightarrow \mathbb{R}^m$: convex functions on X ;
- $\mathbf{h} : \mathbb{R}^n \rightarrow \mathbb{R}^k$: an affine function, i.e., $\mathbf{h}(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$ for some $k \times n$ matrix A and some vector \mathbf{b} in \mathbb{R}^k ;
 - Without loss of generality, we may assume that the matrix A has full rank.
- $\mathbf{0} \in \text{int}\mathbf{h}(X)$, where $\mathbf{h}(X) = \{\mathbf{h}(\mathbf{x}) : \mathbf{x} \in X\}$;
- there exists an $\mathbf{x}' \in X$ such that $\mathbf{g}(\mathbf{x}') < \mathbf{0}$ and $\mathbf{h}(\mathbf{x}') = \mathbf{0}$.

Then we have

$$\inf\{f(\mathbf{x}) : \mathbf{x} \in X, \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}\} = \sup\{\theta(\mathbf{u}, \mathbf{v}) : \mathbf{u} \geq \mathbf{0}\}.$$

Furthermore, if the inf is finite, then $\sup\{\theta(\mathbf{u}, \mathbf{v}) \mid \mathbf{u} \geq \mathbf{0}\}$ is achieved at some $(\bar{\mathbf{u}}, \bar{\mathbf{v}})$ with $\bar{\mathbf{u}} \geq \mathbf{0}$. If the inf is achieved at $\bar{\mathbf{x}}$, then $\sum_{i=1}^m u_i g_i(\bar{\mathbf{x}}) = 0$.

Justification of Strong Duality for SVM

- $X = \mathbb{R}^N \times \mathbb{R}$: a non-empty convex set.
- $F(\mathbf{w}, b) = \frac{1}{2} \|\mathbf{w}\|^2$: a convex function on X .
- $g_i(\mathbf{w}, b) = 1 - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b)$, $1 \leq i \leq m$: affine functions so that they are convex functions on X .
- There exists an $(\mathbf{w}', b') \in X$ such that $\mathbf{g}(\mathbf{w}', b') < \mathbf{0}$.

Then we have

$$\inf\{F(\mathbf{w}, b) : (\mathbf{w}, b) \in X, \mathbf{g}(\mathbf{w}, b) \leq \mathbf{0}\} = \sup\{\theta(\lambda) : \lambda \geq \mathbf{0}\}.$$

- For a linearly separable labeled training sample, the inf is finite and can be achieved at some feasible point $(\mathbf{w}^{SVM}, b^{SVM})$.
Then $\sup\{\theta(\lambda) \mid \lambda \geq \mathbf{0}\}$ is achieved at some $\lambda^{SVM} \geq \mathbf{0}$.
- The primal and dual problems are equivalent.

The SVM Algorithm - Separable Case

- $S = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m)$: a linearly separable labeled training sample of size m with labels $(c(\mathbf{x}_1), c(\mathbf{x}_2), \dots, c(\mathbf{x}_m))$.
- h_S^{SVM} : the hypothesis returned by SVM,

$$\begin{aligned} h_S^{SVM}(\mathbf{x}) &= \text{sgn}(\mathbf{w}^{SVM} \cdot \mathbf{x} + b^{SVM}) \\ &= \text{sgn}\left(\sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot \mathbf{x}) + b^{SVM}\right) \end{aligned}$$

- $b^{SVM} = c(\mathbf{x}_j) - \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot \mathbf{x}_j)$ for any support vector \mathbf{x}_j . Thus we have

$$h_S^{SVM}(\mathbf{x}) = \text{sgn}\left(c(\mathbf{x}_j) + \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot (\mathbf{x} - \mathbf{x}_j))\right)$$

for any support vector \mathbf{x}_j .

- The hypothesis solution h_S^{SVM} depends only on inner products between vectors and not directly on the vectors themselves.

The Maximum Margin ρ_{\max}

- $b^{SVM} = c(\mathbf{x}_j) - \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot \mathbf{x}_j)$ for any support vector \mathbf{x}_j . This implies

$$\begin{aligned} & \sum_{j=1}^m \lambda_j^{SVM} c(\mathbf{x}_j) b^{SVM} \\ &= \sum_{j=1}^m \lambda_j^{SVM} c(\mathbf{x}_j)^2 - \sum_{j=1}^m \lambda_j^{SVM} c(\mathbf{x}_j) \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot \mathbf{x}_j). \end{aligned}$$

- Since $\sum_{j=1}^m \lambda_i^{SVM} c(\mathbf{x}_j) = 0$ and $\mathbf{w}^{SVM} = \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i) \mathbf{x}_i$, we have

$$\sum_{j=1}^m \lambda_j^{SVM} = \|\mathbf{w}^{SVM}\|^2.$$

- $\rho_{\max}^2 = \frac{1}{\|\mathbf{w}^{SVM}\|^2} = \frac{1}{\sum_{j=1}^m \lambda_j^{SVM}}.$

The Contents of This Lecture

- Support vector machines - separable case.
- Support vector machines - general case.
- Margin guarantees.

Non-Linearly Separable Labeled Training Samples

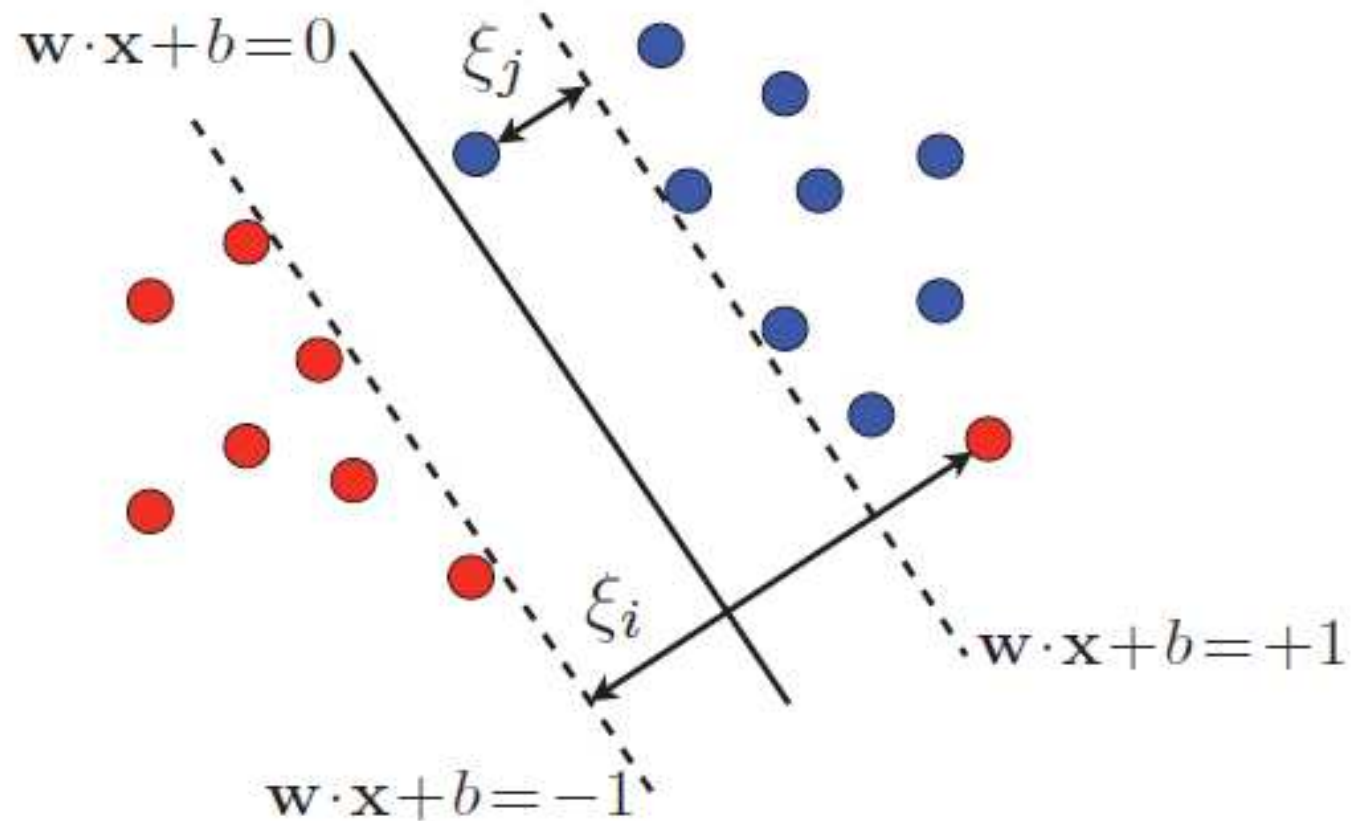
- $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$: a labeled training sample of m items, drawn i.i.d. from the input space according to P , with labels $(c(\mathbf{x}_1), \dots, c(\mathbf{x}_m))$.
- **Problem:** the training data S is often not linearly separable in practice, i.e., for any hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$, there exists $\mathbf{x}_i \in S$ such that

$$c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) \not\geq 1.$$

- **Idea:** relax inequality constraints using slack variables $\eta_i \geq 0$, $i = 1, 2, \dots, m$, such that

$$c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \eta_i.$$

- A slack variable η_i measures the amount by which vector \mathbf{x}_i violates the desired inequality $c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$.



Point \mathbf{x}_i is classified incorrectly and point \mathbf{x}_j is correctly classified, but with a margin less than 1.

Remarks

- Soft margin : $\rho = 1/\|\mathbf{w}\|$.
- For a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$, a vector \mathbf{x}_i with $\eta_i > 0$ can be viewed as an outlier.
- How should we select the hyperplane in the general, separable or non-separable, case?
- There are two conflicting objectives: on one hand, we wish to limit the total amount of slack due to outliers, which can be measured by $\sum_{i=1}^m \eta_i$ or $\sum_{i=1}^m \eta_i^p$ for some $p \geq 1$; on the other hand, we seek a hyperplane with a large soft margin, though a larger soft margin can lead to more outliers and thus larger amounts of slack.

The Primal Problem for SVM - General Case

$$\begin{aligned}
 &\text{Minimize} && F(\mathbf{w}, b, \eta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \eta_i \\
 &\text{Subject to} && 1 - \eta_i - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) \leq 0, i = 1, \dots, m \\
 &&& -\eta_i \leq 0, i = 1, \dots, m \\
 &&& (\mathbf{w}, b, \eta) \in \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^m.
 \end{aligned}$$

- A quadratic programming (QP) problem.

Remarks

- The parameter $C > 0$ determines the trade-off between margin-maximization (or minimization of $\|w\|^2$) and the minimization of the slack penalty $\sum_{i=1}^m \eta_i$.
- The parameter C is typically determined via n -fold cross-validation.

Qualification of the Primal Problem - General Case

- The object function $F(\mathbf{w}, b, \eta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \eta_i$ is infinitely differentiable and convex so that it is pseudoconvex at any feasible point.
- The inequality constraint functions $g_i(\mathbf{w}, b, \eta) = 1 - \eta_i - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b)$ and $h_i(\mathbf{w}, b, \eta) = -\eta_i$, $1 \leq i \leq m$, are affine functions so that they are infinitely differentiable and convex and then quasiconvex at any feasible point.

$$\bullet \quad \nabla F = \begin{bmatrix} \mathbf{w} \\ 0 \\ C\mathbf{1} \end{bmatrix}, \quad \nabla g_i = \begin{bmatrix} -c(\mathbf{x}_i)\mathbf{x}_i \\ -c(\mathbf{x}_i) \\ -\mathbf{e}_i \end{bmatrix}, \quad \text{and} \quad \nabla h_i = \begin{bmatrix} \mathbf{0} \\ 0 \\ -\mathbf{e}_i \end{bmatrix}.$$

- The Kuhn-Tucker necessary conditions are:

$$\begin{aligned} \nabla F + \sum_{i=1}^m \lambda_i \nabla g_i + \sum_{i=1}^m \mu_i \nabla h_i &= \mathbf{0} \\ \Leftrightarrow \mathbf{w} &= \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \mathbf{x}_i, 0 = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i), C = \lambda_i + \mu_i, i \in [1, m] \\ \lambda_i g_i(\mathbf{w}, b, \eta) &= 0, \quad i \in [1, m] \\ \mu_i \eta_i &= 0, \quad i \in [1, m] \\ \lambda_i, \mu_i &\geq 0, \quad i \in [1, m]. \end{aligned}$$

- Any feasible point (\mathbf{w}, b, η) which satisfies the Kuhn-Tucker necessary conditions in above is a global minimum solution.
- The weight vector \mathbf{w} solution of the general, separable or non-separable, SVM problem is also a linear combination of the training set vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$.

Support Vectors

- Support vectors: any vector \mathbf{x}_i which appears in the linear combination $\mathbf{w} = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \mathbf{x}_i$, i.e., $\lambda_i \neq 0$.
- If $\lambda_i \neq 0$, we must have $c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \eta_i$ by the complementary slackness conditions.
- If $\eta_i = 0$, the support vector \mathbf{x}_i lies in the marginal hyperplane $\mathbf{w} \cdot \mathbf{x} + b = c(\mathbf{x}_i)$.
- If $\eta_i > 0$, the support vector \mathbf{x}_i is an outlier. In this case, $\mu_i = 0$ and then $\lambda_i = C$.

Remarks

- Support vectors fully define the maximum-margin hyperplane or SVM solution.
- Support vectors \mathbf{x}_i are either outliers, in which case $\lambda_i = C$, or vectors lying on the marginal hyperplanes.
- Vectors in the sample neither outliers nor lying on the marginal hyperplanes do not affect the solution to the SVM problem.
- As in the separable case, note that while the solution \mathbf{w} of the SVM problem is usually unique, the support vectors are not.

How to Determine Optimal Lagrangian Variables λ_i^{SVM} ?

- Once optimal Lagrangian variables λ_i^{SVM} are determined, we can compute

$$\mathbf{w}^{SVM} = \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i) \mathbf{x}_i$$

and for any support vector \mathbf{x}_j lying on the marginal hyperplanes, we have

$$b^{SVM} = c(\mathbf{x}_j) - \mathbf{w}^{SVM} \cdot \mathbf{x}_j = c(\mathbf{x}_j) - \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i) (\mathbf{x}_i \cdot \mathbf{x}_j).$$

- We will use the Lagrangian dual problem to determine optimal λ_i^{SVM} .

Lagrangian Dual Function for SVM - General Case

- $X = \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^m$: a nonempty open convex set.
- Lagrangian function: for all $\mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \eta \in \mathbb{R}^m$, and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^m$,

$$\begin{aligned}
 & L(\mathbf{w}, b, \eta, \lambda, \mu) \\
 = & F(\mathbf{w}, b, \eta) + \sum_{i=1}^m \lambda_i g_i(\mathbf{w}, b, \eta) + \sum_{i=1}^m \mu_i h_i(\mathbf{w}, b, \eta) \\
 = & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \eta_i + \sum_{i=1}^m \lambda_i (1 - \eta_i - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b)) \\
 & - \sum_{i=1}^m \mu_i \eta_i.
 \end{aligned}$$

- For any fixed $\lambda, \mu \in \mathbb{R}^m$, the gradient ∇L of the Lagrangian

function w.r.t. (\mathbf{w}, b, η) is

$$\begin{aligned}\nabla L &= \nabla F + \sum_{i=1}^m \lambda_i \nabla g_i + \sum_{i=1}^m \mu_i \nabla h_i \\ &= \begin{bmatrix} \mathbf{w} \\ 0 \\ C\mathbf{1} \end{bmatrix} - \sum_{i=1}^m \lambda_i \begin{bmatrix} c(\mathbf{x}_i)\mathbf{x}_i \\ c(\mathbf{x}_i) \\ \mathbf{e}_i \end{bmatrix} - \sum_{i=1}^m \mu_i \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{e}_i \end{bmatrix}\end{aligned}$$

and the Hessian matrix is

$$\mathbf{H} = \begin{bmatrix} I_{N \times N} & \mathbf{0}_{N \times (m+1)} \\ \mathbf{0}_{(m+1) \times N} & \mathbf{0}_{(m+1) \times (m+1)} \end{bmatrix}$$

which is positive semi-definite.

- For any fixed $\lambda, \mu \in \mathbb{R}^m$, the Lagrangian function is differentiable and convex over a non-empty open convex set X

so that $(\hat{\mathbf{w}}, \hat{b}, \hat{\eta})$ is an optimal solution to the minimization of $L(\mathbf{w}, b, \eta, \lambda, \mu)$ subject to $(\mathbf{w}, b, \eta) \in X$ if and only if $\nabla L(\hat{\mathbf{w}}, \hat{b}, \hat{\eta}, \lambda, \mu) = \mathbf{0}$ if and only if

$$\hat{\mathbf{w}} = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \mathbf{x}_i, \quad 0 = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i), \quad \text{and} \quad C = \lambda_i + \mu_i, \quad i \in [1, m].$$

- Note that for any fixed $\lambda, \mu \in \mathbb{R}^m$, $\sum_{i=1}^m \lambda_i c(\mathbf{x}_i) \neq 0$ or $C \neq \lambda_i + \mu_i$ for some $i \in [1, m]$ if and only if the infimum of the Lagrangian function $L(\mathbf{w}, b, \eta, \lambda, \mu)$ is $-\infty$.

- Lagrangian dual function: for any $\lambda, \mu \in \mathbb{R}^m$,

$$\begin{aligned}
& \theta(\lambda, \mu) \\
&= \inf_{(\mathbf{w}, b, \eta) \in X} L(\mathbf{w}, b, \eta, \lambda, \mu) \\
&= \begin{cases} \frac{1}{2} \|\hat{\mathbf{w}}\|^2 + C \sum_{i=1}^m \hat{\eta}_i + \sum_{i=1}^m \lambda_i (1 - \hat{\eta}_i - c(\mathbf{x}_i)(\hat{\mathbf{w}} \cdot \mathbf{x}_i + \hat{b})) \\ - \sum_{i=1}^m \mu_i \hat{\eta}_i, & \text{if } \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) = 0, C = \lambda_i + \mu_i, i \in [1, m], \\ -\infty, & \text{otherwise} \end{cases} \\
&= \begin{cases} \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j c(\mathbf{x}_i) c(\mathbf{x}_j) (\mathbf{x}_i \cdot \mathbf{x}_j), \\ \quad \text{if } \sum_{i=1}^m \lambda_i c(\mathbf{x}_i) = 0, C = \lambda_i + \mu_i, i \in [1, m], \\ -\infty, & \text{otherwise.} \end{cases}
\end{aligned}$$

Lagrangian Dual Problem for SVM - General Case

$$\text{Maximize} \quad \theta(\lambda, \mu) = \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j c(\mathbf{x}_i) c(\mathbf{x}_j) (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{Subject to} \quad \lambda_i, \mu_i \geq 0, i = 1, \dots, m$$

$$\lambda_i + \mu_i - C = 0, i = 1, \dots, m$$

$$\sum_{i=1}^m \lambda_i c(\mathbf{x}_i) = 0$$

$$(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^m$$

Or equivalently,

$$\text{Maximize} \quad \theta(\lambda) = \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j c(\mathbf{x}_i) c(\mathbf{x}_j) (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{Subject to} \quad \lambda_i \geq 0, i = 1, \dots, m$$

$$C - \lambda_i \geq 0, i = 1, \dots, m$$

$$\sum_{i=1}^m \lambda_i c(\mathbf{x}_i) = 0$$

$$\lambda \in \mathbb{R}^m$$

- A quadratic programming (QP) problem.

Qualification of the Dual Problem

- The object function

$$\theta(\lambda) = \sum_{i=1}^m \lambda_i - \frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j c(\mathbf{x}_i) c(\mathbf{x}_j) (\mathbf{x}_i \cdot \mathbf{x}_j)$$

is infinitely differentiable and concave so that it is pseudoconcave at any feasible point.

- The inequality constraint functions $g_i(\lambda) = \lambda_i, 1 \leq i \leq m$, $\tilde{g}_i(\lambda) = C - \lambda_i, 1 \leq i \leq m$, and the equality constraint function $h(\lambda) = \sum_{i=1}^m \lambda_i c(\mathbf{x}_i)$ are affine functions so that they are infinitely differentiable, concave and convex and then quasiconcave and quasiconvex at any feasible point.

- $\nabla\theta(\lambda) = \mathbf{1} - \mathbf{A}\lambda$, where $\mathbf{A} = [c(\mathbf{x}_i)\mathbf{x}_i \cdot c(\mathbf{x}_j)\mathbf{x}_j]$ is the Gram matrix of the vectors $c(\mathbf{x}_i)\mathbf{x}_i, i = 1, 2, \dots, m$.
- $\nabla g_i(\lambda) = \mathbf{e}_i, i = 1, 2, \dots, m, \nabla \tilde{g}_i(\lambda) = -\mathbf{e}_i, i = 1, 2, \dots, m$, and $\nabla h(\lambda) = [c(\mathbf{x}_1), \dots, c(\mathbf{x}_m)]^T$.
- The Kuhn-Tucker necessary conditions are:

$$\nabla\theta + \sum_{i=1}^m u_i \nabla g_i + \sum_{i=1}^m \tilde{u}_i \nabla \tilde{g}_i + v \nabla h = \mathbf{0}$$

$$\Leftrightarrow \mathbf{A}\lambda = \mathbf{1} + \mathbf{u} - \tilde{\mathbf{u}} + v \begin{bmatrix} c(\mathbf{x}_1) \\ \vdots \\ c(\mathbf{x}_m) \end{bmatrix}$$

$$u_i \lambda_i = 0, \tilde{u}_i (C - \lambda_i) = 0, \quad i = 1, 2, \dots, m$$

$$u_i, \tilde{u}_i \geq 0, \quad i = 1, 2, \dots, m.$$

- Any feasible point λ which satisfies the Kuhn-Tucker necessary conditions in above is a global maximum solution.

Justification of Strong Duality for SVM - General Case

- $X = \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^m$: a non-empty convex set.
- $F(\mathbf{w}, b, \eta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \eta_i$: a convex function on X .
- $g_i(\mathbf{w}, b, \eta) = 1 - \eta_i - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i + b), 1 \leq i \leq m$: affine functions so that they are convex functions on X .
- $h_i(\mathbf{w}, b, \eta) = -\eta_i, 1 \leq i \leq m$: affine functions so that they are convex functions on X .
- There exists an $(\mathbf{w}', b', \eta') \in X$ such that $\mathbf{g}(\mathbf{w}', b', \eta') < \mathbf{0}$ and $\mathbf{h}(\mathbf{w}', b', \eta') < \mathbf{0}$.

Then we have

$$\begin{aligned} & \inf \{ F(\mathbf{w}, b, \eta) : (\mathbf{w}, b, \eta) \in X, \mathbf{g}(\mathbf{w}, b, \eta) \leq \mathbf{0}, \mathbf{h}(\mathbf{w}, b, \eta) \leq \mathbf{0} \} \\ &= \sup \{ \theta(\lambda, \mu) : (\lambda, \mu) \geq \mathbf{0} \}. \end{aligned}$$

- For a non-trivial labeled training sample, the inf is finite and can be achieved at some feasible point $(\mathbf{w}^{SVM}, b^{SVM}, \eta^{SVM})$. Then $\sup\{\theta(\lambda) \mid \lambda \geq \mathbf{0}\}$ is achieved at some $(\lambda^{SVM}, \mu^{SVM}) \geq \mathbf{0}$.
- The primal and dual problems are equivalent.

The SVM Algorithm - General Case

- $S = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m)$: a non-trivial labeled training sample of size m with labels $(c(\mathbf{x}_1), c(\mathbf{x}_2), \dots, c(\mathbf{x}_m))$.
- h_S^{SVM} : the hypothesis returned by SVM,

$$\begin{aligned} h_S^{SVM}(\mathbf{x}) &= \text{sgn}(\mathbf{w}^{SVM} \cdot \mathbf{x} + b^{SVM}) \\ &= \text{sgn}\left(\sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot \mathbf{x}) + b^{SVM}\right) \end{aligned}$$

- $b^{SVM} = c(\mathbf{x}_j) - \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot \mathbf{x}_j)$ for any support vector \mathbf{x}_j with $0 < \lambda_j < C$. Thus we have

$$h_S^{SVM}(\mathbf{x}) = \text{sgn}\left(c(\mathbf{x}_j) + \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot (\mathbf{x} - \mathbf{x}_j))\right)$$

for any support vector \mathbf{x}_j with $0 < \lambda_j < C$.

- The hypothesis solution h_S^{SVM} depends only on inner products between vectors and not directly on the vectors themselves.

The SVM Soft Margin ρ_{SVM}

- $b^{SVM} = c(\mathbf{x}_j) - c(\mathbf{x}_j)\eta_j^{SVM} - \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i)(\mathbf{x}_i \cdot \mathbf{x}_j)$ for any support vector \mathbf{x}_j , i.e., $\lambda_j^{SVM} > 0$. This implies

$$\begin{aligned}
 & \sum_{j=1}^m \lambda_j^{SVM} c(\mathbf{x}_j) b^{SVM} \\
 = & \sum_{j=1}^m \lambda_j^{SVM} (1 - \eta_j^{SVM}) c(\mathbf{x}_j)^2 \\
 & - \sum_{j=1}^m \lambda_j^{SVM} c(\mathbf{x}_j) \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i) (\mathbf{x}_i \cdot \mathbf{x}_j).
 \end{aligned}$$

- Since $\sum_{j=1}^m \lambda_j^{SVM} c(\mathbf{x}_j) = 0$ and $\mathbf{w}^{SVM} = \sum_{i=1}^m \lambda_i^{SVM} c(\mathbf{x}_i) \mathbf{x}_i$,

we have

$$\sum_{j=1}^m \lambda_j^{SVM} (1 - \eta_j^{SVM}) = \|\mathbf{w}^{SVM}\|^2.$$

- $\rho_{SVM}^2 = \frac{1}{\|\mathbf{w}^{SVM}\|^2} = \frac{1}{\sum_{j=1}^m \lambda_j^{SVM} (1 - \eta_j^{SVM})}.$

The Contents of This Lecture

- Support vector machines - separable case.
- Support vector machines - non-separable case.
- Margin guarantees.

Binary Linear Classification Problem

- $\mathcal{X} \subseteq \mathbb{R}^N$: the input space.
- $\mathcal{Y}' = \mathcal{Y} = \{-1, +1\}$: the output, label space with loss function $L(y', y) = 1_{y' \neq y}$.
- c : a fixed but unknown target concept in the concept class \mathcal{C} .
- $\mathcal{H} = \{\mathbf{x} \mapsto \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b) \mid \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$: the hypothesis set of all linear classifiers.
- $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$: a sample of m items, drawn i.i.d. from the input space according to P , with labels $(c(\mathbf{x}_1), \dots, c(\mathbf{x}_m))$.
- **Problem:** find a linear hypothesis (binary linear classifier) $h : \mathcal{X} \rightarrow \{-1, +1\}$ in \mathcal{H} with small generalization error

$$R(h) = E[1_{h(\mathbf{x}) \neq c(\mathbf{x})}] = P(h(\mathbf{x}) \neq c(\mathbf{x})).$$

VC-Dimension Generalization Bound - Binary Linear Classification

- $\mathcal{S} \subseteq \mathbb{R}^N$: the input space, not contained in any hyperplane.
- $c : \mathcal{S} \rightarrow \{-1, +1\}$: a fixed but unknown target concept in the concept class \mathcal{C} .
- $\mathcal{H} = \{\mathbf{x} \mapsto \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b) \mid \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$: the hypothesis set of all linear classifiers.
 - Since the input space \mathcal{S} is not contained in any hyperplane, $\text{VCdim}(\mathcal{H}) = N + 1$.
- $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$: a sample of size m drawn i.i.d. from the input space \mathcal{S} according to an unknown distribution P , with labels $(c(\mathbf{x}_1), \dots, c(\mathbf{x}_m))$.

For any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\forall h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{2(N+1) \ln \frac{em}{N+1}}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Proof. This is a direct consequence of Corollary 3.4. □

Remarks

- When the dimension N of the input space is large compared to the sample size m , this VC-dimension generalization bound is uninformative.
- Informative bound which does not depend on the dimension N of the input space will be derived.

Geometric Margin of a Point to a Linear Classifier

The geometric margin $\rho_h(\mathbf{x})$ of a point \mathbf{x} in \mathbb{R}^N with respect to a linear classifier $h : x \mapsto \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$ is its distance to the hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$:

$$\rho_h(\mathbf{x}) = \frac{|\mathbf{w} \cdot \mathbf{x} + b|}{\|\mathbf{w}\|}.$$

Geometric Margin of a Finite Set of Points to a Linear Classifier

The geometric margin $\rho_h(A)$ of a finite set $A = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ of points in \mathbb{R}^N with respect to a linear classifier $h : \mathbf{x} \mapsto \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$ is the minimum geometric margin over the points in the set:

$$\rho_h(A) = \min_{1 \leq i \leq m} \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|}.$$

Canonical Representation of a Separating Linear Classifier to a Finite Set of Points

- $A = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$: a finite set of points in \mathbb{R}^N .
- h : a separating linear classifier to A , i.e, no points of A being in the boundary hyperplane of h .

A representation $h : x \mapsto \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$ of the separating linear classifier h to the set A is called canonical to A if

$$\min_{1 \leq i \leq m} |\mathbf{w} \cdot \mathbf{x}_i + b| = 1.$$

The geometric margin of the set A w.r.t. the canonically represented separating linear classifier $h : \mathbf{x} \mapsto \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b)$ to A is

$$\rho_h(A) = \min_{1 \leq i \leq m} \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}.$$

VC-Dimension of a Set of Separating Linear Classifiers to a Finite Set with Margin Guarantee

Theorem 4.2: Let

- $A \subseteq \mathbb{R}^N$: a finite input space with $r \triangleq \max_{\mathbf{x} \in A} \|\mathbf{x}\|_2$.
- \mathcal{H} : the set of all separating linear classifiers to A with geometric margin at least $1/\Lambda$ whose boundary hyperplane contains the origin $\mathbf{0}$, i.e.,

$$\mathcal{H} = \{\mathbf{x} \mapsto \text{sgn}(\mathbf{w} \cdot \mathbf{x}) \mid \min_{\mathbf{x} \in A} |\mathbf{w} \cdot \mathbf{x}_i| = 1 \text{ and } \|\mathbf{w}\| \leq \Lambda\}.$$

- Every separating hyperplane to the set A has a unique canonical representation to A up to ± 1 .
- Each linear classifier (hypothesis) h in \mathcal{H} is a function from the input space A to the output (label) space $\{-1, +1\}$.

Then $d = \text{VC dim}(\mathcal{H}) \leq r^2 \Lambda^2$.

Proof. Assume

- $B = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d\}$: a d -subset of A that can be shattered by \mathcal{H} ;
- $\mathbf{y} = (y_1, y_2, \dots, y_d) \in \{-1, +1\}^d$: a dichotomy of B ;
- $\mathbf{x} \mapsto \mathbf{w}_{\mathbf{y}} \cdot \mathbf{x}$: a linear classifier in \mathcal{H} which realizes the dichotomy \mathbf{y} of B .
 - $\mathbf{w}_{\mathbf{y}}$ depends on \mathbf{y} .

Then we have

$$1 \leq y_i(\mathbf{w}_{\mathbf{y}} \cdot \mathbf{x}_i) \quad \forall i \in [1, d]$$

and, summing up over i , yield

$$d \leq \mathbf{w}_{\mathbf{y}} \cdot \sum_{i=1}^d y_i \mathbf{x}_i \leq \|\mathbf{w}_{\mathbf{y}}\| \left\| \sum_{i=1}^d y_i \mathbf{x}_i \right\|.$$

By taking equally weighted sum over all possible dichotomies \mathbf{y} and

noting that $\|\mathbf{w}_y\| \leq \Lambda$, we have

$$d \leq \Lambda \sum_{\mathbf{y} \in \{-1, +1\}^d} \frac{1}{2^d} \sqrt{\sum_{i=1}^d y_i \mathbf{x}_i \cdot \sum_{j=1}^d y_j \mathbf{x}_j}$$

$$\leq \Lambda \sqrt{\sum_{\mathbf{y} \in \{-1, +1\}^d} \frac{1}{2^d} \sum_{i,j=1}^d y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)}$$

since $f(x) = \sqrt{x}$ is a concave function on $[0, \infty)$

$$= \Lambda \sqrt{\sum_{i,j=1}^d (\mathbf{x}_i \cdot \mathbf{x}_j) \frac{1}{2^d} \sum_{\mathbf{y} \in \{-1, +1\}^d} y_i y_j}$$

$$= \Lambda \sqrt{\sum_{i=1}^d (\mathbf{x}_i \cdot \mathbf{x}_i)}$$

$$\leq \Lambda \sqrt{dr^2} = \Lambda r \sqrt{d}$$

since

$$\frac{1}{2^d} \sum_{\mathbf{y} \in \{-1, +1\}^d} y_i y_j = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

Thus we have $d \leq \Lambda^2 r^2$.

□

Rademacher Complexity of a Set of Linear Functions on Bounded Input Space with Bounded Weight Vector

Theorem 4.3: Let

- $\mathcal{S} = \bar{B}(r; \mathbf{0}) = \{\mathbf{x} : \|\mathbf{x}\| \leq r\} \subseteq \mathbb{R}^N$: the bounded input space, associated with a probability space $(\bar{B}(r; \mathbf{0}), \mathcal{F}, P)$.
- $\mathcal{H} = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \mid \|\mathbf{w}\| \leq \Lambda\}$: the set of all linear functions with bounded weight vector.
- $S = (\mathbf{x}_1, \dots, \mathbf{x}_m)$: a sample of m points drawn i.i.d. from the input space $\bar{B}(r; \mathbf{0})$ according to an unknown distribution P .

Then the empirical Rademacher complexity of \mathcal{H} w.r.t. the sample S can be upper bounded as follows:

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \sqrt{\frac{r^2 \Lambda^2}{m}}.$$

Proof.

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(\mathbf{x}_i) \\
&= \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sup_{\|\mathbf{w}\| \leq \Lambda} \frac{1}{m} \sum_{i=1}^m \sigma_i (\mathbf{w} \cdot \mathbf{x}_i) \\
&= \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sup_{\|\mathbf{w}\| \leq \Lambda} \frac{1}{m} \mathbf{w} \cdot \sum_{i=1}^m \sigma_i \mathbf{x}_i \\
&\leq \frac{\Lambda}{m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \frac{1}{2^m} \sqrt{\sum_{i=1}^d \sigma_i \mathbf{x}_i \cdot \sum_{j=1}^d \sigma_j \mathbf{x}_j} \\
&\leq \frac{\Lambda}{m} \sqrt{\sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \frac{1}{2^m} \sum_{i,j=1}^m \sigma_i \sigma_j (\mathbf{x}_i \cdot \mathbf{x}_j)},
\end{aligned}$$

again since $f(x) = \sqrt{x}$ is a concave function on $[0, \infty)$. Now we

have

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{H}) &\leq \frac{\Lambda}{m} \sqrt{\sum_{i,j=1}^m (\mathbf{x}_i \cdot \mathbf{x}_j) \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sigma_i \sigma_j} \\
&= \frac{\Lambda}{m} \sqrt{\sum_{i=1}^m (\mathbf{x}_i \cdot \mathbf{x}_i)} \\
&\leq \frac{\Lambda}{m} \sqrt{mr^2} = \sqrt{\frac{\Lambda^2 r^2}{m}}
\end{aligned}$$

since

$$\frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sigma_i \sigma_j = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

Thus we have $\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \sqrt{\frac{r^2 \Lambda^2}{m}}$. □

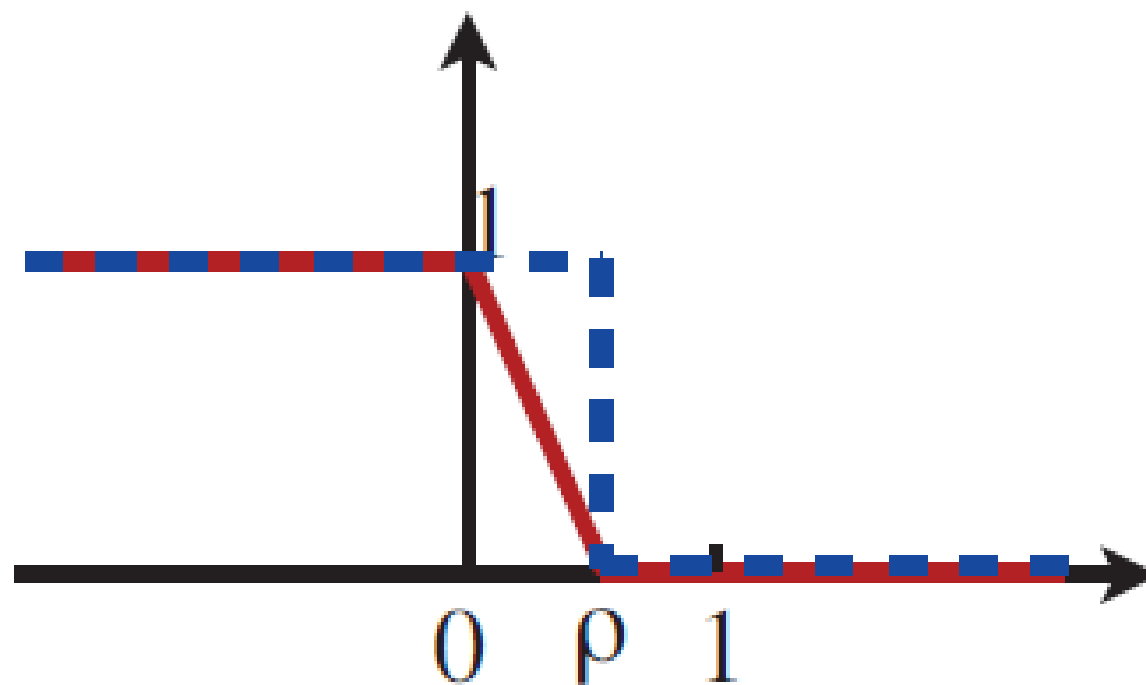
ρ -Margin Loss Function

- $\rho > 0$: a given confidence margin.
- $\Phi_\rho(x) : \mathbb{R} \rightarrow [0, 1]$: a soft inverse limiter with margin ρ , defined as

$$\Phi_\rho(x) = \begin{cases} 1, & \text{if } x \leq 0, \\ 1 - x/\rho, & \text{if } 0 \leq x \leq \rho, \\ 0, & \text{if } x \geq \rho. \end{cases}$$

The ρ -margin loss function $L_\rho : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ is defined as

$$L_\rho(y', y) \triangleq \Phi_\rho(y'y).$$



Three functions $\Phi_0(x) \leq \Phi_\rho(x)$ (in red) $\leq \Phi_0(x - \rho)$ (in blue) for constructing different loss functions.

Remarks

- When using a real-valued function h as a hypothesis to approximate a concept c which is a $\{-1, +1\}$ -valued function, the 0-1 loss function used will be

$$L(y', y) = 1_{\text{sgn}(y') \neq \text{sgn}(y)} = 1_{y'y \leq 0} = \Phi_0(y'y),$$

where $\Phi_0(x)$ is the hard inverse limiter,

$$\Phi_0(x) = \begin{cases} 1, & \text{if } x \leq 0, \\ 0, & \text{if } x > 0. \end{cases}$$

- The 0-1 loss function $L(y', y) = 1_{y'y \leq 0}$ is always no greater than the ρ -margin loss function $L_\rho(y', y)$.

Empirical ρ -Margin Loss

- \mathcal{I} : the input space of all possible items ω , associated with a probability space $(\mathcal{I}, \mathcal{F}, P)$.
- $c : \mathcal{I} \rightarrow \{-1, +1\}$: a fixed but unknown target concept in the concept class \mathcal{C} .
- \mathcal{Y}' : the output space, which is usually a bounded subset of \mathbb{R} .
- \mathcal{H} : a hypothesis set of \mathcal{Y}' -valued functions on the input space \mathcal{I} .
- $L_\rho(y', y) = \Phi_\rho(y'y)$: the ρ -margin loss function.
- $S = (\omega_1, \dots, \omega_m)$: a sample of size m drawn i.i.d. from \mathcal{I} according to an unknown distribution P , with labels $(c(\omega_1), \dots, c(\omega_m))$.
- h : an arbitrary hypothesis in \mathcal{H} .

The empirical ρ -margin loss of an hypothesis h w.r.t. the concept c on the labeled sample S is defined as

$$\hat{R}_{S,\rho}(h) \triangleq \frac{1}{m} \sum_{i=1}^m L_{\rho}(h(\omega_i), c(\omega_i)) = \frac{1}{m} \sum_{i=1}^m \Phi_{\rho}(h(\omega_i)c(\omega_i)).$$

Remarks

- Since the 0-1 loss function $L(y', y) = 1_{y'y \leq 0}$ is always no greater than the ρ -margin loss function $L_\rho(y', y)$, the empirical error is

$$\begin{aligned}\hat{R}_S(h) &= \frac{1}{m} \sum_{i=1}^m L(h(\omega_i), c(\omega_i)) \\ &\leq \frac{1}{m} \sum_{i=1}^m L_\rho(h(\omega_i), c(\omega_i)) = \hat{R}_{S,\rho}(h).\end{aligned}$$

Talagrand's Lemma

Lemma 4.2: Let

- \mathcal{I} : the input space of all possible items ω , associated with a probability space $(\mathcal{I}, \mathcal{F}, P)$.
- $\mathcal{Y}' \subseteq \mathbb{R}$: the output space, which is a subset of \mathbb{R} .
- \mathcal{H} : a hypothesis set of \mathcal{Y}' -valued measurable functions on the input space \mathcal{I} .
- $\Phi : \mathcal{Y}' \rightarrow \mathbb{R}$: an α -Lipschitz function, i.e., there is an $\alpha > 0$ such that $|\Phi(x) - \Phi(y)| \leq \alpha|x - y|$, $\forall x, y \in \mathcal{Y}'$.
- $S = (\omega_1, \omega_2, \dots, \omega_m)$: a sample of m items drawn i.i.d. from \mathcal{I} according to P .

Assume that

- $\sup_{h \in \mathcal{H}} \left(\sum_{i=1}^j \sigma_i(\Phi \circ h)(\omega_i) + \sum_{i=j+1}^m \alpha \sigma_i h(\omega_i) \right)$ is finite for all $\sigma_i \in \{-1, +1\}, i \in [1, m]$ and for all $j \in [0, m]$.

Then we have

$$\hat{\mathfrak{R}}_S(\Phi \circ \mathcal{H}) \leq \alpha \hat{\mathfrak{R}}_S(\mathcal{H}),$$

where both $\hat{\mathfrak{R}}_S(\Phi \circ \mathcal{H})$ and $\hat{\mathfrak{R}}_S(\mathcal{H})$ are finite.

Proof. By the definition of empirical Rademacher complexity,

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\Phi \circ \mathcal{H}) &= \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i(\Phi \circ h)(\omega_i) \\
&= \frac{1}{2^{m-1}} \sum_{\sigma_1, \sigma_2, \dots, \sigma_{m-1} \in \{-1, +1\}} \frac{1}{2} \sum_{\sigma_m \in \{-1, +1\}} \frac{1}{m} \\
&\quad \sup_{h \in \mathcal{H}} (u_{m-1}(h) + \sigma_m(\Phi \circ h)(\omega_m)),
\end{aligned}$$

where $u_{m-1}(h) \triangleq \sum_{i=1}^{m-1} \sigma_i(\Phi \circ h)(\omega_i)$. Since $\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i(\Phi \circ h)(\omega_i)$ is finite for any given $\sigma_1, \sigma_2, \dots, \sigma_m$ by assumption, for any $\epsilon > 0$, there exist $h_1, h_2 \in \mathcal{H}$ such that

$$\begin{aligned}
\sup_{h \in \mathcal{H}} (u_{m-1}(h) + (\Phi \circ h)(\omega_m)) - \epsilon &\leq u_{m-1}(h_1) + (\Phi \circ h_1)(\omega_m), \\
\sup_{h \in \mathcal{H}} (u_{m-1}(h) - (\Phi \circ h)(\omega_m)) - \epsilon &\leq u_{m-1}(h_2) - (\Phi \circ h_2)(\omega_m)
\end{aligned}$$

and then

$$\begin{aligned}
& \frac{1}{2} \sum_{\sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} (u_{m-1}(h) + \sigma_m (\Phi \circ h)(\omega_m)) - \epsilon \\
& \leq \frac{1}{2} (u_{m-1}(h_1) + (\Phi \circ h_1)(\omega_m)) + \frac{1}{2} (u_{m-1}(h_2) - (\Phi \circ h_2)(\omega_m)) \\
& \leq \frac{1}{2} (u_{m-1}(h_1) + u_{m-1}(h_2) + s\alpha(h_1(\omega_m) - h_2(\omega_m))) \\
& \quad \text{by Lipschitz property, where } s = \text{sgn}(h_1(\omega_m) - h_2(\omega_m)) \\
& = \frac{1}{2} (u_{m-1}(h_1) + s\alpha h_1(\omega_m)) + \frac{1}{2} (u_{m-1}(h_2) - s\alpha h_2(\omega_m)) \\
& \leq \frac{1}{2} \sup_{h \in \mathcal{H}} (u_{m-1}(h) + s\alpha h(\omega_m)) + \frac{1}{2} \sup_{h \in \mathcal{H}} (u_{m-1}(h) - s\alpha h(\omega_m)) \\
& = \frac{1}{2} \sum_{\sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} (u_{m-1}(h) + \alpha \sigma_m h(\omega_m)).
\end{aligned}$$

Since the inequality holds for any $\epsilon > 0$, we have

$$\begin{aligned} & \frac{1}{2} \sum_{\sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} (u_{m-1}(h) + \sigma_m(\Phi \circ h)(\omega_m)) \\ & \leq \frac{1}{2} \sum_{\sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} (u_{m-1}(h) + \alpha \sigma_m h(\omega_m)). \end{aligned}$$

Now we have

$$\begin{aligned} \hat{\mathfrak{R}}_S(\Phi \circ \mathcal{H}) & \leq \frac{1}{2^{m-1}} \sum_{\sigma_1, \sigma_2, \dots, \sigma_{m-1} \in \{-1, +1\}} \frac{1}{2} \sum_{\sigma_m \in \{-1, +1\}} \frac{1}{m} \\ & \quad \sup_{h \in \mathcal{H}} \left(\sum_{i=1}^{m-1} \sigma_i(\Phi \circ h)(\omega_i) + \alpha \sigma_m h(\omega_m) \right) \\ & = \frac{1}{2^{m-1}} \sum_{\sigma_1, \dots, \sigma_{m-2}, \sigma_m \in \{-1, +1\}} \frac{1}{2} \sum_{\sigma_{m-1} \in \{-1, +1\}} \frac{1}{m} \\ & \quad \sup_{h \in \mathcal{H}} (u_{m-2}(h) + \sigma_{m-1}(\Phi \circ h)(\omega_{m-1})), \end{aligned}$$

where $u_{m-2}(h) \triangleq \sum_{i=1}^{m-2} \sigma_i(\Phi \circ h)(\omega_i) + \alpha \sigma_m h(\omega_m)$. Since $\sup_{h \in \mathcal{H}} \left(\sum_{i=1}^{m-1} \sigma_i(\Phi \circ h)(\omega_i) + \alpha \sigma_i h(\omega_i) \right)$ is finite for any given $\sigma_1, \sigma_2, \dots, \sigma_m$ by assumption, by proceeding similar argument in

above, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\Phi \circ \mathcal{H}) &\leq \frac{1}{2^{m-1}} \sum_{\sigma_1, \dots, \sigma_{m-2}, \sigma_m \in \{-1, +1\}} \frac{1}{2} \sum_{\sigma_{m-1} \in \{-1, +1\}} \frac{1}{m} \\
&\quad \sup_{h \in \mathcal{H}} (u_{m-2}(h) + \alpha \sigma_{m-1} h(\omega_{m-1})) \\
&= \frac{1}{2^{m-1}} \sum_{\sigma_1, \dots, \sigma_{m-2}, \sigma_m \in \{-1, +1\}} \frac{1}{2} \sum_{\sigma_{m-1} \in \{-1, +1\}} \frac{1}{m} \\
&\quad \sup_{h \in \mathcal{H}} \left(\sum_{i=1}^{m-2} \sigma_i (\Phi \circ h)(\omega_i) + \alpha \sum_{i=m-1}^m \sigma_i h(\omega_i) \right) \\
&= \frac{1}{2^{m-1}} \sum_{\sigma_1, \dots, \sigma_{m-3}, \sigma_{m-1} \sigma_m \in \{-1, +1\}} \frac{1}{2} \sum_{\sigma_{m-2} \in \{-1, +1\}} \frac{1}{m} \\
&\quad \sup_{h \in \mathcal{H}} (u_{m-3}(h) + \sigma_{m-2} (\Phi \circ h)(\omega_{m-2})),
\end{aligned}$$

where $u_{m-3}(h) \triangleq \sum_{i=1}^{m-3} \sigma_i (\Phi \circ h)(\omega_i) + \alpha \sum_{i=m-1}^m \sigma_i h(\omega_i)$. By

continuing similar argument, we have

$$\begin{aligned}\hat{\mathfrak{R}}_S(\Phi \circ \mathcal{H}) &\leq \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \alpha \sum_{i=1}^m \sigma_i h(\omega_i) \\ &= \alpha \hat{\mathfrak{R}}_S(\mathcal{H}).\end{aligned}$$

□

Remarks

- By assuming that

$$\sup_{h \in \mathcal{H}} \left(\sum_{i=1}^j \sigma_i (\Phi \circ h)(\omega_i) + \sum_{i=j+1}^m \alpha \sigma_i h(\omega_i) \right)$$

is finite for all $\sigma_i \in \{-1, +1\}$, $i \in [1, m]$, for all $j \in [0, m]$ and for all random samples $S = (\omega_1, \dots, \omega_m)$ of size m and by taking average over the random sample S of size m , we have

$$\mathfrak{R}_m(\Phi \circ \mathcal{H}) \leq \alpha \mathfrak{R}_m(\mathcal{H}).$$

- The soft inverse limiter $\Phi_\rho(x)$ with margin $\rho > 0$ is a $1/\rho$ -Lipschitz function since its maximum slope is $1/\rho$.

Margin-Based Generalization Bound for Binary Classification

Theorem 4.4: Let

- \mathcal{I} : the input space of all possible items ω , associated with a probability space $(\mathcal{I}, \mathcal{F}, P)$.
- $c : \mathcal{I} \rightarrow \{-1, +1\}$: a fixed but unknown target concept in the concept class \mathcal{C} .
- $\mathcal{Y}' \subseteq \mathbb{R}$: the output space, which is a subset of \mathbb{R} .
- \mathcal{H} : a hypothesis set of \mathcal{Y}' -valued measurable functions on the input space \mathcal{I} such that $\sup_{h \in \mathcal{H}} |h(\omega)| < +\infty \ \forall \ \omega \in \mathcal{I}$.
- $S = (\omega_1, \omega_2, \dots, \omega_m)$: a sample of m items drawn i.i.d. from \mathcal{I} according to an unknown distribution P with labels $(c(\omega_1), c(\omega_2), \dots, c(\omega_m))$.

- $\rho > 0$: a given confidence margin.
- $L_\rho(y', y) = \Phi_\rho(y'y) : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$: the ρ -margin loss function.
- $g_h : \mathcal{I} \times \{-1, +1\} \rightarrow [0, 1]$: the loss function associated with h under the ρ -margin loss function L_ρ , defined as $g_h(\omega, y) \triangleq L_\rho(h(\omega), y) = \Phi_\rho(h(\omega)y)$.
- $\mathcal{G} = \{g_h \mid h \in \mathcal{H}\}$: the family of loss functions associated with hypotheses in \mathcal{H} under the ρ -margin loss function L_ρ .
- $\mathcal{Z} = \mathcal{I} \times \{-1, +1\}$: the input set of loss functions g_h , associated with a probability space $(\mathcal{Z}, \tilde{\mathcal{F}}, \tilde{P})$ where \tilde{P} is an extension of P from on \mathcal{F} to on $\tilde{\mathcal{F}} = \mathcal{F} \times 2^{\{-1, +1\}}$.
- $\tilde{S} = ((\omega_1, c(\omega_1)), \dots, (\omega_m, c(\omega_m)))$: the labeled sample corresponding to S .
- $\hat{A}_{\tilde{S}}(g_h) = \frac{1}{m} \sum_{i=1}^m g_h(\omega_i, c(\omega_i)) = \frac{1}{m} \sum_{i=1}^m L_\rho(h(\omega_i), c(\omega_i)) = \hat{R}_{S, \rho}(h)$, the empirical ρ -margin loss of h w.r.t. c on sample S .

- $E_{z \sim \tilde{P}}[g_h(z)] = E_{\tilde{S} \sim \tilde{P}_m}[\hat{A}_{\tilde{S}}(g_h)] = E_{S \sim P_m}[\hat{R}_{S,\rho}(h)] \geq E_{S \sim P_m}[\hat{R}_S(h)] = R(h).$

For any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all h in \mathcal{H} :

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho} \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}},$$

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho} \hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

Proof. By the Rademacher complexity bound for the family \mathcal{G} in Theorem 3.1, for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all g_h in \mathcal{G} :

$$\begin{aligned} E_{z \sim \tilde{P}}[g_h(z)] &\leq \frac{1}{m} \sum_{i=1}^m g_h(\omega_i, c(\omega_i)) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \\ E_{z \sim \tilde{P}}[g_h(z)] &\leq \frac{1}{m} \sum_{i=1}^m g_h(\omega_i, c(\omega_i)) + 2\hat{\mathfrak{R}}_{\tilde{\mathcal{S}}}(\mathcal{G}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}. \end{aligned}$$

Let

$$\tilde{\mathcal{H}} \triangleq \{z = (\omega, y) \mapsto h(\omega)y \mid h \in \mathcal{H}\},$$

which is a family of $(-\mathcal{Y}' \cup \mathcal{Y}')$ -valued functions on the input set $\mathcal{Z} = \mathcal{I} \times \{-1, +1\}$. It is clear that $\mathcal{G} = \Phi_\rho \circ \tilde{\mathcal{H}}$. Since Φ_ρ is a bounded $1/\rho$ -Lipschitz function and $\sup_{h \in \mathcal{H}} |h(\omega)|$ is finite for all $\omega \in \mathcal{I}$, $\sup_{h \in \mathcal{H}} \left(\sum_{i=1}^j \sigma_i \Phi_\rho(h(\omega_i)c(\omega)) + \sum_{i=j+1}^m \frac{1}{\rho} \sigma_i h(\omega_i)c(\omega) \right)$ is finite for all $\sigma_i \in \{-1, +1\}, i \in [1, m]$, for all $j \in [0, m]$ and for all

sample $S = (\omega_1, \dots, \omega_m)$ of size m , we have

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) \leq \frac{1}{\rho} \hat{\mathfrak{R}}_{\tilde{S}}(\tilde{\mathcal{H}}) \text{ and then } \mathfrak{R}_m(\mathcal{G}) \leq \frac{1}{\rho} \mathfrak{R}_m(\tilde{\mathcal{H}})$$

by Talagrand's lemma. The empirical Rademacher complexity of $\tilde{\mathcal{H}}$ is

$$\begin{aligned} \hat{\mathfrak{R}}_{\tilde{S}}(\tilde{\mathcal{H}}) &= \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i c(\omega_i) h(\omega_i) \\ &= \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \dots, \sigma_m \in \{-1, +1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(\omega_i) = \hat{\mathfrak{R}}_S(\mathcal{H}) \end{aligned}$$

and then $\mathfrak{R}_m(\tilde{\mathcal{H}}) = \mathfrak{R}_m(\mathcal{H})$. Now with

$$\frac{1}{m} \sum_{i=1}^m g_h(\omega_i, c(\omega_i)) = \hat{R}_{S, \rho}(h) \text{ and } R(h) \leq E_{z \sim \tilde{P}}[g_h(z)],$$

the theorem is proved. □

Remarks

- The margin-based generalization bound for binary classification shows the trade-off between two terms: the larger the desired margin ρ , the smaller the middle term; however, the first term, the empirical ρ -margin loss $\hat{R}_{S,\rho}(h)$, increases as a function of ρ .
- The bounds of Theorem 4.4 can be generalized to hold uniformly for all $\rho > 0$ at the cost of an additional term $\sqrt{\frac{\ln \log_2 \frac{2}{\rho}}{m}}$.

Margin-Based Generalization Bound for Linear Hypotheses on Bounded Input Space with Bounded Weight Vector

Corollary 4.1: Let

- $\mathcal{S} = \bar{B}(r; \mathbf{0}) = \{\mathbf{x} : \|\mathbf{x}\| \leq r\} \subseteq \mathbb{R}^N$: a bounded input space, associated with a probability space $(\bar{B}(r; \mathbf{0}), \mathcal{F}, P)$.
- $c : \mathcal{S} \rightarrow \{-1, +1\}$: a fixed but unknown target concept in the concept class \mathcal{C} .
- $\mathcal{H} = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \mid \|\mathbf{w}\| \leq \Lambda\}$: the set of all linear functions with bounded weight vector.
 - It is clear that $\sup_{h \in \mathcal{H}} |h(\mathbf{x})| \leq \Lambda \|\mathbf{x}\| < +\infty \forall \mathbf{x} \in \mathcal{S}$.
- $S = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m)$: a sample of m points drawn i.i.d. from the input space $\bar{B}(r; \mathbf{0})$ according to an unknown distribution P with labels $(c(\mathbf{x}_1), c(\mathbf{x}_2), \dots, c(\mathbf{x}_m))$.

- $\rho > 0$: a given confidence margin.
- $L_\rho(y', y) = \Phi_\rho(y'y) : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$: the ρ -margin loss function.
- $\hat{R}_{S,\rho}(h) = \frac{1}{m} \sum_{i=1}^m L_\rho(h(\mathbf{x}_i), c(\mathbf{x}_i)) = \frac{1}{m} \sum_{i=1}^m \Phi_\rho(h(\mathbf{x}_i)c(\mathbf{x}_i))$: the empirical ρ -margin loss of a linear hypothesis h in \mathcal{H} w.r.t. the concept c on the sample S .
- $R(h) = E_{\mathbf{x} \sim P}[1_{\text{sgn}(h(\mathbf{x})) \neq c(\mathbf{x})}]$: the generalization error of linear hypothesis $h \in \mathcal{H}$.

For any $\delta > 0$, with probability at least $1 - \delta$, all h in \mathcal{H} :

$$R(h) \leq \hat{R}_{S,\rho}(h) + 2\sqrt{\frac{r^2 \Lambda^2 / \rho^2}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Proof. This is a direct consequence of Theorems 4.3 and 4.4. \square

Remarks

- The margin-based generalization bound for linear hypotheses does not depend directly on the dimension of the input space, but only on the margin.
- It suggests that a small generalization error can be achieved when ρ/r is large (small second term) while the empirical ρ -margin loss is relatively small (first term).
 - The latter occurs when few points are either classified incorrectly or correctly, but with margin less than ρ .
- The learning guarantee in Corollary 4.1 hinges upon the hope of a good margin value ρ : if there exists a relatively large margin value $\rho > 0$ for which the empirical ρ -margin loss is small, then a small generalization error is guaranteed by the corollary.

- This favorable margin ρ depends on the distribution: while the learning bound is distribution-independent, the existence of a good margin is in fact distribution-dependent.

Strong Justification for SVM

- For $\rho = 1$, the soft inverse limiter Φ_1 with margin 1 is upper bounded by the hinge function $x \mapsto \max(1 - x, 0)$:

$$\Phi_1(x) \leq \max(1 - x, 0) \quad \forall x \in \mathbb{R}$$

and then the empirical ρ -margin loss $\hat{R}_{S,\rho}(h)$ of a linear hypothesis $h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$ is upper bounded by the average amount of slack penalty:

$$\begin{aligned} \hat{R}_{S,\rho}(h) &= \frac{1}{m} \sum_{i=1}^m \Phi_1(h(\mathbf{x}_i)c(\mathbf{x}_i)) \\ &\leq \frac{1}{m} \sum_{i=1}^m \max(1 - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i), 0) \\ &= \frac{1}{m} \sum_{i=1}^m \eta_i. \end{aligned}$$

- The margin-based generalization bound implies that with probability at least $1 - \delta$, for any linear function $h : \mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x}$ with $\|\mathbf{w}\| \leq \Lambda$ on bounded input space $\bar{B}(r; \mathbf{0})$,

$$R(h) \leq \frac{1}{m} \sum_{i=1}^m \eta_i + 2\sqrt{\frac{r^2 \Lambda^2}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}},$$

where $\eta_i = \max(1 - c(\mathbf{x}_i)(\mathbf{w} \cdot \mathbf{x}_i), 0)$ are the slack penalty over the training set.

- The objective function minimized by the SVM algorithm has precisely the form of this upper bound: the first term corresponds to the slack penalty over the training set and the second to the minimization of the $\|\mathbf{w}\|$ which is equivalent to that of $\|\mathbf{w}\|^2$.
- We have been using a parameter C in SVM to adjust the relative strength in the minimization of either term.

Searching for Large-Margin Separating Hyperplanes in High-Dimension Space

- Since margin-based generalization bound does not directly depend on the dimension of the input space and do guarantee good generalization with a favorable margin, it suggests seeking large-margin separating hyperplanes in a very high-dimensional space.
- The next lecture provides a way of doing this, in addition to overcoming the very high cost of computation with very high-dimensional vectors as well as further generalization of SVM to nonlinear separation.