Foundations of Machine Learning
Support Vector Machines

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu
Binary Classification Problem

- **Training data**: sample drawn i.i.d. from set $X \subseteq \mathbb{R}^N$ according to some distribution $D$,

$$S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in X \times \{-1, +1\}.$$

- **Problem**: find hypothesis $h: X \mapsto \{-1, +1\}$ in $H$ (classifier) with small generalization error $R(h)$.

- choice of hypothesis set $H$ : learning guarantees of previous lecture.

  → linear classification (hyperplanes) if dimension $N$ is not too large.
This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees
Linear Separation

- **classifiers**: \( H = \{ x \mapsto \text{sgn}(w \cdot x + b) : w \in \mathbb{R}^N, b \in \mathbb{R} \} \).

- **geometric margin**: \( \rho = \min_{i \in [1,m]} \frac{|w \cdot x_i + b|}{\|w\|} \).

- **which separating hyperplane?**
Optimal Hyperplane: Max. Margin

(Vapnik and Chervonenkis, 1965)

\[ w \cdot x + b = 0 \]

\[ w \cdot x + b = -1 \]

\[ w \cdot x + b = +1 \]

\[ \rho = \max_{w,b : y_i(w \cdot x_i + b) \geq 0 \text{ } \in [1,m]} \min \frac{|w \cdot x_i + b|}{\|w\|} \]
Maximum Margin

\[ \rho = \max_{w, b : y_i(w \cdot x_i + b) \geq 0 \quad i \in [1, m]} \min_{i \in [1, m]} \frac{|w \cdot x_i + b|}{\|w\|} \]

\[ = \max_{w, b : y_i(w \cdot x_i + b) \geq 0 \quad i \in [1, m]} \min_{i \in [1, m]} \frac{|w \cdot x_i + b|}{\|w\|} \]

\[ \min_{i \in [1, m]} |w \cdot x_i + b| = 1 \]

\[ = \max_{w, b : y_i(w \cdot x_i + b) \geq 0 \quad i \in [1, m]} \frac{1}{\|w\|} \]

\[ \min_{i \in [1, m]} |w \cdot x_i + b| = 1 \]

\[ = \max_{w, b : y_i(w \cdot x_i + b) \geq 1} \frac{1}{\|w\|} \]

(scale-invariance)
Optimization Problem

Constrained optimization:

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

subject to \( y_i(w \cdot x_i + b) \geq 1, i \in [1, m] \).

Properties:

- Convex optimization.
- Unique solution for linearly separable sample.
Optimal Hyperplane Equations

- **Lagrangian:** for all $w, b, \alpha_i \geq 0$,

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

- **KKT conditions:**

$$\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{m} \alpha_i y_i x_i.$$

$$\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \iff \sum_{i=1}^{m} \alpha_i y_i = 0.$$

$$\forall i \in [1, m], \alpha_i [y_i(w \cdot x_i + b) - 1] = 0.$$
Support Vectors

- **Complementarity conditions:**

\[ \alpha_i [y_i (w \cdot x_i + b) - 1] = 0 \implies \alpha_i = 0 \lor y_i (w \cdot x_i + b) = 1. \]

- **Support vectors:** vectors \( x_i \) such that

\[ \alpha_i \neq 0 \land y_i (w \cdot x_i + b) = 1. \]

- **Note:** support vectors are not unique.
Moving to The Dual

Plugging in the expression of \( w \) in \( L \) gives:

\[
L = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i x_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i.
\]

Thus,

\[
L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j).
\]
Equivalent Dual Opt. Problem

- **Constrained optimization:**

\[
\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)
\]

subject to: \(\alpha_i \geq 0 \land \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1,m].\)

- **Solution:**

\[
h(x) = \text{sgn}(\sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b),
\]

with \(b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i)\) for any SV \(x_i\).
**Leave-One-Out Error**

- **Definition:** let $h_S$ be the hypothesis output by learning algorithm $L$ after receiving sample $S$ of size $m$. Then, the leave-one-out error of $L$ over $S$ is:

$$
\hat{R}_{\text{loo}}(L) = \frac{1}{m} \sum_{i=1}^{m} 1_{h_{S \setminus \{x_i\}}(x_i) \neq f(x_i)}.
$$

- **Property:** unbiased estimate of expected error of hypothesis trained on sample of size $m-1$,

$$
E_{S \sim D^m} [\hat{R}_{\text{loo}}(L)] = \frac{1}{m} \sum_{i=1}^{m} E_S [1_{h_{S \setminus \{x_i\}}(x_i) \neq f(x_i)}] = E_S [1_{h_{S \setminus \{x\}}(x) \neq f(x)}]
$$

$$
= E_{S' \sim D^{m-1}} [E_{x \sim D} [1_{h_{S'}(x) \neq f(x)}]] = E_{S' \sim D^{m-1}} [R(h_{S'})].
$$
Leave-One-Out Analysis

- **Theorem:** Let $h_S$ be the optimal hyperplane for a sample $S$ and let $N_{SV}(S)$ be the number of support vectors defining $h_S$. Then,

$$\mathbb{E}_{S \sim D^m} [R(h_S)] \leq \mathbb{E}_{S \sim D^{m+1}} \left[ \frac{N_{SV}(S)}{m + 1} \right].$$

- **Proof:** Let $S \sim D^{m+1}$ be a sample linearly separable and let $x \in S$. If $h_{S-\{x\}}$ misclassifies $x$, then $x$ must be a SV for $h_S$. Thus,

$$\hat{R}_{1oo}(\text{opt.-hyp.}) \leq \frac{N_{SV}(S)}{m + 1}.$$
Notes

- Bound on expectation of error only, not the probability of error.

- Argument based on sparsity (number of support vectors). We will see later other arguments in support of the optimal hyperplanes based on the concept of margin.
This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees
Support Vector Machines

(Cortes and Vapnik, 1995)

- **Problem:** data often not linearly separable in practice. For any hyperplane, there exists $x_i$ such that

$$y_i [\mathbf{w} \cdot \mathbf{x}_i + b] \not\geq 1.$$ 

- **Idea:** relax constraints using slack variables $\xi_i \geq 0$

$$y_i [\mathbf{w} \cdot \mathbf{x}_i + b] \geq 1 - \xi_i.$$
Soft-Margin Hyperplanes

- Support vectors: points along the margin or outliers.
- Soft margin: $\rho = 1/\|\mathbf{w}\|$. 
Optimization Problem

(Cortes and Vapnik, 1995)

- Constrained optimization:

\[
\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i
\]

subject to \( y_i(w \cdot x_i + b) \geq 1 - \xi_i \land \xi_i \geq 0, \ i \in [1, m] \).

- Properties:
  - \( C \geq 0 \) trade-off parameter.
  - Convex optimization.
  - Unique solution.
Parameter $C$: trade-off between maximizing margin and minimizing training error. How do we determine $C$?

The general problem of determining a hyperplane minimizing the error on the training set is NP-complete (as a function of the dimension).

Other convex functions of the slack variables could be used: this choice and a similar one with squared slack variables lead to a convenient formulation and solution.
SVM - Equivalent Problem

- Optimization:

\[
\min_{\mathbf{w}, b} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{m} \left( 1 - y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \right)_+. 
\]

- Loss functions:
  - hinge loss:
    \[
    L(h(x), y) = (1 - yh(x))_+. 
    \]
  - quadratic hinge loss:
    \[
    L(h(x), y) = (1 - yh(x))^2_+. 
    \]
Hinge Loss

Hinge loss $\xi^1$

'Quadratic' hinge loss $\xi^2$

0/1 loss function
SVMs Equations

- **Lagrangian:** for all \( w, b, \alpha_i \geq 0, \beta_i \geq 0, \)

\[
L(w, b, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i [y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{m} \beta_i \xi_i.
\]

- **KKT conditions:**

\[
\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \iff w = \sum_{i=1}^{m} \alpha_i y_i x_i.
\]

\[
\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \iff \sum_{i=1}^{m} \alpha_i y_i = 0.
\]

\[
\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \iff \alpha_i + \beta_i = C.
\]

\[
\forall i \in [1, m], \alpha_i [y_i(w \cdot x_i + b) - 1 + \xi_i] = 0 \]

\[
\beta_i \xi_i = 0.
\]
Support Vectors

- **Complementarity conditions:**

  \[ \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] = 0 \implies \alpha_i = 0 \lor y_i (w \cdot x_i + b) = 1 - \xi_i. \]

- **Support vectors:** vectors \( x_i \) such that

  \[ \alpha_i \neq 0 \land y_i (w \cdot x_i + b) = 1 - \xi_i. \]

- **Note:** support vectors are not unique.
Moving to The Dual

- **Plugging in the expression of** \( w \text{ in } L \text{ gives:} \)

\[
L = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i x_i \right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i. \]

\[
- \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)
\]

- **Thus,**

\[
L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j).
\]

- **The condition** \( \beta_i \geq 0 \text{ is equivalent to } \alpha_i \leq C. \)
Dual Optimization Problem

Constrained optimization:

\[ \max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \]

subject to: \( 0 \leq \alpha_i \leq C \) \( \land \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m] \).

Solution:

\[ h(x) = \text{sgn} \left( \sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x) + b \right), \]

with \( b = y_i - \sum_{j=1}^{m} \alpha_j y_j (x_j \cdot x_i) \) for any \( x_i \) with \( 0 < \alpha_i < C \).
This Lecture

- Support Vector Machines - separable case
- Support Vector Machines - non-separable case
- Margin guarantees
High-Dimension

Learning guarantees: for hyperplanes in dimension $N$ with probability at least $1 - \delta$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2(N + 1) \log \frac{em}{N+1}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$  

- bound is uninformative for $N \gg m$.  
- but SVMs have been remarkably successful in high-dimension.  
- can we provide a theoretical justification?  
- analysis of underlying scoring function.
Confidence Margin

**Definition:** the confidence margin of a real-valued function $h$ at $(x, y) \in X \times Y$ is $\rho_h(x, y) = yh(x)$.

- interpreted as the hypothesis’ confidence in prediction.
- if correctly classified coincides with $|h(x)|$.
- relationship with geometric margin for linear functions $h : x \mapsto w \cdot x + b$: for $x$ in the sample,

\[
|\rho_h(x, y)| \geq \rho_{\text{geom}}\|w\|.
\]
Confidence Margin Loss

- **Definition:** for any confidence margin parameter $\rho > 0$ the $\rho$-margin loss function $\Phi_\rho$ is defined by

\[
\Phi_\rho(yh(x))
\]

- For a sample $S = (x_1, \ldots, x_m)$ and real-valued hypothesis $h$, the empirical margin loss is

\[
\hat{R}_\rho(h) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(y_i h(x_i)) \leq \frac{1}{m} \sum_{i=1}^{m} 1_{y_i h(x_i) < \rho}
\]
General Margin Bound

**Theorem**: Let $H$ be a set of real-valued functions. Fix $\rho > 0$. For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in H$:

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

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

**Proof**: Let $\tilde{H} = \{z = (x, y) \mapsto yh(x) : h \in H\}$. Consider the family of functions taking values in $[0, 1]$:

$$\tilde{\mathcal{H}} = \{\Phi_\rho \circ f : f \in \tilde{H}\}.$$
• By the theorem of Lecture 3, with probability at least $1 - \delta$, for all $g \in \widetilde{\mathcal{H}}$,

$$E[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2 \mathcal{R}_m(\widetilde{\mathcal{H}}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$ 

• Thus,

$$E[\Phi_\rho(yh(x))] \leq \hat{R}_\rho(h) + 2 \mathcal{R}_m(\Phi_\rho \circ \tilde{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$ 

• Since $\Phi_\rho$ is $\frac{1}{\rho}$-Lipschitz, by Talagrand’s lemma,

$$\mathcal{R}_m(\Phi_\rho \circ \tilde{H}) \leq \frac{1}{\rho} \mathcal{R}_m(\tilde{H}) = \frac{1}{\rho m} \mathbb{E} \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i y_i h(x_i) \right] = \frac{1}{\rho} \mathcal{R}_m(H).$$ 

• Since $1_{yh(x) < 0} \leq \Phi_\rho(yh(x))$, this shows the first statement, and similarly the second one.
Rademacher Complexity of Linear Hypotheses

**Theorem:** Let $S \subseteq \{x : \|x\| \leq R\}$ be a sample of size $m$ and let $H = \{x \mapsto w \cdot x : \|w\| \leq \Lambda\}$. Then,

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

**Proof:**

$$
\begin{align*}
\hat{\mathcal{R}}_S(H) &= \frac{1}{m} \mathbb{E} \left[ \sup_{\|w\| \leq \Lambda} \sum_{i=1}^{m} \sigma_i w \cdot x_i \right] = \frac{1}{m} \mathbb{E} \left[ \sup_{\|w\| \leq \Lambda} w \cdot \sum_{i=1}^{m} \sigma_i x_i \right] \\
&\leq \frac{\Lambda}{m} \mathbb{E} \left[ \left\| \sum_{i=1}^{m} \sigma_i x_i \right\| \right] \leq \frac{\Lambda}{m} \left[ \mathbb{E} \left[ \left\| \sum_{i=1}^{m} \sigma_i x_i \right\|^2 \right] \right]^{1/2} \\
&\leq \frac{\Lambda}{m} \left[ \mathbb{E} \left[ \sum_{i=1}^{m} \|x_i\|^2 \right] \right]^{1/2} \leq \frac{\Lambda \sqrt{mR^2}}{m} = \sqrt{\frac{R^2 \Lambda^2}{m}}.
\end{align*}
$$
Corollary: Let $\rho > 0$ and $H = \{x \mapsto w \cdot x : \|w\| \leq \Lambda\}$. Assume that $X \subseteq \{x : \|x\| \leq R\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \hat{R}_\rho(h) + 2\sqrt{\frac{R^2 \Lambda^2 / \rho^2}{m}} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$ 

Proof: Follows directly general margin bound and bound on $\hat{R}_S(H)$ for linear classifiers.

- Finer relative deviation margin bounds (Cortes, MM, Suresh; ICML 2021).
High-Dimensional Feature Space

Observations:

- generalization bound does not depend on the dimension but on the margin.
- this suggests seeking a large-margin hyperplane in a higher-dimensional feature space.

Computational problems:

- taking dot products in a high-dimensional feature space can be very costly.
- solution based on kernels (next lecture).
References


Appendix
Saddle Point

Let \((w^*, b^*, \alpha^*)\) be the saddle point of the Langrangian. Multiplying both sides of the equation giving \(b^*\) by \(\alpha_i^* y_i\) and taking the sum leads to:

\[
\sum_{i=1}^{m} \alpha_i^* y_i b = \sum_{i=1}^{m} \alpha_i^* y_i^2 - \sum_{i,j=1}^{m} \alpha_i^* \alpha_j^* y_i y_j (x_i \cdot x_j).
\]

Using \(y_i^2 = 1, \sum_{i=1}^{m} \alpha_i^* y_i = 0,\) and \(w^* = \sum_{i=1}^{m} \alpha_i^* y_i x_i\) yields

\[
0 = \sum_{i=1}^{m} \alpha_i^* - \|w^*\|^2.
\]

Thus, the margin is also given by:

\[
\rho^2 = \frac{1}{\|w^*\|^2} = \frac{1}{\|\alpha^*\|_1}.
\]
Talagrand's Contraction Lemma

(Ledoux and Talagrand, 1991; pp. 112-114)

Theorem: Let \( \Phi : \mathbb{R} \rightarrow \mathbb{R} \) be an \( L \)-Lipschitz function. Then, for any hypothesis set \( H \) of real-valued functions,

\[
\hat{\mathcal{R}}_S(\Phi \circ H) \leq L \hat{\mathcal{R}}_S(H).
\]

Proof: fix sample \( S = (x_1, \ldots, x_m) \). By definition,

\[
\mathcal{R}_S(\Phi \circ H) = \frac{1}{m} \mathbb{E} \left[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i(\Phi \circ h)(x_i) \right]
\]

\[
= \frac{1}{m} \mathbb{E} \sum_{\sigma_1, \ldots, \sigma_{m-1}} \left[ \mathbb{E}_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m(\Phi \circ h)(x_m) \right] \right],
\]

with \( u_{m-1}(h) = \sum_{i=1}^{m-1} \sigma_i(\Phi \circ h)(x_i) \).
Talagrand's Contraction Lemma

Now, assuming that the suprema are reached, there exist $h_1, h_2 \in H$ such that

\[
\mathbb{E}_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \right]
\]

\[
= \frac{1}{2} [u_{m-1}(h_1) + (\Phi \circ h_1)(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - (\Phi \circ h_2)(x_m)]
\]

\[
\leq \frac{1}{2} [u_{m-1}(h_1) + u_{m-1}(h_2) + sL(h_1(x_m) - h_2(x_m))]
\]

\[
= \frac{1}{2} [u_{m-1}(h_1) + sLh_1(x_m)] + \frac{1}{2} [u_{m-1}(h_2) - sLh_2(x_m)]
\]

\[
\leq \mathbb{E}_{\sigma_m} \left[ \sup_{h \in H} u_{m-1}(h) + \sigma_m Lh(x_m) \right],
\]

where $s = \text{sgn}(h_1(x_m) - h_2(x_m))$. 
Talagrand’s Contraction Lemma

- When the suprema are not reached, the same can be shown modulo $\epsilon$, followed by $\epsilon \to 0$.

- Proceeding similarly for other $\sigma_i$s directly leads to the result.
VC Dimension of Canonical Hyperplanes

Theorem: Let $S \subseteq \{ x : \| x \| \leq R \}$. Then, the VC dimension $d$ of the set of canonical hyperplanes
\[
\{ x \mapsto \text{sgn}(w \cdot x) : \min_{x \in S} |w \cdot x| = 1 \land \|w\| \leq \Lambda \}
\]
verifies
\[
d \leq R^2 \Lambda^2.
\]

Proof: Let $\{x_1, \ldots, x_d\}$ be a set fully shattered. Then, for all $y \in \{-1, +1\}^d$, there exists $w$ such
\[
\forall i \in [1, d], 1 \leq y_i (w \cdot x_i).
\]
• Summing up the inequalities gives

\[ d \leq \mathbf{w} \cdot \sum_{i=1}^{d} y_i \mathbf{x}_i \leq \| \mathbf{w} \| \| \sum_{i=1}^{d} y_i \mathbf{x}_i \| \leq \Lambda \| \sum_{i=1}^{d} y_i \mathbf{x}_i \|. \]

• Taking the expectation over \( y \sim U \) (uniform) yields

\[
\begin{align*}
  d &\leq \Lambda \mathbb{E}_{y \sim U} \left[ \| \sum_{i=1}^{d} y_i \mathbf{x}_i \| \right] \leq \Lambda \left[ \mathbb{E}_{y \sim U} \left[ \| \sum_{i=1}^{d} y_i \mathbf{x}_i \|^2 \right] \right]^{1/2} \text{ (Jensen’s ineq.)} \\
  &= \Lambda \left[ \sum_{i,j=1}^{d} \mathbb{E}[y_i y_j] (\mathbf{x}_i \cdot \mathbf{x}_j) \right]^{1/2} \\
  &= \Lambda \left[ \sum_{i=1}^{d} (\mathbf{x}_i \cdot \mathbf{x}_i) \right]^{1/2} \leq \Lambda [dR^2]^{1/2} = \Lambda R \sqrt{d}.
\end{align*}
\]

• Thus, \( \sqrt{d} \leq \Lambda R \).