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,

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

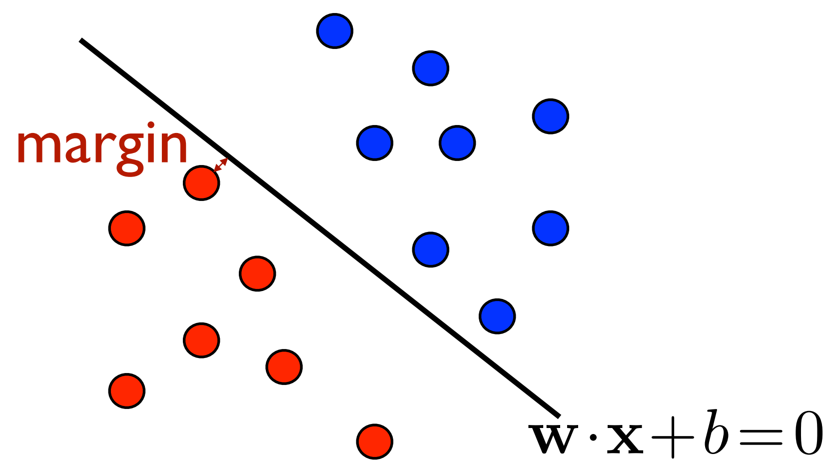
- Problem: find hypothesis $h : X \mapsto \{-1, +1\}$ in H (classifier) with small generalization error $R(h)$.
  - choice of hypothesis set H: learning guarantees of previous lecture.

    → 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

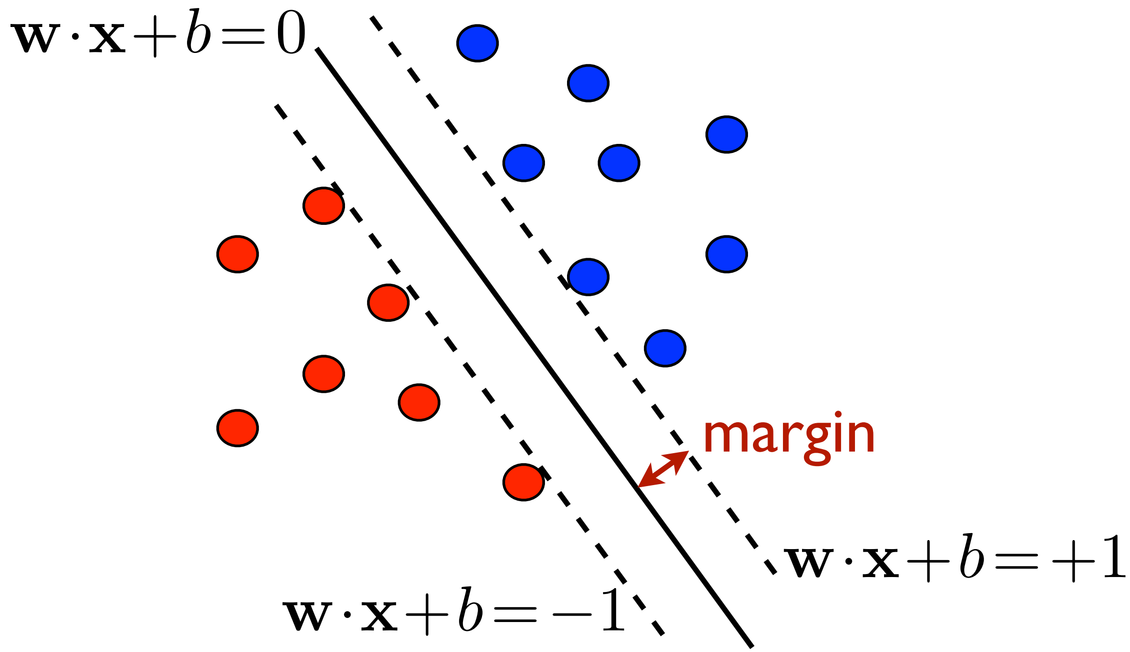
Linear Separation



- classifiers: $H = \{\mathbf{x} \mapsto \operatorname{sgn}(\mathbf{w}\cdot\mathbf{x}+b) : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$.
- geometric margin: $\rho = \min_{i\in[1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i+b|}{\|\mathbf{w}\|}$.
- which separating hyperplane?

Optimal Hyperplane: Max. Margin

(Vapnik and Chervonenkis, 1965)



$$\rho = \max_{\mathbf{w},b\,:\, y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 0} \ \min_{i\in[1,m]} \frac{|\mathbf{w}\cdot\mathbf{x}_i + b|}{\|\mathbf{w}\|}.$$

Maximum Margin

$$\begin{aligned}
\rho &= \max_{\mathbf{w}, b\colon y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 0} \ \min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|} \\
&= \max_{\substack{\mathbf{w}, b\colon y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 0 \\ \min_{i \in [1,m]} |\mathbf{w}\cdot\mathbf{x}_i + b| = 1}} \ \min_{i \in [1,m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i + b|}{\|\mathbf{w}\|} \qquad \text{(scale-invariance)} \\
&= \max_{\substack{\mathbf{w}, b\colon y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 0 \\ \min_{i \in [1,m]} |\mathbf{w}\cdot\mathbf{x}_i + b| = 1}} \ \frac{1}{\|\mathbf{w}\|} \\
&= \max_{\mathbf{w}, b\colon y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1} \ \frac{1}{\|\mathbf{w}\|}. \qquad \text{(min. reached)}
\end{aligned}$$

Optimization Problem

- Constrained optimization:

$$\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to} \ \ y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1, i \in [1, m].$$

- Properties:
  - Convex optimization.
  - Unique solution for linearly separable sample.

Optimal Hyperplane Equations

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

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

- **KKT conditions**:

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \iff \boxed{\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i.}$$

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

$$\boxed{\forall i \in [1, m], \;\; \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0.}$$

Support Vectors

- Complementarity conditions:

$$\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0 \implies \alpha_i = 0 \vee y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1.$$

- Support vectors: vectors $\mathbf{x}_i$ such that

$$\alpha_i \neq 0 \wedge y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1.$$

  - Note: support vectors are not unique.

Moving to The Dual

- Plugging in the expression of $\mathbf{w}$ in L gives:

$$L = \underbrace{\frac{1}{2}\left\|\sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i\right\|^2 - \sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)}_{-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (\mathbf{x}_i\cdot\mathbf{x}_j)} - \underbrace{\sum_{i=1}^{m}\alpha_i y_i b}_{0} + \sum_{i=1}^{m}\alpha_i.$$

- Thus,

$$L = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j).$$

Equivalent Dual Opt. Problem

- Constrained optimization:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subject to: } \alpha_i \geq 0 \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].$$

- Solution:

$$h(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b \Big),$$

with $b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i)$ for any SV $\mathbf{x}_i$.

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:

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

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

$$\boxed{\mathop{\mathrm{E}}_{S \sim D^m}[\widehat{R}_{\text{loo}}(L)]} = \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathrm{E}}_{S}[1_{h_{S-\{x_i\}}(x_i) \neq f(x_i)}] = \mathop{\mathrm{E}}_{S}[1_{h_{S-\{x\}}(x) \neq f(x)}]$$

$$= \mathop{\mathrm{E}}_{S' \sim D^{m-1}}[\mathop{\mathrm{E}}_{x \sim D}[1_{h_{S'}(x) \neq f(x)}]] \boxed{= \mathop{\mathrm{E}}_{S' \sim D^{m-1}}[R(h_{S'})]}.$$

Leave-One-Out Analysis

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

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

- 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,

$$\widehat{R}_{\rm loo}(\text{opt.-hyp.}) \leq \frac{N_{\rm SV}(S)}{m+1}.$$

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)

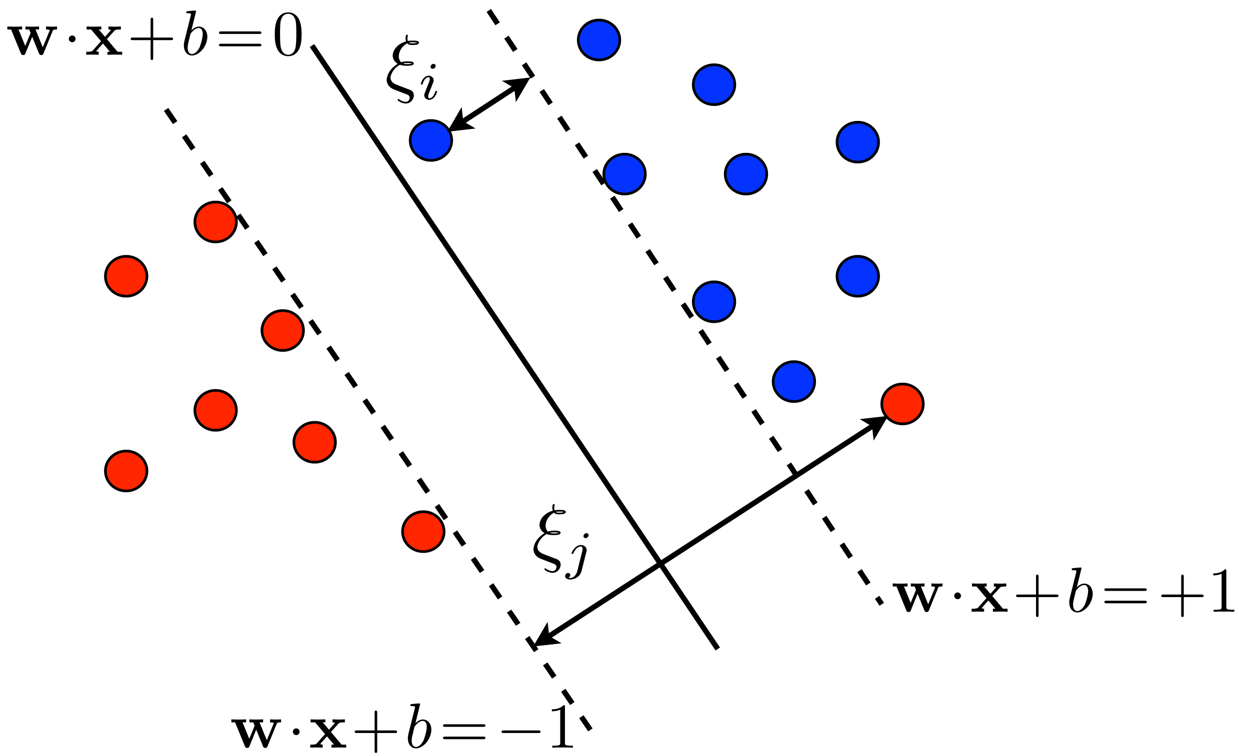
- Problem: data often not linearly separable in practice. For any hyperplane, there exists $\mathbf{x}_i$ such that

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

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

$$y_i \left[\mathbf{w} \cdot \mathbf{x}_i + b\right] \geq 1 - \xi_i.$$

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:

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

$$\text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i \;\wedge\; \xi_i \geq 0, i \in [1, m].$$

- Properties:
  - $C \geq 0$ trade-off parameter.
  - Convex optimization.
  - Unique solution.

Notes

- 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:

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

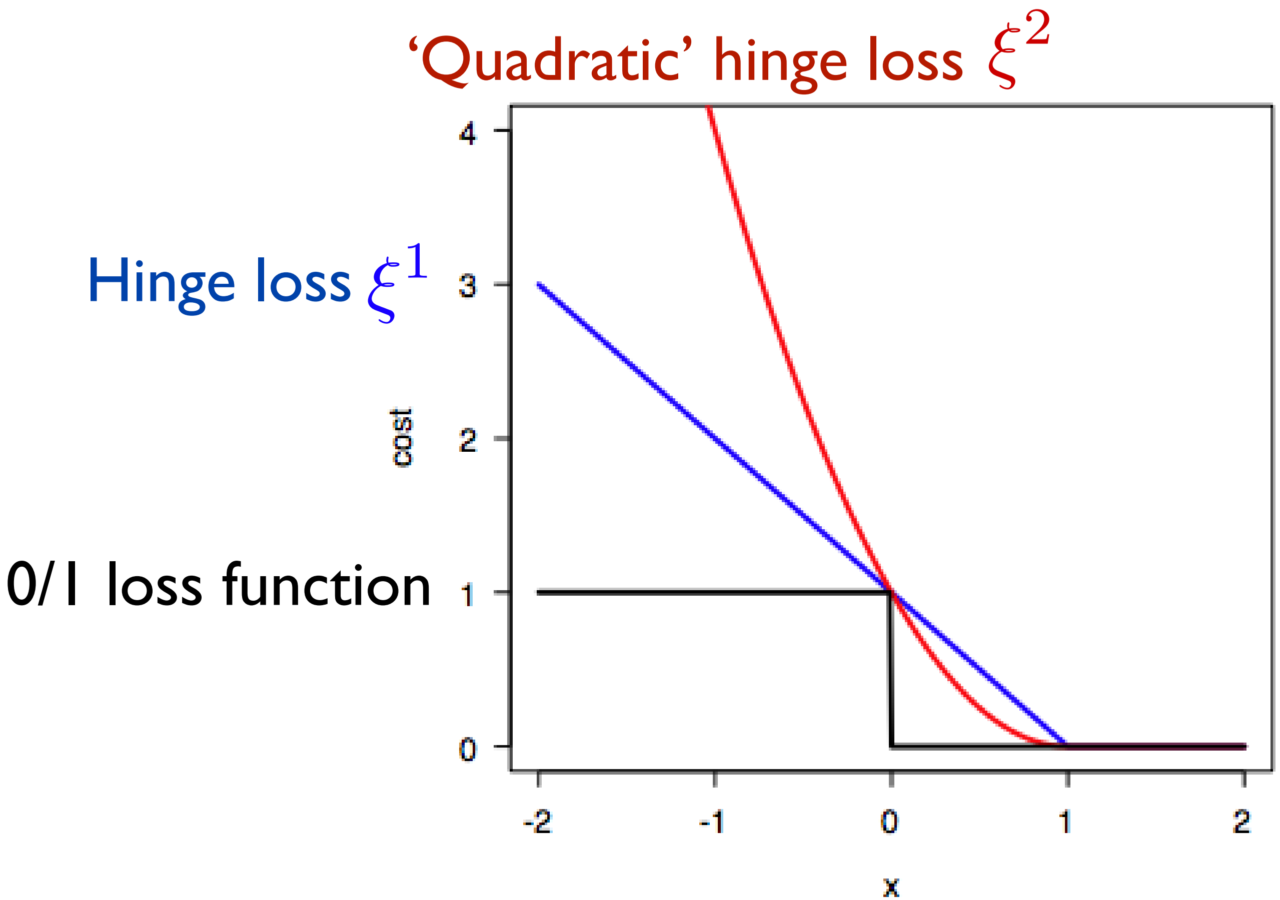
- Loss functions:
  - hinge loss:

$$L(h(x), y) = (1 - yh(x))_+.$$

  - quadratic hinge loss:

$$L(h(x), y) = (1 - yh(x))_+^2.$$

Hinge Loss



‘Quadratic’ hinge loss ξ^2
Hinge loss ξ^1
0/1 loss function
cost
x
4
3
2
1
0
-2
-1
0
1
2

SVMs Equations

- **Lagrangian**: for all $\mathbf{w}, b, \alpha_i \geq 0, \beta_i \geq 0,$

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

- **KKT conditions**:

$$\nabla_w L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i = 0 \iff \boxed{\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i.}$$

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

$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \iff \boxed{\alpha_i + \beta_i = C.}$$

$$\boxed{\begin{aligned} \forall i \in [1, m],\ \alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i] &= 0 \\ \beta_i \xi_i &= 0. \end{aligned}}$$

Support Vectors

- Complementarity conditions:

$$\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i] = 0 \implies \alpha_i = 0 \vee y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i.$$

- Support vectors: vectors $\mathbf{x}_i$ such that

$$\alpha_i \neq 0 \wedge y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i.$$

  - Note: support vectors are not unique.

Moving to The Dual

- Plugging in the expression of w in L gives:

$$L = \underbrace{\frac{1}{2}\left\|\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i\right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)}_{-\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)} - \underbrace{\sum_{i=1}^{m} \alpha_i y_i b}_{0} + \sum_{i=1}^{m} \alpha_i.$$

- Thus,

$$L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j).$$

- The condition $\beta_i \geq 0$ is equivalent to $\alpha_i \leq C$.

Dual Optimization Problem

- Constrained optimization:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subject to: } 0 \leq \alpha_i \boxed{\leq C} \wedge \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [1, m].$$

- Solution:

$$h(x) = \operatorname{sgn}\Big(\sum_{i=1}^{m} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b\Big),$$

with $b = y_i - \sum_{j=1}^{m} \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i)$ for any $\mathbf{x}_i$ with $0 < \alpha_i \boxed{< C.}$

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 $1-\delta$,

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

  - 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.

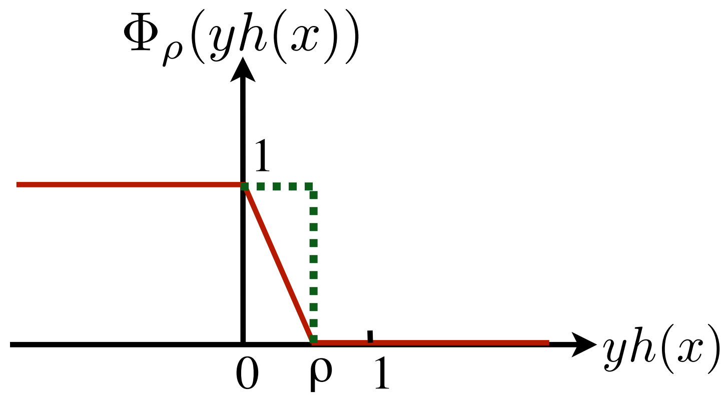
Confidence Margin

- **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 $h\colon \mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} + b$: for x in the sample,

$$|\rho_h(x, y)| \geq \rho_{\text{geom}} \|\mathbf{w}\|.$$

Confidence Margin Loss

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



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

$$\widehat{R}_\rho(h) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(y_i h(x_i)) \leq \boxed{\frac{1}{m} \sum_{i=1}^{m} 1_{y_i h(x_i) < \rho}.}$$

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 $h \in H$:

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

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

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

$$\widetilde{\mathcal{H}} = \{\Phi_\rho \circ f \colon f \in \widetilde{H}\}.$$

- By the theorem of Lecture 3, with probability at least $1-\delta$, for all $g\in\widetilde{\mathcal{H}}$,
$$\mathrm{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathfrak{R}_m(\widetilde{\mathcal{H}}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$
- Thus,
$$\mathrm{E}[\Phi_\rho(yh(x))] \leq \widehat{R}_\rho(h) + 2\mathfrak{R}_m\big(\Phi_\rho \circ \widetilde{H}\big) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$
- Since Φ_ρ is $\frac{1}{\rho}$ - Lipschitz, by Talagrand's lemma,
$$\mathfrak{R}_m\big(\Phi_\rho \circ \widetilde{H}\big) \leq \frac{1}{\rho}\mathfrak{R}_m(\widetilde{H}) = \frac{1}{\rho m}\mathop{\mathrm{E}}_{\sigma,S}\Big[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i y_i h(x_i)\Big] = \frac{1}{\rho}\mathfrak{R}_m\big(H\big).$$
- 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 $S \subseteq \{x : \|\mathbf{x}\| \leq R\}$ be a sample of size m and let $H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} : \|\mathbf{w}\| \leq \Lambda\}$. Then,

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

- **Proof**:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H) &= \frac{1}{m} \operatorname*{E}_{\sigma} \left[\sup_{\|\mathbf{w}\| \leq \Lambda} \sum_{i=1}^{m} \sigma_i \mathbf{w} \cdot \mathbf{x}_i \right] = \frac{1}{m} \operatorname*{E}_{\sigma} \left[\sup_{\|\mathbf{w}\| \leq \Lambda} \mathbf{w} \cdot \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right] \\
&\leq \frac{\Lambda}{m} \operatorname*{E}_{\sigma} \left[\left\| \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right\| \right] \leq \frac{\Lambda}{m} \left[\operatorname*{E}_{\sigma} \left[\left\| \sum_{i=1}^{m} \sigma_i \mathbf{x}_i \right\|^2 \right] \right]^{1/2} \\
&\leq \frac{\Lambda}{m} \left[\operatorname*{E}_{\sigma} \left[\sum_{i=1}^{m} \|\mathbf{x}_i\|^2 \right] \right]^{1/2} \leq \frac{\Lambda \sqrt{m R^2}}{m} = \sqrt{\frac{R^2 \Lambda^2}{m}}.
\end{aligned}$$

Margin Bound - Linear Classifiers

- **Corollary**: Let $\rho > 0$ and $H = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \colon \|\mathbf{w}\| \leq \Lambda\}$. Assume that $X \subseteq \{\mathbf{x} \colon \|\mathbf{x}\| \leq R\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

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

- **Proof**: Follows directly general margin bound and bound on $\widehat{\mathfrak{R}}_S(H)$ 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

- Corinna Cortes and Vladimir Vapnik, Support-Vector Networks, *Machine Learning*, 20, 1995.
- Koltchinskii, Vladimir and Panchenko, Dmitry. Empirical margin distributions and bounding the generalization error of combined classifiers. The Annals of Statistics, 30(1), 2002.
- Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer, New York.
- Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Basederlin, 1982.
- Vladimir N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1995.
- Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

Appendix

Saddle Point

- Let $(\mathbf{w}^*, b^*, \alpha^*)$ 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:

$$\sum_{i=1}^{m} \alpha_i^* y_i b = \sum_{i=1}^{m} \alpha_i^* y_i^2 - \sum_{i,j=1}^{m} \alpha_i^* \alpha_j^* y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j).$$

- Using $y_i^2 = 1, \sum_{i=1}^{m} \alpha_i^* y_i = 0$, and $\mathbf{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \mathbf{x}_i$ yields

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

- Thus, the margin is also given by:

$$\boxed{\rho^2 = \frac{1}{\|\mathbf{w}^*\|_2^2} = \frac{1}{\|\alpha^*\|_1}.}$$

Talagrand's Contraction Lemma

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

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

$$\widehat{\mathfrak{R}}_S(\Phi \circ H) \leq L\, \widehat{\mathfrak{R}}_S(H).$$

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

$$\begin{aligned}
\mathfrak{R}_S(\Phi \circ H) &= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i (\Phi \circ h)(x_i) \right] \\
&= \frac{1}{m} \operatorname*{E}_{\sigma_1, \ldots, \sigma_{m-1}} \left[\operatorname*{E}_{\sigma_m} \left[\sup_{h \in H} u_{m-1}(h) + \sigma_m (\Phi \circ h)(x_m) \right] \right],
\end{aligned}$$

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

$$\begin{aligned}
&\mathop{\mathrm{E}}_{\sigma_m}\Big[\sup_{h\in H} u_{m-1}(h) + \sigma_m(\Phi\circ h)(x_m)\Big]\Big] \\
&= \frac{1}{2}[u_{m-1}(h_1) + (\Phi\circ h_1)(x_m)] + \frac{1}{2}[u_{m-1}(h_2) - (\Phi\circ h_2)(x_m)] \\
&\leq \frac{1}{2}[u_{m-1}(h_1) + u_{m-1}(h_2) + sL(h_1(x_m) - h_2(x_m))] \\
&= \frac{1}{2}[u_{m-1}(h_1) + sLh_1(x_m)] + \frac{1}{2}[u_{m-1}(h_2) - sLh_2(x_m)] \\
&\leq \mathop{\mathrm{E}}_{\sigma_m}\Big[\sup_{h\in H} u_{m-1}(h) + \sigma_m L h(x_m)\Big],
\end{aligned}$$

where $s = \mathrm{sgn}(h_1(x_m) - h_2(x_m))$.

Talagrand's Contraction Lemma

- When the suprema are not reached, the same can be shown modulo ϵ, followed by $\epsilon \to 0$.
- Proceeding similarly for other σ_is directly leads to the result.

VC Dimension of Canonical Hyperplanes

- **Theorem**: Let $S \subseteq \{\mathbf{x} : \|\mathbf{x}\| \leq R\}$. Then, the VC dimension d of the set of canonical hyperplanes $\{x \mapsto \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}) \colon \min_{x \in S} |\mathbf{w} \cdot \mathbf{x}| = 1 \wedge \|\mathbf{w}\| \leq \Lambda\}$ verifies

$$d \leq R^2 \Lambda^2.$$

- **Proof**: Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$ be a set fully shattered. Then, for all $\mathbf{y} \in \{-1, +1\}^d$, there exists $\mathbf{w}$ such

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

- 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 $\mathbf{y} \sim U$ (uniform) yields

$$d \leq \Lambda \mathop{\mathrm{E}}_{\mathbf{y} \sim U}[\|\sum_{i=1}^{d} y_i \mathbf{x}_i\|] \leq \Lambda \Big[\mathop{\mathrm{E}}_{\mathbf{y} \sim U}[\|\sum_{i=1}^{d} y_i \mathbf{x}_i\|^2] \Big]^{1/2} \text{ (Jensen's ineq.)}$$

$$= \Lambda \Big[\sum_{i,j=1}^{d} \mathrm{E}[y_i y_j](\mathbf{x}_i \cdot \mathbf{x}_j) \Big]^{1/2}$$

$$= \Lambda \Big[\sum_{i=1}^{d} (\mathbf{x}_i \cdot \mathbf{x}_i) \Big]^{1/2} \leq \Lambda \big[dR^2 \big]^{1/2} = \Lambda R \sqrt{d}.$$

- Thus, $\sqrt{d} \leq \Lambda R$.