Computational and Statistical Learning Theory

TTIC 31120

Prof. Nati Srebro

Lecture 11:

SVMs

ℓ_1 Regularization

Regularized Linear Prediction

$$\mathcal{H}_B = \{x \mapsto \langle w, \phi(x)\rangle \mid w \in \mathbb{R}^d, \|w\|_2 \le B\}$$

$$\mathcal{R}_S(\mathcal{H}_B) = \sqrt{\frac{B^2\overline{R^2}(S)}{m}}, \qquad \overline{R^2}(S) = \frac{1}{m}\sum_i \|\phi(x_i)\|_2^2$$

- SVMs:
  - Hypothesis class: ℓ_2-regularized linear predictors $\mathcal{H}_B$
  - Hinge loss $loss^{hinge}(\hat{y}; y) = [1 - \hat{y}y]_+$

- ℓ_2 control + Lipschtiz loss ➔ dimension independent statistical control

  ➔ can use very high (even infinite) dimensional feature spaces

- How do we cope with high dimensional $\phi(x)$ computationally?

Representer Theorem

$$\widehat{w} = \arg\min_{w} L_S(w) + \lambda\|w\|^2$$

Representer Theorem: for any

$$f(w, S) = f(\langle w, \phi(x_1)\rangle, \langle w, \phi(x_1)\rangle, \ldots, \langle w, \phi(x_m)\rangle, S)$$

(i.e. $f(w, S)$ depends on w only through inner products with $\phi(x_i)$), there exists a solution to:

$$\arg\min_{w} f(w) + \lambda\|w\|^2$$

of the form: $w = \sum_i \alpha_i \phi(x_i)$

Proof: let w be a solution, and consider its projection $w_{\|}$ onto $span\big(\phi(x_1), .., \phi(x_m)\big)$. We have that $\langle w_{\|}, \phi(x_i)\rangle = \langle w, \phi(x_i)\rangle$ and so $f(w, S) = f(w_{\|}, S)$ while $\|w_{\|}\| \leq \|w\|$.

- This holds for any loss function!

Kernalization

$$L_S(w) + \lambda\|w\|^2 = \frac{1}{m}\sum_i loss(\langle w, \phi(x_i)\rangle, y_i) + \lambda\|w\|^2$$

$$w = \sum_j \alpha_j \phi(x_j)$$

$$= \frac{1}{m}\sum_i loss\left(\langle \sum_j \alpha_j \phi(x_j), \phi(x_i)\rangle, y_i\right) + \lambda\langle \sum_j \alpha_j \phi(x_j), \sum_j \alpha_j \phi(x_j)\rangle$$

$$= \frac{1}{m}\sum_i loss\left(\sum_j \alpha_j \underbrace{\langle \phi(x_i), \phi(x_j)\rangle}_{K(x,x')=\langle x,x'\rangle}, y_i\right) + \lambda \sum_{ij} \alpha_i\alpha_j \underbrace{\langle \phi(x_i), \phi(x_j)\rangle}_{K(x,x')=\langle x,x'\rangle}$$

Solve:

$$\arg\min_\alpha \frac{1}{m}\sum_i loss\left(\sum_j \alpha_j K(x_i, x_j), y_i\right) + \lambda \sum_{ij} \alpha_i\alpha_j K(x_i, x_j)$$

Predict using:

$$x \mapsto \sum_j \alpha_j K(x_j, x)$$

- Holds for any loss function!
  (or any objective that depends only on $\langle w, \phi(x)\rangle$ and $\|w\|$)

Complexity In Terms of Kernel

- Recall sample complexity scales as:

$$\propto \underbrace{\|w\|_2^2}_{\sum_{ij}\alpha_i\alpha_j K(x_i,x_j)} \cdot \mathbb{E}[\underbrace{\|\phi(x)\|_2^2}_{K(x,x)}]$$

- When considering a kernel, norm should always be measured relative to $K(x,x)$

- On a sample with Gram matrix $K_{ij} = K(x_i, x_j)$, complexity of a predictor $x \to \sum_i \alpha_i K(x_i, x)$ can be measured as:

$$(\alpha^T K \alpha) \cdot trace(K)$$

Support Vector Machines

- $\|w\|_2$ control and any Lipschitz loss
  ➔ Generalization
- $\|w\|_2$ and any loss (or any objective of $\langle w, \phi(x) \rangle$)
  ➔ Kernalization
- Any convex loss
  ➔ Convex ERM, tractability

- What's special about the hinge loss?
  - Tightest convex Lipschitz relaxation of ramp loss
  - Sparsity of α (ie notion of *support* vectors)

- **Theorem**: Let $w^* = \arg\min L_S^{hinge}(w) + \lambda\|w\|^2$. Then $w^* = \sum \alpha_i \phi(x_i)$ for $\alpha_i \in \mathbb{R}$ s.t. $y_i \langle w, \phi(x_i) \rangle > 1 \Rightarrow \alpha_i = 0$

Proof: changing $\phi(x_i)$ does not change the optimum

(Using duality, can show $0 \leq y_i \alpha_i \leq {}^2/_\lambda$ with $y_i \langle w, \phi(x_i) \rangle < 1 \Rightarrow \alpha_i = {}^2/_\lambda$)

SVMs: Summary

- If we want to use lots of features (even infinite), two problems:
  - Generalization
  - Computation

- Solutions:
  - Dimension-independent scale sensitive capacity control in terms of $\|w\|_2$ ensures generalization even in infinite dimensions
  - Kernalization allows implicitly working with infinite number of features, as long as we can calculate inner products efficiently
  - Bonus: with hinge loss, representation is sparse and depends only on violations and points on the margin

A different regularizer: ℓ_1

$$\mathcal{H}_B = \{x \mapsto \langle w, \phi(x) \rangle \mid w \in \mathbb{R}^d, \|w\|_1 \leq B \}$$

Radamacher Complexity of Convex Hull

$$conv(\mathcal{F}) = \left\{ g = \sum_{f \in \mathcal{F}} \alpha_f \cdot f \;\middle|\; \alpha_f > 0, \sum_f \alpha_f = 1 \right\}$$

- **Theorem**: $\mathcal{R}_S\big(conv(\mathcal{F})\big) = \mathcal{R}_S(\mathcal{F})$
- **Proof**: $\mathcal{R}_S\big(conv(\mathcal{F})\big) = \mathbb{E}_\xi \left[\sup_{g \in conv(\mathcal{F})} \frac{1}{m} \sum_i \xi_i g(x_i) \right]$

$$= \mathbb{E}_\xi \left[\sup_{\alpha_f \geq 0, \sum \alpha_f = 1} \frac{1}{m} \sum_i \xi_i \sum_{f \in \mathcal{F}} \alpha_f \cdot f(x_i) \right]$$

$$= \mathbb{E}_\xi \left[\sup_{\alpha_f \geq 0, \sum \alpha_f = 1} \sum_{f \in \mathcal{F}} \alpha_f \cdot \left(\frac{1}{m} \sum_i \xi_i f(x_i) \right) \right]$$

$$= \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_i \xi_i f(x_i) \right] = \mathcal{R}_S(\mathcal{F})$$

A different regularizer: ℓ_1

$$\mathcal{H}_B = \{x \mapsto \langle w, \phi(x)\rangle \mid w \in \mathbb{R}^d, \|w\|_1 \le B\}$$

- **Conclusion**:

$$\mathcal{R}_S(\mathcal{H}_B) = \mathcal{R}_S(B \cdot \mathcal{H}_1) = B \cdot \mathcal{R}_S(\mathcal{H}_1)$$

$$= B \cdot \mathcal{R}_S(conv(\{x \mapsto \phi(x)[j] \mid j = 1, \dots, d\}) \le B(\sup|\phi(x_i)[j]|)\sqrt{\frac{\log d}{m}}$$

$$= \sqrt{\frac{B^2 \sup\|\phi(x_i)\|_\infty^2 \log d}{m}}$$

- In fact, even if $d = \infty$:

$$\mathcal{R}_S(\mathcal{H}_B) = O\left(\sqrt{\frac{B^2 \sup\|\phi(x_i)\|_\infty^2 \,\mathrm{VCdim}(\{x \mapsto \phi(x)[j]\})}{m}}\right)$$

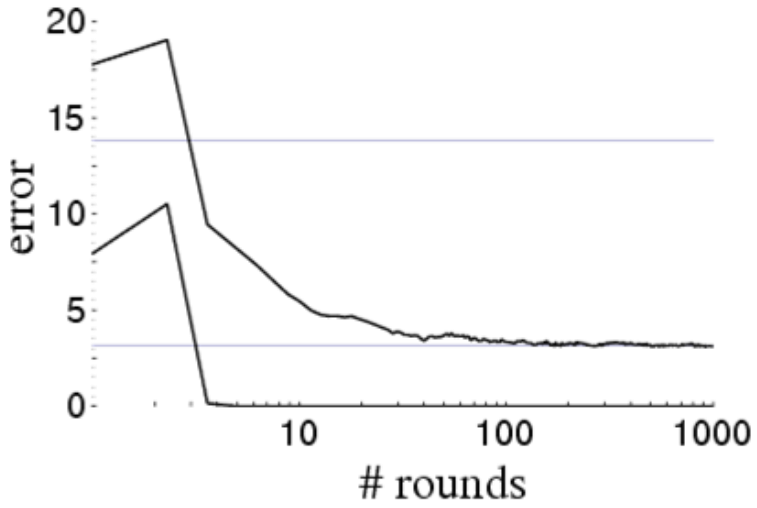
Boosting

- $x \mapsto \phi(x)[h]$, weak binary predictor $h(x) = \pm 1$

  ➔ $\|\phi(x)\|_\infty = 1$

- Can ensure generalization in terms of $\|w\|_1 = \sum \alpha_t$

- What's going on after $L_S^{01}(w) = 0$?



Boosting and the ℓ_1 margin

- What happens to $\|w_T\|_1$ as T increases?

  $\|w_T\|_1 = \sum_{t=1}^{T} \alpha_t = \sum_{t=1}^{T} \frac{1}{2} \log\left(\frac{1}{\epsilon_t} - 1\right) \geq \frac{T}{2} \log \frac{1+2\gamma}{1-2\gamma}$

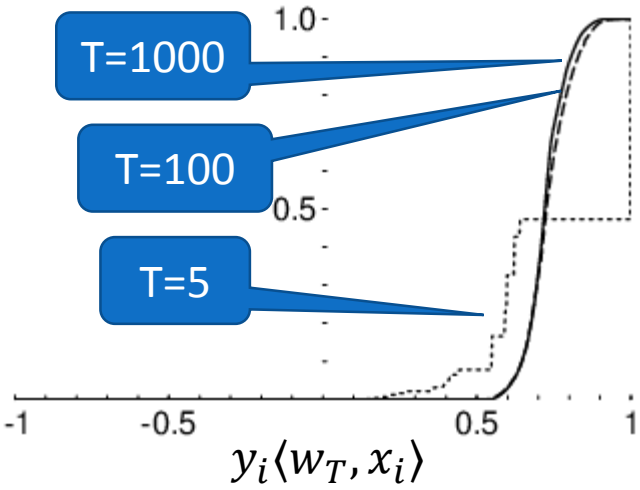
  ➔ and so $\|w_T\|_1$ increases!

- Consider $margin(w) = \min_{(x_i,y_i)\in S} y_i \langle w_T, x_i \rangle$

  $L_S^{01}(w) = 0 \Leftrightarrow margin(w) > 0$

  $L_{S(w)}^{mrg}\left(\frac{w}{\gamma}\right) = 0 \Leftrightarrow margin(w) \geq \gamma$

  Complexity parameter: $\left\|\frac{w}{\gamma}\right\|_1 = \frac{\|w\|_1}{magin(w)}$



Exp-Loss, Margin and AdaBoost

- Recall: $L_S^{\exp}(h) = \frac{1}{m}\sum_i e^{-y_i h(x_i)} \leq e^{-2\gamma^2 T}$
- And so:

$$\forall_i e^{-y_i h(x_i)} \leq m L_S^{exp}(h) \leq m e^{-2\gamma^2 T}$$

$$\rightarrow y_i h(x_i) \geq 2\gamma^2 T - \log(m)$$

- Contrast with $\|w_T\|_1 \geq \frac{T}{2}\log\frac{1+2\gamma}{1-2\gamma} \approx 2\gamma T$

- Theorem: After $T = \frac{3\log 2m}{4\gamma^4}$ iterations, $L_S^{mrg}\left(\frac{1}{\gamma}\frac{w_T}{\|w_T\|_1}\right) = 0$

$$\rightarrow \frac{margin(w)}{\|w\|_1} \geq \gamma$$

Weak Learning and the ℓ_1 Margin

- Consider a base class $\mathcal{B} = \{f: \mathcal{X} \to \pm 1\}$ and the corresponding feature map $\phi: \mathcal{X} \to \mathbb{R}^{\mathcal{B}}$ defined as $\phi(x)[f] = h[f]$.

- Goal: relate weak learnability using predictors in $\mathcal{B}$ to ℓ_1-margin using $\phi(x)$

- Weak learnability:
  - $h: \mathcal{X} \to \pm 1$ is γ-weakly learnable using $\mathcal{B}$ if for any distribution $\mathcal{D}(\mathcal{X})$, there exists $f \in \mathcal{B}$ s.t. $\Pr_{x \sim \mathcal{D}}[f(x) = h(x)] \geq \frac{1}{2} + \frac{\gamma}{2}$
  - $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ is γ-weakly learnable using $\mathcal{B}$ if all $h \in \mathcal{H}$ are γ-weakly learnable

- Assume that $\mathcal{B}$ is symmetric, i.e. for any $f \in \mathcal{B}$, also $-f \in \mathcal{B}$
  - This allows us to consider only $w \geq 0$, and so $\|w\|_1 = \sum w[f]$
  - If $w[f] < 0$, instead use $w[-f] > 0$

  (without assuming $\mathcal{B}$ is symmetric, we will need to talk about margin attainable only with $w \geq 0$)