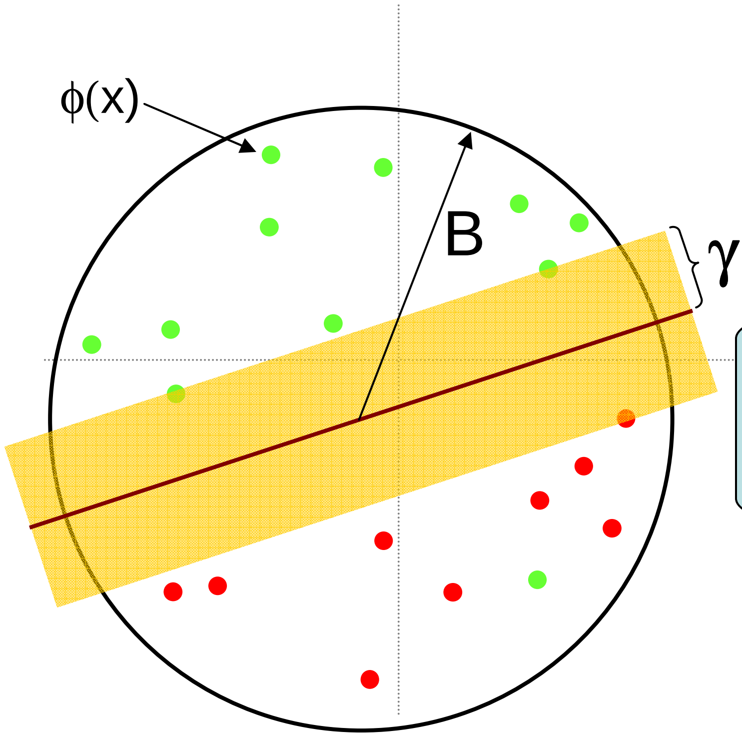
Learning Bounds for Support Vector Machines with Learned Kernels

Nati Srebro
TTI-Chicago

Shai Ben-David
University of Waterloo

Mostly based on a paper presented at COLT'06

Kernelized
Large-Margin Linear Classification



$$K(x_1,x_2) = \langle\, \phi(x_1)\,,\, \phi(x_2)\, \rangle$$

- Implicitly defines a Hilbert space in which we seek large-margin separation
- **Represents our prior knowledge, or bias**

$K(x,x) \leq B^2$

$$\text{estimation error} = E[\text{error}] - \text{training error} \leq \sqrt{\frac{\mathcal{O}\left((B/\gamma)^2\right) - \log \delta}{n}}$$

failure probability

sample size

sample complexity $\approx (B/\gamma)^2$

Learning the Kernel

- Success of learning rests on choice of a “good” Kernel, appropriate for the task
  - How can we know which kernel is “good” for the task at hand?
- Jointly learn classifier *and* Kernel, using the training data: Search for a kernel from some family $\mathcal{K}$ of allowed kernels
  - Learn bandwidth, or covariance matrix of Gaussian kernel; other kernel parameters **[Cristianini+98][Chapelle+02][Keerthi02]** etc
  - Linear, or convex, combination of base kernels **[Lacnkriet+02,04][Crammer+03]**; applications, esp. in Bioinformatics **[Sonnenburg+05][Ben-Hur&Noble05]** etc
- More flexibility: lower approximation error, but higher estimation error

What is the sample complexity cost of this flexibility?

Outline

With a fixed kernel:

$$\text{estimation error} \leq \sqrt{\frac{\mathcal{O}\left((B/\gamma)^2\right) - \log\delta}{n}}$$

How does this change when the kernel is learned from some family $\mathcal{K}$?

What is the "cost" of learning the kernel?

- Main result: Learning bound for general kernel families
  - Additive increase to the sample complexity
- Examples: bounds for specific families
- Learn $\sum_i \alpha_i K_i$ or just use $\sum_i K_i$?
- Group Lasso (block-L_1)
- On demand: proof technique (very simple) and why using the Rademacher complexity **can't** work

Previous Bounds: Specific Kernel Families

$$\mathcal{K}_{\text{convex}}(K_1, \ldots, K_k) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{k} \lambda_i K_i \mid \lambda_i \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

estimation error $\leq \sqrt{2 \frac{k \bullet (\frac{B}{\gamma})^2 - \log \delta}{n}}$

[Lanckriet+ JMLR 2004]

$$\mathcal{K}^{\ell}_{\text{Gaussian}} \stackrel{\text{def}}{=} \left\{ (x_1, x_2) \mapsto e^{-(x_1 - x_2)' A (x_1 - x_2)} \mid \text{psd } A \in \mathbb{R}^{\ell \times \ell} \right\}$$

estimation error $\leq \sqrt{2 \frac{C_\ell \bullet (\frac{B}{\gamma})^2 - \log \delta}{n}}$

unspecified function of input dimensionality

[Micchelli+ 2005]

Suggests a multiplicative increase in the required sample size.

Finite Cardinality $\mathcal{K}=\{K_1,K_2,\ldots,K_{|\mathcal{K}|}\}$

For a single kernel K:

$$\Pr\Bigg(\underbrace{\exists \text{ margin-}\gamma \text{ classifier } \textbf{w.r.t. K} \ \ \text{estimation error} > \sqrt{\frac{\mathcal{O}\big((B/\gamma)^2\big) - \log\delta}{n}}}_{\text{"}\textit{bad event}\text{" for a kernel K}}\Bigg) < \delta$$

For a finite kernel family $\mathcal{K}$, set $\delta \leftarrow \delta/|\mathcal{K}|$, and take a union bound over "bad events":

$$\Pr\Bigg(\exists\, \mathbf{K}\in\mathcal{K} \ \ \exists \text{ margin-}\gamma \text{ class. } \textbf{w.r.t. K} \ \ \text{estimation error} > \sqrt{\frac{\mathcal{O}\big((B/\gamma)^2\big) - \log\delta/|\mathcal{K}|}{n}}\Bigg) < |\mathcal{K}|\frac{\delta}{|\mathcal{K}|}$$

$$\Pr\Bigg(\exists\, \mathbf{K}\in\mathcal{K} \ \ \exists \text{ margin-}\gamma \text{ class. } \textbf{w.r.t. K} \ \ \text{estimation error} > \sqrt{\frac{\mathcal{O}\big((B/\gamma)^2 + \log|\mathcal{K}|\big) - \log\delta}{n}}\Bigg) < \delta$$

Main Result

An additive bound for general kernel families, in terms of their *pseodo-dimension*:

For any K chosen from $\mathcal{K}$, and any classifier with margin γ with respect to $\mathcal{K}$:

$$16+8d_\phi \log \frac{128en^3B^2}{\gamma^2 d_\phi}+2048\left(\frac{B}{\gamma}\right)^2 \log \frac{\gamma en}{8B} \log \frac{128nB}{\gamma^2}$$

$$\text{estimation error} \leq \sqrt{\frac{\tilde{O}\Big((B/\gamma)^2 + d_\phi(\mathcal{K})\Big) - \log\delta}{n}}$$

sample complexity $\approx (B/\gamma)^2 + d_\phi(\mathcal{K})$

$d_\phi(\mathcal{K})$ = pseudo-dimension of $\mathcal{K}$
= VC-dimension of subgraphs of K $\in \mathcal{K}$

$\{ (x_1,x_2,t) \mid K(x_1,x_2)<t \}$

Bounds for Specific Kernel Families

$$\mathcal{K}_{\text{convex}}(K_1,\ldots,K_k) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{k} \lambda_i K_i \mid \lambda_i \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

Previous result: $\text{estimation error} \leq \sqrt{2\frac{k \bullet (\frac{B}{\gamma})^2 - \log \delta}{n}}$ **[Lanckriet+ JMLR 2004]**

$$\mathcal{K}_{\text{linear}}(K_1,\ldots,K_k) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{k} \lambda_i K_i \mid K_{\vec{\lambda}} \text{ is psd and } \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

No previous bounds

Applying our result:

$d_\phi(\mathcal{K}_{\text{linear}}),\ d_\phi(\mathcal{K}_{\text{convex}}) \leq k$

$$\text{estimation error} \leq \sqrt{\frac{\tilde{O}\left((B/\gamma)^2 + k\right) - \log \delta}{n}}$$

Bounds for Specific Kernel Families

$$\mathcal{K}^{\ell}_{\text{Gaussian}} \stackrel{\text{def}}{=} \left\{(x_1, x_2) \mapsto e^{-(x_1-x_2)' A (x_1-x_2)} \mid \text{psd } A \in \mathbb{R}^{\ell \times \ell}\right\}$$

Previous result:

estimation error $\leq \sqrt{2 \frac{C_\ell \bullet (\frac{B}{\gamma})^2 - \log \delta}{n}}$

unspecified function of input dimensionality

[Micchelli+ 2005]

Applying our result:

$d_\phi(K_{\text{Gaussian}}) \leq \ell(\ell+1)/2$

input dimensionality

estimation error $\leq \sqrt{\frac{\tilde{O}\left((B/\gamma)^2 + \ell^2\right) - \log \delta}{n}}$

Only diagonal A: ℓ

Only rank(A)≤k: $k\ell \log_2(8ek\ell)$

Additive vs. Multiplicative

$$\mathcal{K}_{\text{convex}}(K_1,\ldots,K_k) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{k} \lambda_i K_i \mid \lambda_i \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

Sample complexity analysis:

If $\exists$ predictor with error *err* at margin γ relative to some K $\in \mathcal{K}$,
How many sample needed to get error *err*+ε ?

Answer according to multiplicative bound: $\mathcal{O}\left(\frac{k(B/\gamma)^2}{\epsilon^2}\right)$

Answer according to our (additive) bound: $\tilde{\mathcal{O}}\left(\frac{(B/\gamma)^2 + k}{\epsilon^2}\right)$

Relaxed approach: Just use $\sum_i K_i$

Feature Space View

Instead of multiple kernels K_i, can think of implied feature spaces directly:

$$K_i(x,x') = \left\langle \phi_i(x), \phi_i(x') \right\rangle$$

$$\phi(x) = \underbrace{\sqrt{\alpha_1}\cdot\phi_1(x)}_{w_1}\ \underbrace{\sqrt{\alpha_2}\cdot\phi_2(x)}_{w_2}\ \ldots\ \underbrace{\sqrt{\alpha_k}\cdot\phi_k(x)}_{w_k}$$

$$w =$$

Weighting each feature space by $\sqrt{\alpha_i} \Rightarrow K = \sum_i \alpha_i K_i$

Relaxed approach: use unweighted feature space $\phi(x)$

- $K=\sum_i K_i$
- $\|w\|^2=\sum_i \|w_i\|^2$ required in unweighted space $\leq \|w\|^2$ in any weighted space
- $B^2_K = kB^2$
- Estimation error bound: $\mathcal{O}\left(\sqrt{\frac{kB^2\|w\|^2}{n}}\right)$

Additive vs. Multiplicative

$$\mathcal{K}_{\text{convex}}(K_1,\ldots,K_k) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{k} \lambda_i K_i \mid \lambda_i \geq 0 \text{ and } \sum_{i=1}^{k} \lambda_i = 1 \right\}$$

Sample complexity analysis:

If $\exists$ predictor with error *err* at margin γ relative to some K $\in \mathcal{K}$,
How many sample needed to get error *err*+ε ?

Answer according to multiplicative bound: $\mathcal{O}\left(\frac{k(B/\gamma)^2}{\epsilon^2}\right)$

Answer according to our (additive) bound: $\tilde{\mathcal{O}}\left(\frac{(B/\gamma)^2+k}{\epsilon^2}\right)$

Relaxed approach: Just use $\sum_i K_i$

- margin γ relative to some K $\in \mathcal{K} \rightarrow$ margin γ relative to $\sum_i K_i$
- $B^2_{\sum K_i} = \sup_x K(x,x) \leq k \cdot B^2_K$

Sample complexity: $\mathcal{O}\left(\frac{k(B/\gamma)^2}{\epsilon^2}\right)$

Learn $\sum_i \alpha_i K_i$ or use $\sum_i K_i$?

Relative to margin γ for some $\sum_i \alpha_i K_i$:

Learn $\sum_i \alpha_i K_i$: error of learned predictor $\leq$ error of best margin γ predictor with some $\sum_i \alpha_i K_i$ $+ \sqrt{\frac{\tilde{\mathcal{O}}\left((B/\gamma)^2 + k\right)}{n}}$

Use $\sum_i K_i$: error of learned predictor $\leq$ error of best margin γ predictor with some $\sum_i \alpha_i K_i$ $+ \sqrt{\frac{\mathcal{O}\left(k(B/\gamma)^2\right)}{n}}$

- Do we have enough samples to afford the factor of *k*?
- Is decrease in estimation error worth the computational cost?
  (maybe not if we have enough data and the estimation error is small anyway)

Relative to margin γ for $\sum_i (1/k) K_i$:

Use $\sum_i K_i$: error of learned predictor $\leq$ error of best margin γ predictor with $\sum_i (1/k) K_i$ $+ \sqrt{\frac{\mathcal{O}\left((B/\gamma)^2\right)}{n}}$

Flexibility with setting weights $\Rightarrow$ Lower approximation error
$\Rightarrow$ but $\sqrt{k/n}$ increase to estimation error

- Is the decrease in approximation error worth the increase in estimation error?
  (and the extra computational cost)

Alternate View: Group Lasso

Instead of multiple kernels K_i, can think of implied feature spaces directly:

$$\phi(x) = \underbrace{\sqrt{\alpha_1}\cdot\phi_1(x)}_{w_1} \underbrace{\sqrt{\alpha_2}\cdot\phi_2(x)}_{w_2} \;\ldots\; \underbrace{\sqrt{\alpha_k}\cdot\phi_k(x)}_{w_k}$$

$w =$

$K_i(x, x') = \left\langle \phi_i(x), \phi_i(x') \right\rangle$

Weighting each feature space by $\sqrt{\alpha_i} \Rightarrow K = \sum_i \alpha_i K_i$

Relaxed approach: use unweighted feature space $\phi(x)$

- $K = \sum_i K_i$, $B^2_K = kB^2$
- $||w||^2 = \sum_i ||w_i||^2$ required in unweighted space $\leq ||w||^2$ in any weighted space
- Estimation error bound: $\mathcal{O}\left(\sqrt{\frac{kB^2 \sum_i \|w_i\|^2}{n}}\right)$

[Bach et al 04] Learning with $\mathcal{K}_{convex}$ equivalent to using unweighted feature space $\phi(x)$ and Block-L_1 regularizer $\sum_i ||w_i||$

$||w||^2 = \sum_i ||w_i||^2 \leq (\sum_i ||w_i||)^2$

est error for group lasso $\leq \tilde{\mathcal{O}}\left(\sqrt{\frac{B^2 \left(\sum_i \|w_i\|\right)^2 + k}{n}}\right)$

Proof Sketch

bound pseudodimension $d_\phi(\mathcal{K})$

standard result on covering numbers in terms of d_ϕ

covering of $\mathcal{K}$ of size $(\cdots)^{d_\phi(\mathcal{K})}$

standard results on covering numbers of the unit sphere

covering of $\mathcal{F}_K$ of size $(\cdots)^{(B/\varepsilon)^2}$

Construct covering for $\mathcal{F}_\mathcal{K}$ as “cross-product”:
for each kernel K in the covering of $\mathcal{K}$, take the covering of $\mathcal{F}_K$.

Lemma: if K, K’ are similar as real-valued functions, every K-classifier can be approximated by K’-classifier

covering of $\mathcal{F}_\mathcal{K}$ of size $(\cdots)^{d_\phi(\mathcal{K})} \cdot (\cdots)^{(B/\varepsilon)^2}$

generalization error bounds in terms of log(covering number)

Rademacher vs. Covering Numbers

- Other bound rely on calculating the Rademacher complexity $\mathcal{R}[\mathcal{F}_{\mathcal{K}}]$ of the class of classifiers (unit norm) classifiers with respect to any K $\in \mathcal{K}$
  - $\mathcal{R}[\mathcal{F}_{\mathcal{K}}]$ scales with the scale of functions in $\mathcal{F}_{\mathcal{K}}$, i.e. with B.
  - Generalization error bounds depend on $\mathcal{R}[\mathcal{F}_{\mathcal{K}}]/\gamma$

  $\Rightarrow$ **Bounds based on the Rademacher Complexity necessarily have a multiplicative dependence on B/γ**

- Covering numbers allow us to combine scale-sensitive and finite-dimensionality (scale insensitive) arguments

  (at the cost of messier log-factors)

Learning Bounds for SVMs with Learned Kernels

Nati Srebro Shai Ben-David

- Bound on estimation error for large margin classifier with respect to kernel which is chosen, from family $\mathcal{K}$, based on training data:

pseudodimension of $\mathcal{K}$, as family of real-valued functions

$$\sqrt{\frac{\tilde{O}\left(d_\phi(\mathcal{K}) + (B/\gamma)^2\right) - \log \delta}{n}}$$

- Valid for generic kernalized L_2-regularized learning
- Easy to obtain bounds for further kernel families
- For $\mathcal{K}_{convex}$: using $\sum_i K_i$ may require k times more data