Similarity-Based Theoretical Foundation for Sparse Parzen Window Prediction

Nina Balcan Avrim Blum
Carnegie Mellon

Nati Srebro
Toyota Technological Institute—Chicago

Sparse Parzen Window Prediction

- We are concerned with predictors of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x'_i)$$

  $x'_1,\ldots,x'_n$ are ***landmarks*** (often used also as training data) and $K(x,x')$ encodes similarity.

- SVMs: learn α by minimizing objective related to dual large margin problem in implicit Hilbert space.
- Parzen/Soft Nearest-Neighbor: $\alpha_i = y'_i$
- Learn α_i by directly minimizing empirical loss
- Also want **sparsity**, i.e. many α_i=0, and so only few landmarks actually used for prediction

The Learning Rule

- Use $|\alpha|_1 = \sum_i |\alpha_i|$ as surrogate for sparsity
- Hinge loss: $[1-y\cdot f(x)]_+ = \max(0,1-y\cdot f(x))$
- Yields the popular learning rule:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{m} [1 - y_i f(x_i)]_+ \\ \text{s.t.} \quad & \sum_{j=1}^{n} |\alpha_j| \leq M \end{aligned} \qquad (*)$$

where $(x_1,y_1),\ldots,(x_m,y_m)$ are **labeled** training examples, which might, or might not, be the same as the landmarks (recall landmarks need not be labeled).

The Learning Rule: References

- Bennett and Campbell (SIGKDD Explot. Newsl. 2000), **Support vector machines: hype or hallelujah?**
- Roth (ICANN'01), **Sparse kernel regressors**.
- Guigue, Rakotomamonjy & Canu (ECML'05), **Kernel basis pursuit.**

With different loss functions:

- Singer (NIPS'99), **Leveraged vector machines.**

Combined with $|\alpha|_2$ regularization:

- Osuna and Girosi (1999). **Reducing the run-time complexity in support vector machines.** Advances in kernel methods: Support Vector learning.
- Gunn and Kandola (2002). **Structural modelling with sparse kernels.** Machine Learning, *48*, 137—163

Learning Guarantees?

- Despite popularity of learning rule (*), **no established guarantees in terms of K**!
- For SVMs, guarantees based on large margin in implied feature space.
- Even if SVM condition holds (large margin in implied space), can (*) also be used? No previously known guarantee…
  - In fact, combining $|\alpha|_1$ and $|\alpha|_2$ suggested in order to benefit from SVM guarantees.

- Is there a simple and interpretable condition on K that guarantees learnability using rule (*)?
  - Since (*) doesn't require $K \succcurlyeq 0$, would hope for guarantees that do not rely on $K \succcurlyeq 0$.
  - Do landmarks have to be training examples?

Our Results

- **Natural condition on K that justifies Learning Rule (*)**
  - View K as similarity function
  - No requirement that $K \succcurlyeq 0$
  - Labeled sample complexity (training points) and unlabeled sample complexity (landmarks) yielding generalization error bound.

- If $K \succcurlyeq 0$ and is a good kernel for SVMs

  $\Rightarrow$ also satisfies our condition

  $\Rightarrow$ Learning Rule (*) can be used

Non-PSD Similarity

SVM requires $K \succcurlyeq 0$

- Often not the case for natural similarity, e.g.:
  - "Earth Movers Distance" (especially in vision)
  - BLAST scores for proteins or DNA
  - $K(x_1,x_2)=P_{x'}[d(x_1,x_2) \leq d(x_1,x')]$

Percentile rank of how close x_2 is to x_1, relative to all other points

- Can coerce K to be PSD and use SVM:
  - Gerpel et al (NIPS'98), **Classification of pairwise proximity data**
  - Wu et al (ICML'05), **An analysis of transformations on non-positive semidefinite similarity matrix for kernel machines**
  - Luss and d'Aspremont (NIPS'07), **SVM classification with indifinite kernels**
  - Checn and Ye (ICML'08), **Training SVMs with indefinite kernels**

- But perhaps more natural to use (*)

Our guarantee justifies using (*), even when $K \not\succcurlyeq 0$.

Condition Justifying Learning Rule (*)

Definition: K is a **$(\varepsilon,\gamma,\tau)$-good similarity** function if there exists a probabilistic set R of "reasonable points" such that:

Can think of R(x) as random 0/1 indicator

- There is at least τ mass of reasonable points:

$$\Pr_{x',R(x')}[\ R(x')\ ] \geq \tau$$

- A parzen predictor based on the reasonable points has average hinge loss at most ε relative to margin γ:

$$E_{x,y}[\ [1-y\cdot g(x)/\gamma]_+\ ] \leq \varepsilon$$

where $g(x) = E_{x',y',R(x')}[\ y'\ K(x,x')\ |\ R(x')\ ]$

Theorem: If K is a $(\varepsilon,\gamma,\tau)$-good similarity function, then, for any $\delta,\varepsilon_1>0$, with probability $\geq$ 1-δ over a sample $x'_1,\ldots,x'_n$ of

$$n = \frac{2}{\tau}\left(\log(2/\delta) + 16\frac{\log(2/\delta)}{\epsilon_1^2\gamma^2}\right)$$

random (potentially unlabeled) landmarks, there exists a predictor

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x'_i)$$

With low ℓ_1-norm:

$$|\alpha|_1 = \sum |\alpha_i| \leq 1/\gamma$$

and low expected error:

$$E_{x,y}[\ [1-y\cdot f(x)]_+\ ] \leq \varepsilon + \varepsilon_1$$

Corollary: If K is a $(\varepsilon,\gamma,\tau)$-good similarity function, then, for any $\delta,\varepsilon_1>0$, with probability $\geq 1-\delta$ over a sample $x'_1,\ldots,x'_n$ of

$$n = O\left(\frac{\log(1/\delta)}{\tau\gamma^2\epsilon_1^2}\right)$$

random (potentially unlabeled) landmarks, and a (labeled) sample $(x_1,y_1),\ldots,(x_m,y_m)$ of size

$$m = \tilde{\mathcal{O}}\left(\frac{\log n \log(1/\delta)}{\gamma^2\epsilon_1^2}\right)$$

the predictor obtain by learning rule (*) with $M=1/\gamma$ has expected hinge loss:

$$E_{x,y}[\ [1-y\cdot f(x)]_+\ ] \leq \varepsilon + \varepsilon_1$$

Good for SVM $\Rightarrow$ Good for (*)

Definition: K$\succcurlyeq$0 is a **(ε,γ)-good kernel** if there exists a vector β, $|\beta| \leq 1/\gamma$, in the implied Hilbert space s.t. $E[\,[1-y\cdot\langle\beta,x\rangle]_+\,] \leq \varepsilon$

Theorem: If K$\succcurlyeq$0 is a (ε,γ)-good kernel (for a problem with deterministic labels), then for any ε_1>0, K is also a $\left(\epsilon_0+\epsilon_1, \frac{\gamma^2}{(1+\epsilon_0/\epsilon_1)}, \epsilon_0+2\epsilon_1\right)$-good similarity function.

Actually, might be $(\varepsilon_0+\varepsilon_1, c\gamma^2/(1+\varepsilon_0/\varepsilon_1), (\varepsilon_0+2\varepsilon_1)/c)$-good, for some c>1, which is only better in terms of learning guarantees.

Corollary: If $K \succcurlyeq 0$ is a (ε,γ)-good kernel, then for any $\varepsilon_1,\delta>0$, with probability $\geq 1-\delta$, over a sample $x'_1,\ldots,x'_n$ of

$$n = O\left(\frac{(1+\epsilon/\epsilon_1)^2 \log(1/\delta)}{(\epsilon+\epsilon_1)\gamma^4\epsilon_1^2}\right)$$

random (potentially unlabeled) landmarks, and a (labeled) sample $(x_1,y_1),\ldots,(x_m,y_m)$ of size

$$m = \tilde{O}\left(\frac{(1+\epsilon/\epsilon_1)^2 \log n \log(1/\delta)}{\gamma^4\epsilon_1^2}\right)$$

the predictor obtain by learning rule (*) with $M=1/\gamma$ has expected hinge loss $\leq \varepsilon + \varepsilon_1$

Full details and proofs:
Improved Guarantees for Learning via Similarity Functions, Balcan, Blum and Srebro, COLT 2008