Learning with General Similarity Functions

Maria-Florina Balcan

2-Minute Version

Generic classification problem:  

Problem: pixel representation not so good.

Powerful technique: use a **kernel**, a special kind of similarity function K( , ).

But, standard theory in terms of implicit mappings.

Our Work:

Develop a theory that views K as a measure of similarity.
General sufficient conditions for K to be useful for learning.

[Balcan-Blum, ICML 2006] [Balcan-Blum-Srebro, MLJ 2008] [Balcan-Blum-Srebro, COLT 2008]

Kernel Methods

Prominent method for supervised classification today.

The learning alg. interacts with the data via a similarity fns

What is a Kernel?

A kernel K is a legal def of dot-product: i.e. there exists an implicit mapping Φ such that K( , )=Φ ()·Φ ().

E.g., $K(x,y) = (x \cdot y + 1)^d$

ϕ: (n-dimensional space) $\rightarrow$ n^d-dimensional space

Why Kernels matter?

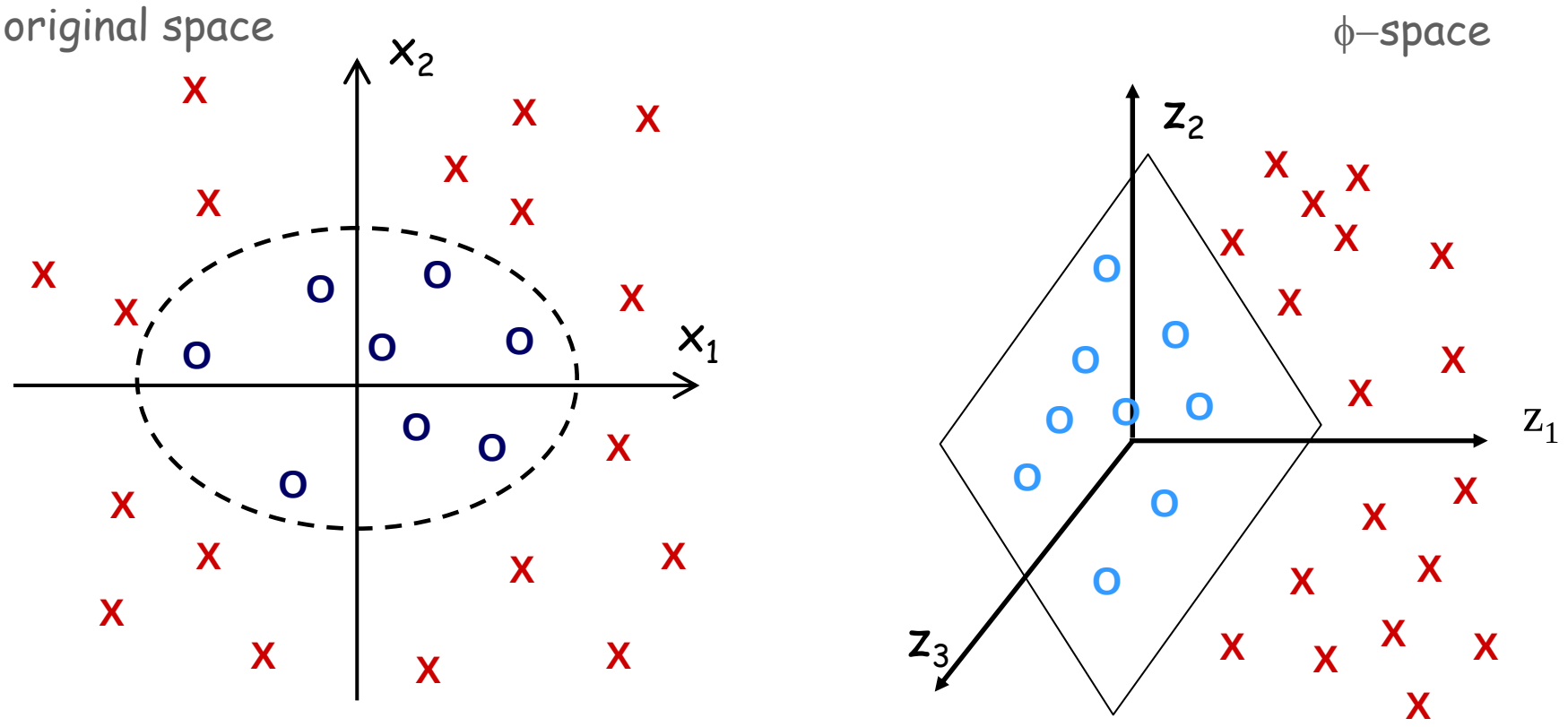
Many algorithms interact with data only via dot-products.

So, if replace $x \cdot y$ with K(x,y), they act implicitly as if data was in the higher-dimensional ϕ-space.

Example

E.g., for n=2, d=2, the kernel $K(x,y) = (x \cdot y)^d$ corresponds to

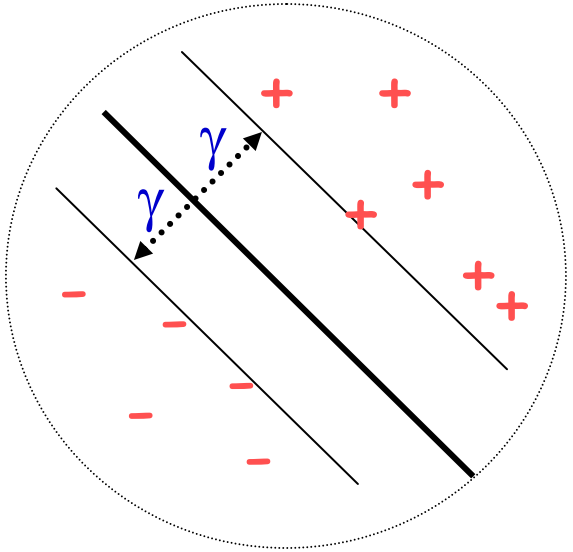
$$(x_1, x_2) \mapsto \phi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$$



Generalize Well if Good Margin

- If data is linearly separable by margin in ϕ-space, then good sample complexity.

If margin γ in ϕ-space, then need sample size of only $\tilde{O}(1/\gamma^2)$ to get confidence in generalization.



$|\phi(x)| \leq 1$

Kernel Methods

Prominent method for supervised classification today

Very useful in practice for dealing with many different types of data.

Significant percentage of ICML, NIPS, COLT.

Limitations of the Current Theory

In practice: kernels are constructed by viewing them as measures of similarity.

Existing Theory: in terms of margins in implicit spaces.

Difficult to think about, not great for intuition.

Kernel requirement rules out many natural similarity functions.

Better theoretical explanation?

Better Theoretical Framework

Yes! We provide a more general and intuitive theory that formalizes the intuition that a good kernel is a good measure of similarity.

[Balcan-Blum, ICML 2006] [Balcan-Blum-Srebro, MLJ 2008]

[Balcan-Blum-Srebro, COLT 2008]

Better theoretical explanation?

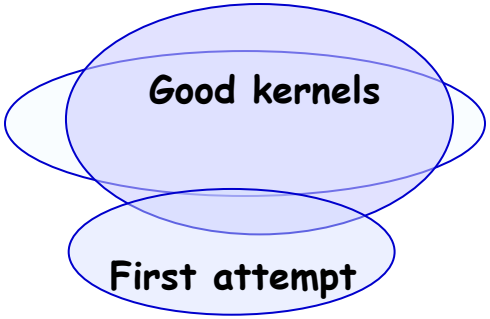
More General Similarity Functions

We provide a notion of a good similarity function:

1) Simpler, in terms of natural direct quantities.

- no implicit high-dimensional spaces
- no requirement that $K(x,y)=\phi(x) \cdot \phi(y)$

K can be used to learn well.



2) Is broad: includes usual notion of good kernel.

has a large margin sep. in ϕ-space

3) Allows one to learn classes that have no good kernels.

A First Attempt

P distribution over labeled examples $(x, c(x))$

Goal: output classification rule good for P

K is good if most x are on average more similar to points y of their own type than to points y of the other type.

K is (ε,γ)-good for P if a $1-\varepsilon$ prob. mass of x satisfy:

$$E_{y\sim P}[K(x,y)|c(y)=c(x)] \geq E_{y\sim P}[K(x,y)|c(y)\neq c(x)]+\gamma$$

gap

Average similarity to points of the same label

Average similarity to points of opposite label

A First Attempt

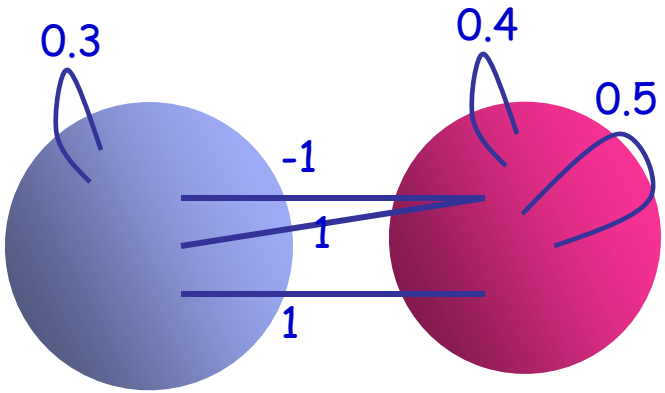
K is (ε,γ)-good for P if a $1-\varepsilon$ prob. mass of x satisfy:

$$E_{y\sim P}[K(x,y)|\phi(y)=\phi(x)] \geq E_{y\sim P}[K(x,y)|\phi(y)\neq\phi(x)]+\gamma$$

Example:

E.g., $K(x,y) \geq 0.2$, $\phi(x) = \phi(y)$

$K(x,y)$ random in $\{-1,1\}$, $\phi(x) \neq \phi(y)$



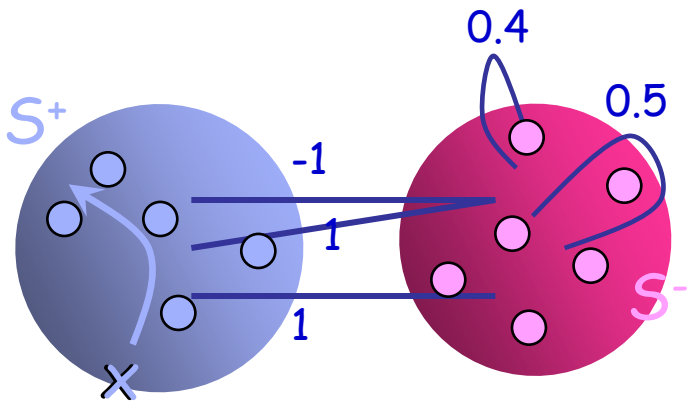
A First Attempt

K is (ε,γ)-good for P if a $1-\varepsilon$ prob. mass of x satisfy:

$$E_{y\sim P}[K(x,y)|\phi(y)=\phi(x)] \geq E_{y\sim P}[K(x,y)|\phi(y)\neq\phi(x)]+\gamma$$

Algorithm

- Draw sets S^+, S^- of positive and negative examples.
- Classify x based on average similarity to S^+ versus to S^-.



A First Attempt

K is (ε,γ)-good for P if a $1-\varepsilon$ prob. mass of x satisfy:

$$E_{y\sim P}[K(x,y)|\phi(y)=\phi(x)] \geq E_{y\sim P}[K(x,y)|\phi(y)\neq\phi(x)]+\gamma$$

Algorithm

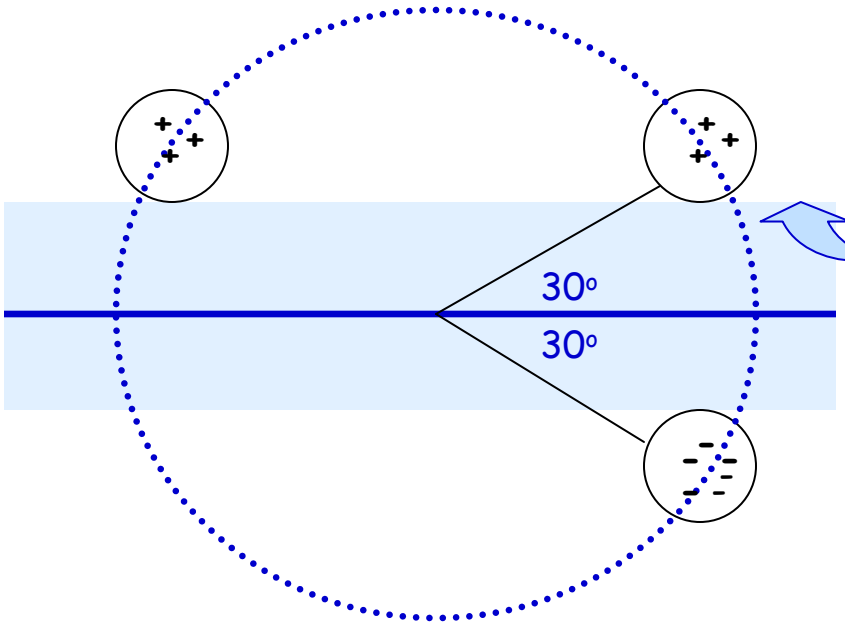
- Draw sets S^+, S^- of positive and negative examples.
- Classify x based on average similarity to S^+ versus to S^-.

Theorem If $|S^+|$ and $|S^-|$ are $\Omega((1/\gamma^2)\ln(1/\delta\varepsilon'))$, then with probability $\geq 1-\delta$, error $\leq \varepsilon+\varepsilon'$.

- For a fixed good x prob. of error w.r.t. x (over draw of S^+, S^-) is $\delta\,\epsilon'$. [Hoeffding]
- At most δ chance that the error rate over GOOD is $\geq \varepsilon'$.
- Overall error rate $\leq \varepsilon+\varepsilon'$.

A First Attempt: Not Broad Enough

$$E_{y\sim P}[K(x,y)|c(y)=c(x)] \geq E_{y\sim P}[K(x,y)|c(y)\neq c(x)]+\gamma$$



more similar to - than to typical +

$\frac{1}{2}$ versus $\frac{1}{4}$

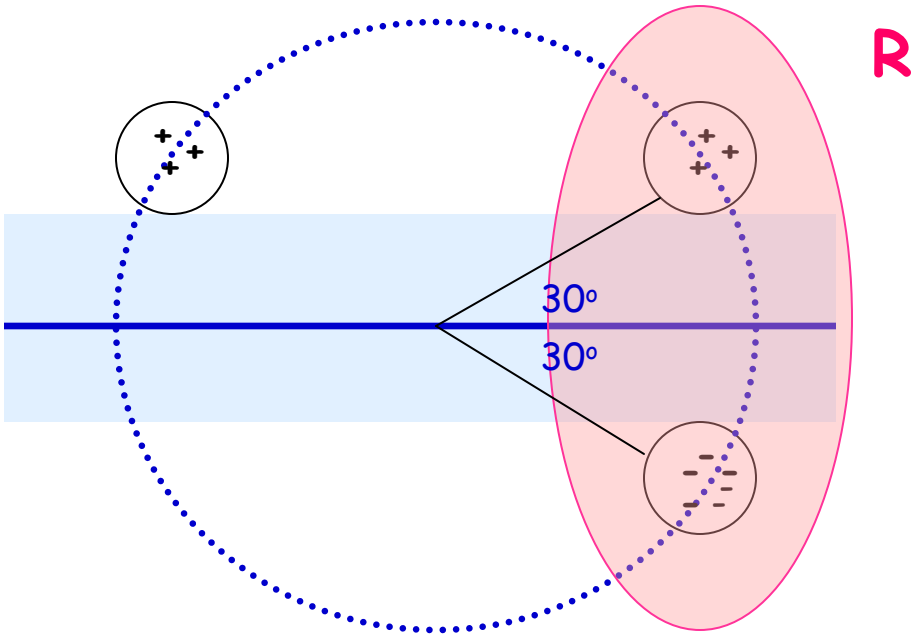
$\frac{1}{2}$ versus $\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (-\frac{1}{2})$

Similarity function $K(x,y)=x \cdot y$

- has a large margin separator; does not satisfy our definition.

A First Attempt: Not Broad Enough

$$E_{y\sim P}[K(x,y)|\phi(y)=\phi(x)] \geq E_{y\sim P}[K(x,y)|\phi(y)\neq\phi(x)]+\gamma$$



Broaden: $\exists$ non-negligible R s.t. most x are on average more similar to $y \in R$ of same label than to $y \in R$ of other label.

[even if do not know R in advance]

Broader Definition

K is $(\varepsilon, \gamma, \tau)$ if $\exists$ a set R of "reasonable" y (allow probabilistic) s.t. $1-\varepsilon$ fraction of x satisfy:

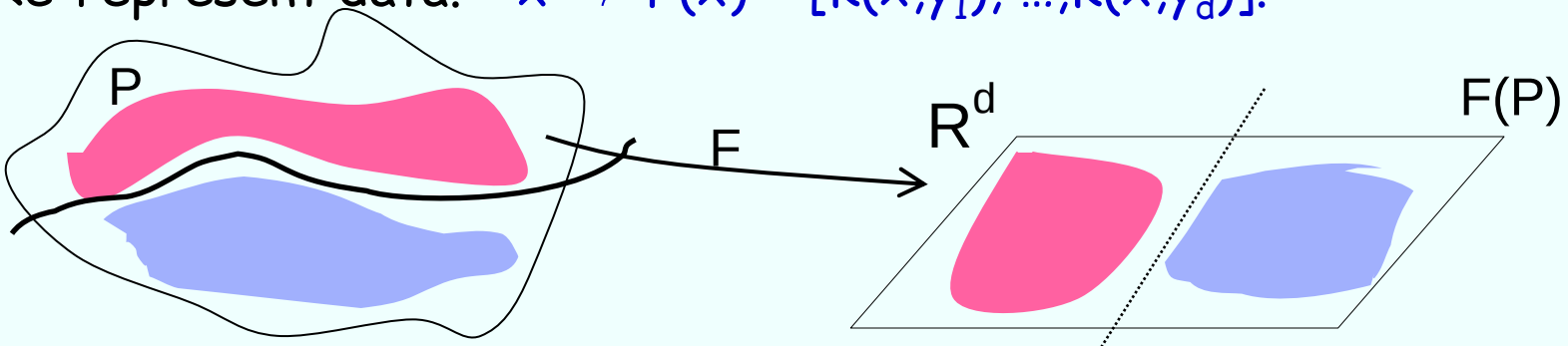
$$E_{y\sim P}[K(x,y)|\phi(y)=\phi(x), R(y)] \geq E_{y\sim P}[K(x,y)|\phi(y)\neq\phi(x), R(y)]+\gamma$$

At least τ prob. mass of reasonable positives & negatives.

Property

- Draw $S=\{y_1, \ldots, y_d\}$ set of landmarks.

  Re-represent data. $x \rightarrow F(x) = [K(x,y_1), \ldots, K(x,y_d)]$.



- If enough landmarks ($d=\Omega(1/\gamma^2\,\tau)$), then with high prob. there exists a good L_1 large margin linear separator.

w=[0,0,1/n+,1/n+,0,0,0,-1/n-,0,0]

Broader Definition

K is $(\varepsilon, \gamma, \tau)$ if $\exists$ a set R of "reasonable" y (allow probabilistic) s.t. $1-\varepsilon$ fraction of x satisfy:

$$E_{y\sim P}[K(x,y)|\phi(y)=\phi(x), R(y)] \geq E_{y\sim P}[K(x,y)|\phi(y)\neq\phi(x), R(y)]+\gamma$$

At least τ prob. mass of reasonable positives & negatives.

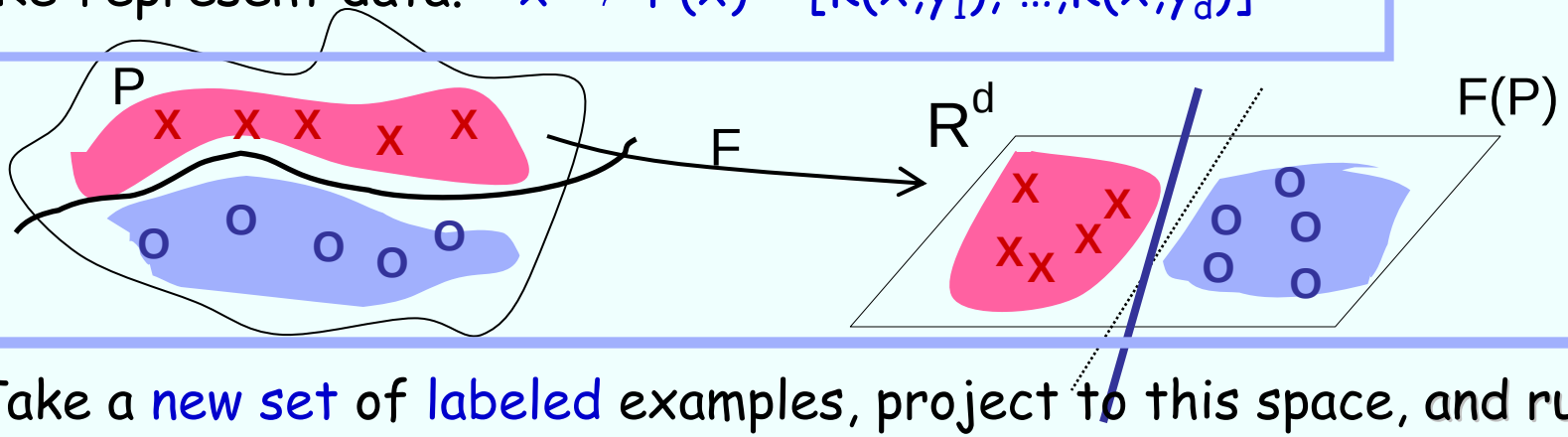
Algorithm

$d_u=\tilde{O}(1/(\gamma^2\tau))$

$d_l=O(1/(\gamma^2\epsilon_{acc} \ln(d_u)))$

- Draw $S=\{y_1, \ldots, y_d\}$ set of landmarks.

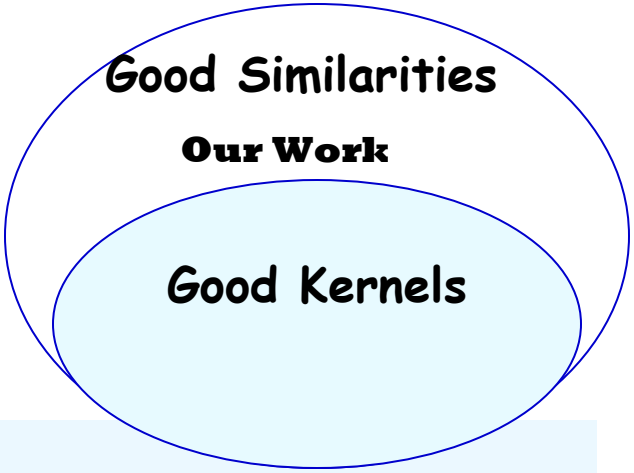
Re-represent data. $x \rightarrow F(x) = [K(x,y_1), \ldots, K(x,y_d)]$



Take a new set of labeled examples, project to this space, and run a good L_1 linear separator alg.

Kernels versus Similarity Functions

Main Technical Contributions



Theorem

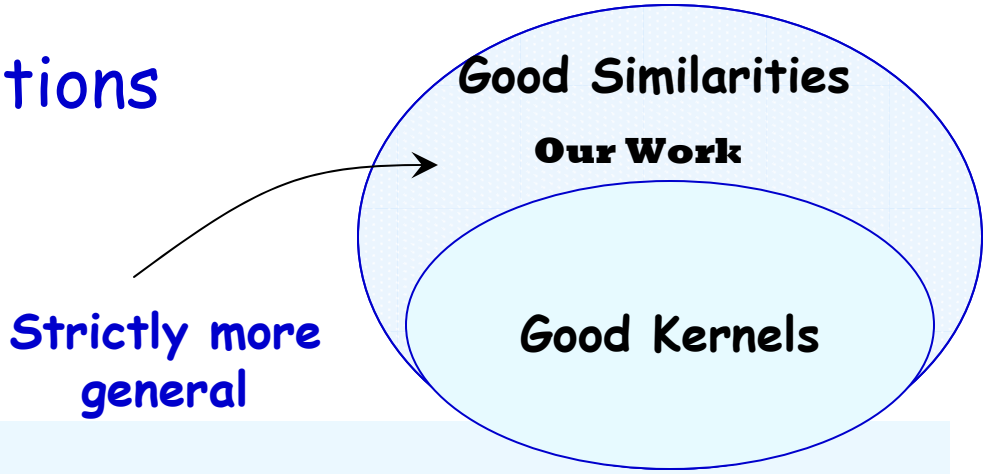
K is a good kernel $\Rightarrow$ K is also a good similarity function.

(but γ gets squared).

If K has margin γ in implicit space, then for any τ,
K is (τ,γ^2,τ)-good in our sense.

Kernels versus Similarity Functions

Main Technical Contributions



Theorem

K is a good kernel $\Rightarrow$ K is also a good similarity function.

(but γ gets squared).

Can also show a Strict Separation.

Theorem

For any class C of n pairwise uncorrelated functions, $\exists$ a similarity function good for all f in C, but no such good kernel function exists.

Kernels versus Similarity Functions

Can also show a Strict Separation.

Theorem

For any class C of n pairwise uncorrelated functions, $\exists$ a similarity function good for all f in C, but no such good kernel function exists.

- In principle, should be able to learn from $O(\varepsilon^{-1}\log(|C|/\delta))$ labeled examples.
- Claim 1: can define generic $(0,1,1/|C|)$-good similarity function achieving this bound. (Assume D not too concentrated)
- Claim 2: There is no (ε,γ) good kernel in hinge loss, even if $\varepsilon=1/2$ and $\gamma=1/|C|^{-1/2}$. So, margin based SC is $d=\Omega(1/|C|)$.

Learning with Multiple Similarity Functions

- Let $K_1, \ldots, K_r$ be similarity functions s. t. some (unknown) convex combination of them is (ε,γ)-good.

Algorithm

- Draw $S=\{y_1, \ldots, y_d\}$ set of landmarks. Concatenate features.

$$F(x) = [K_1(x,y_1), \ldots, K_r(x,y_1), \ldots, K_1(x,y_d), \ldots, K_r(x,y_d)].$$

Guarantee: Whp the induced distribution $F(P)$ in R^{2dr} has a separator of error $\leq \varepsilon + \delta$ at L_1 margin at least $\frac{\gamma}{4}$.

Sample complexity only increases by log(r) factor!

Conclusions

- Theory of learning with similarity fns that provides a formal way of understanding good kernels as good similarity fns.
- Our algorithms work for similarity fns that aren't necessarily PSD (or even symmetric).

Algorithmic Implications

- Can use non-PSD similarities, no need to "transform" them into PSD functions and plug into SVM.

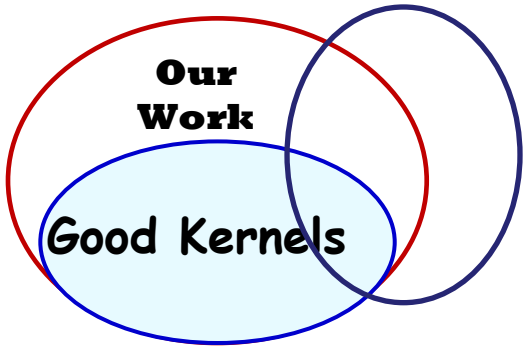
E.g., Liao and Noble, Journal of Computational Biology

Conclusions

- Theory of learning with similarity fns that provides a formal way of understanding good kernels as good similarity fns.
- Our algorithms work for similarity fns that aren't necessarily PSD (or even symmetric).

Open Questions

- Analyze other notions of good similarity fns.



Similarity Functions for Classification

Algorithmic Implications

- Can use non-PSD similarities, no need to "transform" them into PSD functions and plug into SVM.

E.g., Liao and Noble, Journal of Computational Biology

- Give justification to the following rule:

$$\text{minimize} \quad \sum_{i=1}^{d_l} \left[1 - \sum_{j=1}^{d_u} \alpha_j \ell(x_i) K(x_i, \tilde{x}_j) \right] +$$

$$\text{s.t.} \quad \sum_{j=1}^{d_u} |\alpha_j| \leq 1/\gamma$$

- Also show that anything learnable with SVM is learnable this way!