Towards a more intuitive theory of learning with similarity functions

(with extensions to clustering)

Avrim Blum
Carnegie Mellon University

[joint work with Nina Balcan]

Kernel functions have become a great tool in ML

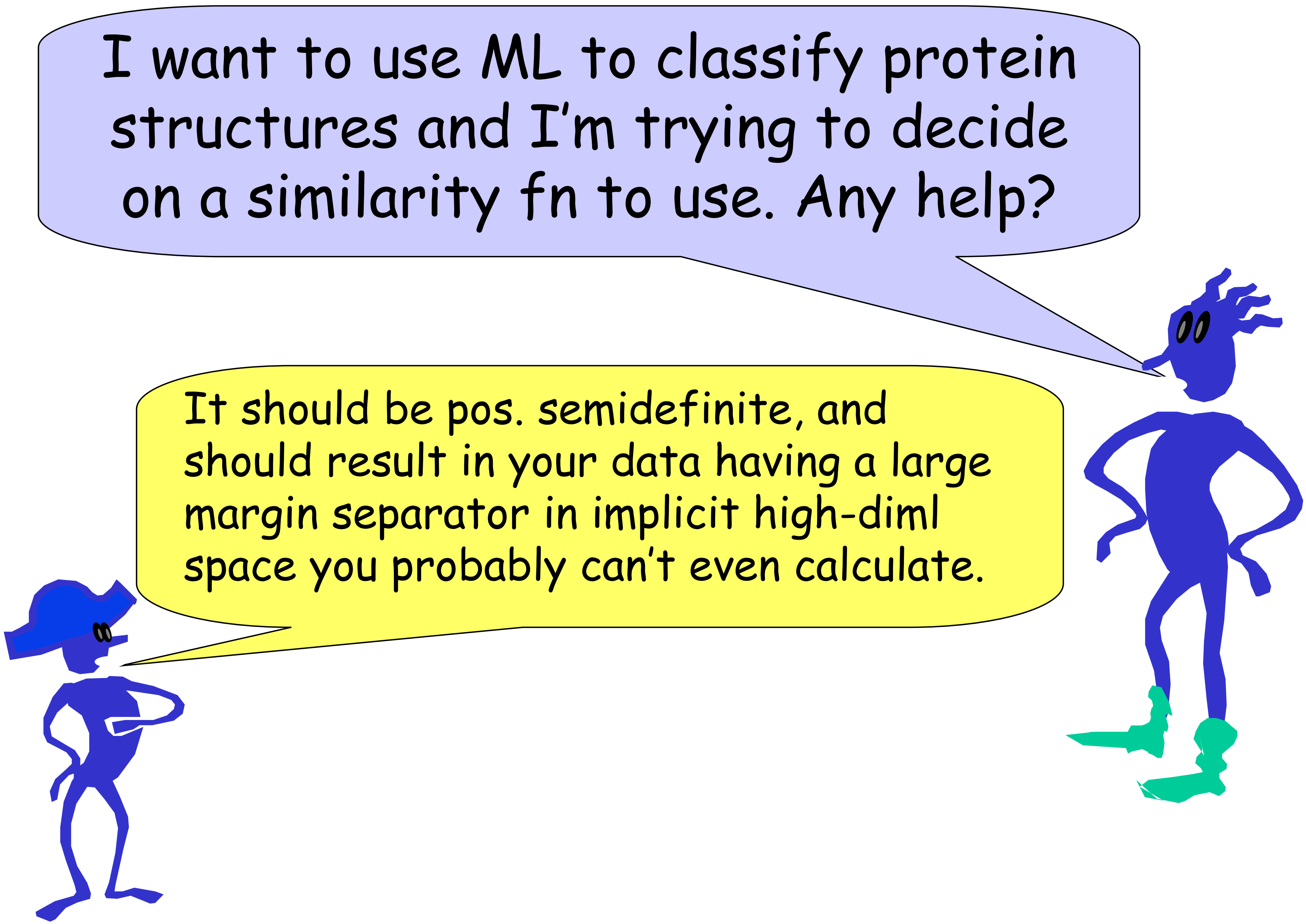
- Useful in practice for dealing with many different kinds of data.
- Elegant theory in terms of margins about what makes a given kernel good for a given learning problem.

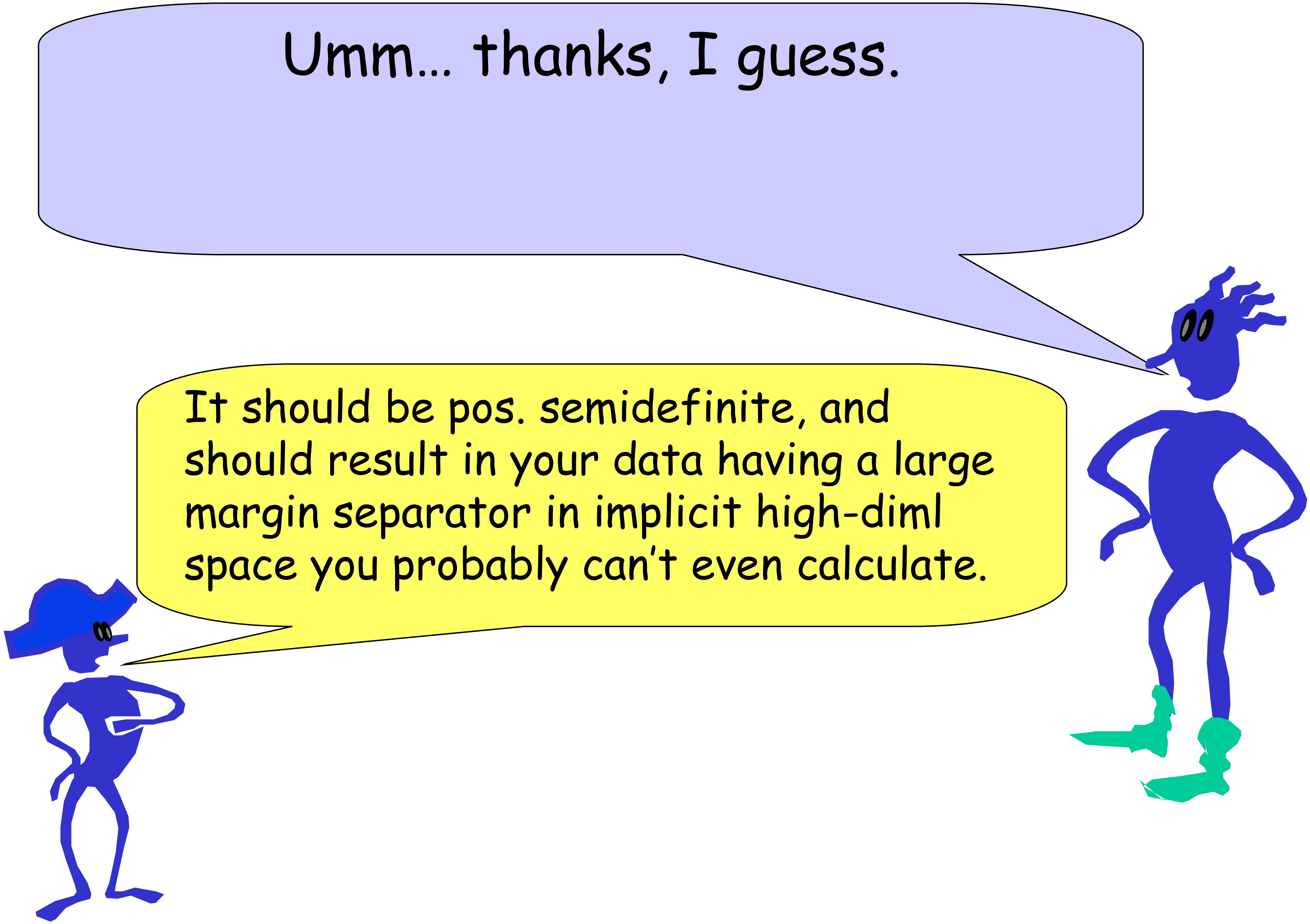
Kernel functions have become a great tool in ML

...but there's something a little funny:

- On the one hand, operationally a kernel is just a similarity function: $K(x,y) \in [-1,1]$, with some extra reqts.

- But Theory talks about margins in implicit high-dimensional ϕ-space. $K(x,y) = \phi(x)\cdot\phi(y)$.


I want to use ML to classify protein structures and I'm trying to decide on a similarity fn to use. Any help?
It should be pos. semidefinite, and should result in your data having a large margin separator in implicit high-diml space you probably can't even calculate.


Umm... thanks, I guess.
It should be pos. semidefinite, and should result in your data having a large margin separator in implicit high-diml space you probably can't even calculate.

Kernel functions have become a great tool in ML

…but there's something a little funny:

- On the one hand, operationally a kernel is just a similarity function: $K(x,y) \in [-1,1]$, with some extra reqts.

- But Theory talks about margins in implicit high-dimensional ϕ-space. $K(x,y) = \phi(x) \cdot \phi(y)$.
  - Not great for intuition (do I expect this kernel or that one to work better for my kind of data)
  - Has a something-for-nothing feel to it. "All the power of the implicit space without having to pay for it". More prosaic explanation?

Goal: definition of "good similarity function" for a learning problem that…

1. Talks in terms of more natural direct properties (no implicit high-diml spaces, no requirement of positive-semidefiniteness, etc)
2. If K satisfies these properties for our given problem, then has implications to learning (can't just say any function is a good one)
3. Is broad: includes usual notion of "good kernel" (one that induces a large margin separator in ϕ-space).

"Learning problem": distrib P over labeled examples x. Assume $\ell(x) \in \{-1,1\}$.

Defn satisfying (1) and (2):

- Say have a learning problem P (distrib over labeled examples).
- $K:(x,y)\rightarrow[-1,1]$ is an (ε,γ)-good similarity function for P if at least a $1-\varepsilon$ prob mass of examples x satisfy:

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

How can we use it?

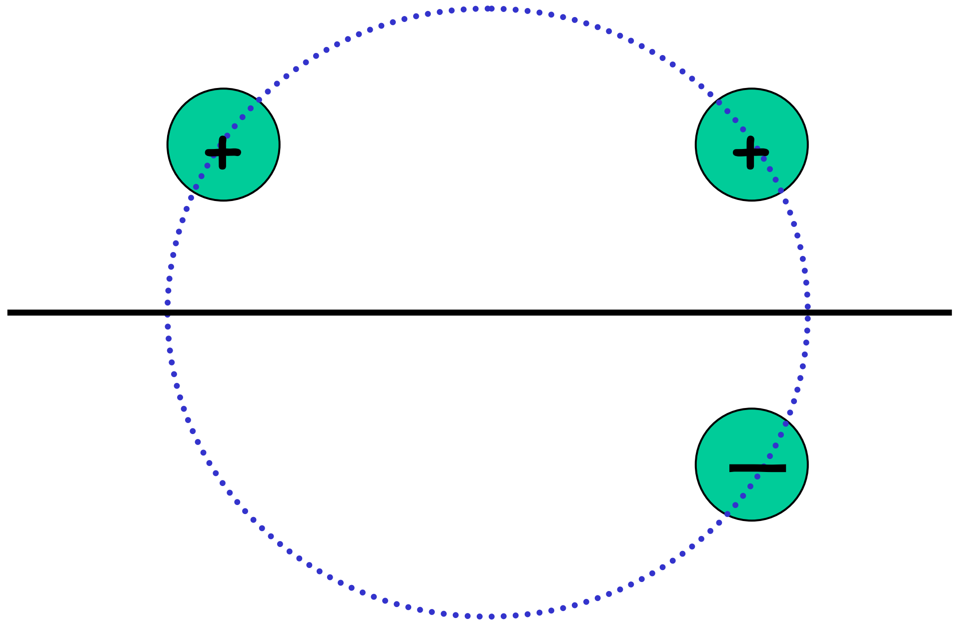
How to use it

At least a 1-ε prob mass of x satisfy:

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

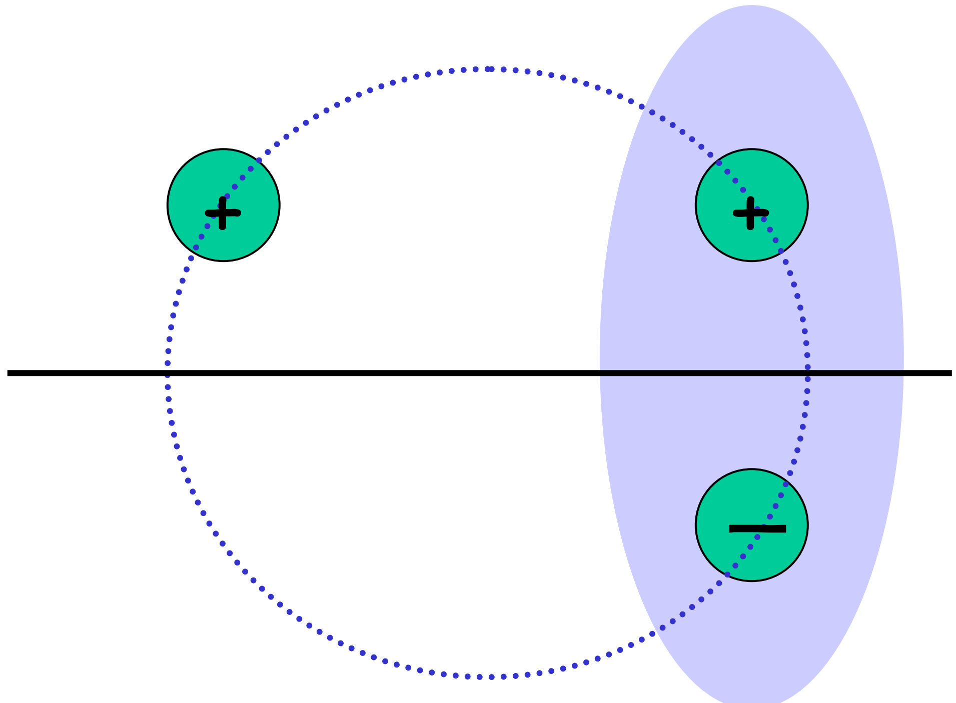
- Draw S^+ of $O(\gamma^{-2}\ln(1/\delta^2))$ positive examples.
- Draw S^- of $O(\gamma^{-2}\ln(1/\delta^2))$ negative examples.
- Classify x based on which gives better score.
- Hoeffding: for any given "good x", prob of error over draw of S^+,S^- at most δ^2.
- So, at most δ chance our draw is bad on more than δ fraction of "good x". So overall error rate $\leq \varepsilon + \delta$.

But not broad enough



- K(x,y)=x·y has good separator but doesn't satisfy defn. (half of positives are more similar to negs that to typical pos)

But not broad enough



- Idea: would work if we didn't pick y's from top-left.
- Broaden to say: OK if $\exists$ large region 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.

Broader defn…

- Say $K:(x,y)\rightarrow[-1,1]$ is an (ε,γ)-good similarity function for P if exists a weighting function $w(y)\in[0,1]$ s.t. at least $1-\varepsilon$ mass of x satisfy:

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

- How to use:
  - Draw $S^+ = \{y_1,\ldots,y_n\}$, $S^- = \{z_1,\ldots,z_n\}$. $n=\tilde{O}(1/\gamma^2)$
  - Use to "triangulate" data:

    $F(x) = [K(x,y_1), \ldots,K(x,y_n), K(x,z_1),\ldots,K(x,z_n)]$.
  - Whp, exists good separator in this space: $w = [w(y_1),\ldots,w(y_n),-w(z_1),\ldots,-w(z_n)]$

Broader defn…

- Say $K:(x,y)\to[-1,1]$ is an (ε,γ)-good similarity function for P if exists a weighting function $w(y)\in[0,1]$ s.t. at least $1-\varepsilon$ mass of x satisfy:

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

  - Whp, exists good separator in this space: $w = [w(y_1),\ldots,w(y_n),-w(z_1),\ldots,-w(z_n)]$
  - So, take new set of examples, project to this space, and run your favorite learning algorithm.

And furthermore

- An (ε,γ)-good kernel [margin $\geq \gamma$ on at least $1-\varepsilon$ fraction of P] is an (ε',γ')-good sim fn under this definition.
- But our current proofs suffer a big penalty: $\varepsilon' = \varepsilon + \varepsilon_{extra}$, $\gamma' = \gamma^4\varepsilon_{extra}$.

And furthermore

- An (ε,γ)-good kernel [margin $\geq \gamma$ on at least $1-\varepsilon$ fraction of P] is an (ε',γ')-good sim fn under this definition.
- But our current proofs suffer a big penalty: $\varepsilon' = \varepsilon + \varepsilon_{extra}$, $\gamma' = \gamma^4\varepsilon_{extra}$.

Proof sketch:

- Set w(y)=0 for the ε fraction of "bad" y's.
- Imagine repeatedly running margin-Perceptron on multiple samples S from remainder.
- Set $w(y) \propto \ell(y) \cdot E[\text{weight}(y) \mid y \in S]$

And furthermore

- An (ε,γ)-good kernel [margin $\geq \gamma$ on at least $1-\varepsilon$ fraction of P] is an (ε',γ')-good sim fn under this definition.
- But our current proofs suffer a big penalty: $\varepsilon' = \varepsilon + \varepsilon_{extra}$, $\gamma' = \gamma^4\varepsilon_{extra}$.

Should be possible to improve bounds.

Maybe one can find better (more intuitive) defs that still capture large margin kernels.

Examples of settings satisfying defs but not legal kernels

- Suppose positives have $K(x,y) \geq 0.8$, negatives have $K(x,y) \geq 0.8$, but for a pos and a neg, $K(x,y)$ are uniform random in $[-1,1]$.
- For a kernel, if a & b are very similar, and a & c are very dissimilar, then b & c have to be pretty dissimilar too. [triangle inequality]
- Natural scenario:
  - Say two people are similar if either they work together or they live together.

Can we use this angle to help think about clustering?

Let's define objective like this:

- Given data set S of n objects.
- Each $x \in S$ has some (unknown) "ground truth" label $\ell(x)$ in $\{1,\ldots,k\}$.
- Goal: produce hypothesis h of low error up to isomorphism of label names:

$$Err(h) = \min_\sigma \Pr_{x \sim S}[\sigma(h(x)) \neq \ell(x)]$$

Like transductive learning from unlabeled data only. (could define inductive version too)

What conditions on a similarity function would be enough to allow one to cluster well?

Let's define objective like this:

- Given data set S of n objects.
- Each $x \in S$ has some (unknown) "ground truth" label $\ell(x)$ in $\{1,\ldots,k\}$.
- Goal: produce hypothesis h of low error up to isomorphism of label names:

$$Err(h) = \min_{\sigma} Pr_{x \sim S}[\sigma(h(x)) \neq \ell(x)]$$

Like transductive learning from unlabeled data only.
(could define inductive version too)

Here is an extremely restrictive condition that trivially works:

Say K is a good similarity function for a clustering problem if:

- $K(x,y) > 0$ for all x,y such that $\ell(x) = \ell(y)$.
- $K(x,y) < 0$ for all x,y such that $\ell(x) \neq \ell(y)$.

If we have such a K, then clustering is pretty trivial.

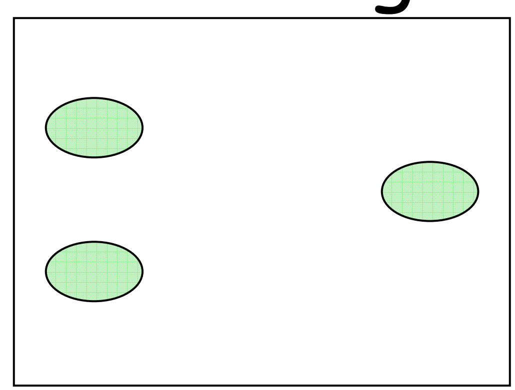
Now, let's try to make this condition a little bit less restrictive….

Proposal #2:

Say K is a good similarity function for a clustering problem if exists c such that:

- $K(x,y) > c$ for all x,y such that $\ell(x) = \ell(y)$.
- $K(x,y) < c$ for all x,y such that $\ell(x) \neq \ell(y)$.

Problem: the same K can be good for two very different clusterings of the same data!



Proposal #2:

Say K is a good similarity function for a clustering problem if exists c such that:

- $K(x,y) > c$ for all x,y such that $\ell(x) = \ell(y)$.
- $K(x,y) < c$ for all x,y such that $\ell(x) \neq \ell(y)$.

Problem: the same K can be good for two very different clusterings of the same data!

Big problem: unlike with learning, can't test your hypotheses!

Let's change our objective a bit…

to be to get a small (polynomial) number of clusterings such that at least one has low error.

- Like list-decoding

Now previous case is fine: exists c such that

- $K(x,y) > c$ for all x,y such that $\ell(x) = \ell(y)$.
- $K(x,y) < c$ for all x,y such that $\ell(x) \neq \ell(y)$.

Sort pairs by decreasing value of $K(x,y)$. Add in edges one at a time as in Kruskal. Output all (at most n) different clusterings produced.

How about our 1st defn for learning?

- $K:(x,y)\rightarrow[-1,1]$ is an (ε,γ)-good similarity function for P if at least a $1-\varepsilon$ prob mass of examples x satisfy:

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

- Extend to multi-class by requiring this to be true separately for all labels $j \neq \ell(x)$.
- ("P" = unif distr over S for transductive)

Can we use this to cluster?

How about our 1st defn for learning?

- $K:(x,y)\rightarrow[-1,1]$ is an (ε,γ)-good similarity function for P if at least a $1-\varepsilon$ prob mass of examples x satisfy:

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

- If # clusters k is small, each has $\Omega(1/k)$ prob mass, γ large, then can do:
  - Pick $O(k/\gamma^2 \log k/\delta)$ random points.
  - Try all $K^{O(k/\gamma^2 \ldots)}$ possible labelings of them.
  - Use to cluster remaining points.
  - Output all different clusterings produced.

How about our 1st defn for learning?

- $K:(x,y)\rightarrow[-1,1]$ is an (ε,γ)-good similarity function for P if at least a $1-\varepsilon$ prob mass of examples x satisfy:

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

- Ought to exist a more efficient algorithm.
- Maybe given x,y, determine if in same cluster by extent to which they agree on similarity to other examples z.
- Other natural defns/sufficient conditions?

How about our 1st defn for learning?

- $K:(x,y)\rightarrow[-1,1]$ is an (ε,γ)-good similarity function for P if at least a $1-\varepsilon$ prob mass of examples x satisfy:

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

- Other natural defns/sufficient conditions?
- E.g., usual notion of "good kernel": draw subsample S' and try <u>all possible</u> large-margin partitions of S'…. again exp'l in $K,1/\gamma$.

Open Problems

- Other/better definitions of "good similarity function" for learning. Ideally prove direct implications to standard algs like SVM etc.

  (But don't want a def like: "K is a good similarity function for P if Algorithm X works…")
- Other/better definitions of "good similarity function" for clustering.