Modern Topics in Learning Theory

Maria-Florina Balcan

04/26/2006

Modern Topics in Learning Theory

- Semi-Supervised Learning
- Active Learning
- Kernels and Similarity Functions
- Tighter Data Dependent Bounds

Outline

- Kernels & Large Margin Classifiers

  Hot topic in recent years

- Kernels as Features [BBV04]
- General Similarity Functions [BB06]

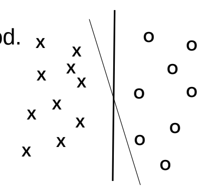
  Hot topic in the near future

Standard Supervised Learning

- X - instance space
- S={(x_i, l_i)} - set of labeled examples
  - x_1, x_2, … - drawn i.i.d. from some distr. D over X and labeled by target concept c
  - $l_i \in \{-1,1\}$ - binary classification
- Do some optimization over S to find h with small error over D.
  - err(h)=$P_{x \in D}[h(x) \neq c(x)]$ → the error of h w.r.t. to c (and D)

Linear Separators

- Well studied and understood.

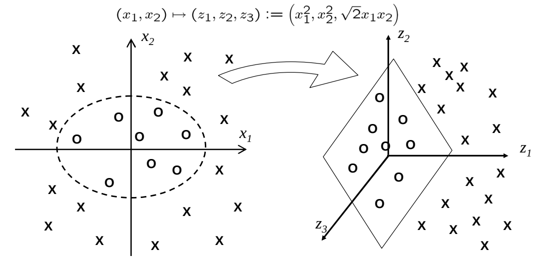


- Instance space: X=R^n
- Hypothesis class - class of linear decision surfaces in R^n
  - h(x)=w · x +b, if h(x) ≥ 0, then label x as positive (+1), otherwise label it as negative (-1)

Nonlinear Classification

- IDEA: Map each point to a higher dimensional feature space and construct linear separator in that higher dimensional space.

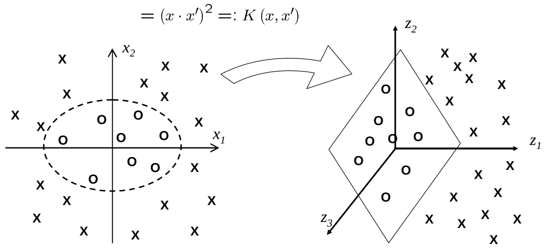
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$$



Nonlinear Classification

$$\phi : R^2 \to R^3 \qquad (x_1, x_2) \mapsto (z_1, z_2, z_3) := \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)$$

$$\phi(x)\cdot\phi(x') = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right)\left(x_1'^2, x_2'^2, \sqrt{2}x_1'x_2'\right)^T$$
$$= (x \cdot x')^2 =: K(x, x')$$



Kernels - Main Idea

- K(·,·) - kernel if it can be viewed as a **legal** definition of inner product:
  - ∃ ϕ: X → R^N such that K(x,y) =ϕ(x) · ϕ(y)
    - range of ϕ - "ϕ-space"
    - N can be very large
  - But think of ϕ as implicit, not explicit!
- Examples
  - Polynomial Kernel: X=R^n, K(x,y) =$(1+x \cdot y)^d$
    - n=3, d=2, ϕ: $R^3 \to R^{10}$

$$\phi(x) = (1, x_1^2, x_2^2, x_3^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)$$

Kernels

- Examples:
  - Linear: K(x,y)=x · y
  - Polynomial: K(x,y) =$(1+x \cdot y)^d$
  - Gaussian: $K(x,y) = \exp\left[-\frac{|x-y|^2}{2\sigma^2}\right]$
- Closure Properties
  - K(x,y)=K_1(x,y)+c
  - K(x,y)=c · K_1(x,y)
  - K(x,y)=K_1(x,y)+K_2(x,y)
  - K(x,y)=K_1(x,y)· K_2(x,y)
- Easily create new kernels using basic ones!

Kernels

K is a kernel iff

- K is symmetric
- for any set of training points $x_1, x_2, \ldots, x_m$ and for any $a_1, a_2, \ldots, a_m \in R$ we have:

$$\sum_{i,j} a_i a_j K(x_i, x_j) \geq 0$$

Kernels - Main Idea

Kernelizing algorithm

- If all computations involving instances are in terms of inner products then:
  - Conceptually, work in a very high diml space and the alg's performance depends only on linear separability in that extended space.
  - Computationally, only need to modify the alg. by replacing each x · y with a K(x,y).
- Examples: Perceptron, Voted Perceptron, SVM.

Lin. Separators: Perceptron algorithm

- **Algorithm:**
  - Start with all-zeroes weight vector w.
  - Given example x, predict positive ⇔ w · x ≥ 0.
  - On a mistake, update as follows:
    - Mistake on positive, then update w ← w + x
    - Mistake on negative, then update w ← w - x
- Easy to kernelize → w is a weighted sum of examples: $w = a_{i_1}x_{i_1} + \cdots + a_{i_k}x_{i_k}$
- So, replace $w \cdot x = a_{i_1}x_{i_1} \cdot x + \cdots + a_{i_k}x_{i_k} \cdot x$ with $a_{i_1}K(x_{i_1}, x) + \cdots + a_{i_k}K(x_{i_k}, x)$

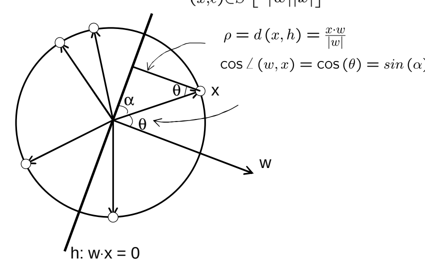
Do we have good generalization?

- Standard SC - the amount of data we need depends on VC-dim of the hypothesis class.
  - VC-dim for the class of linear sep. in R^m is m+1.
- Then, do we pay **a lot** from sample size point of view for going up?

Maria-Florina Balcan

Large Margin Classifiers

- If S is a set of **labeled** examples, then a vector w has margin γ w.r.t. S if $\min_{(x,\ell)\in S}\left[\ell\frac{w\cdot x}{|w||x|}\right]\geq\gamma$



Maria-Florina Balcan

Large Margin Classifiers, cont

- If S is a set of **labeled** examples, then a vector w has margin γ w.r.t. S if $\min_{(x,\ell)\in S}\left[\ell\frac{w\cdot x}{|w||x|}\right]\geq\gamma$
- A vector w has margin γ with respect to P (the combined distribution over labeled examples) if $\Pr_{(x,\ell)\in P}\left[\ell\frac{w\cdot x}{|w||x|}<\gamma\right]=0.$
- **If** large margin, then the amount of data we need depends only on γ and it's independent on the dimension of the (instance) space!

Maria-Florina Balcan

Large Margin Classifiers- Sample Complexity

- **If** large margin, then the amount of data we need depends only on $1/\gamma$ and is independent on the dim of the space!
  - If large margin γ and if our alg. produces a large margin classifier, then the amount of data we need depends only on $1/\gamma$ [Bartlett & Shawe-Taylor '99].
  - If large margin, then Perceptron also behaves well:
    - **Claim**: If the data is consistent with a linear threshold function specified by w^*, then the number of mistakes is at most $(1/\gamma)^2$, where γ is the margin of w^*
  - Another nice justification based on Random Projection [Arriaga & Vempala '99].

Maria-Florina Balcan

Kernels & Large Margins

- If S is a set of **labeled** examples, then a vector w in the ϕ-space has margin γ if:

$$\min_{(x,\ell)\in S}\left[\ell\frac{w\cdot\phi(x)}{|w||\phi(x)|}\right]\geq\gamma$$

- A vector w in the ϕ-space has margin γ with respect to P if: $\Pr_{(x,\ell)\in P}\left[\ell\frac{w\cdot\phi(x)}{|w||\phi(x)|}<\gamma\right]=0$
- A vector w in the ϕ-space has error α at margin γ if:

$$\Pr_{(x,l)\in P}\left[l\frac{w\cdot\phi(x)}{|w||\phi(x)|}<\gamma\right]\leq\alpha \qquad \boxed{(\alpha,\gamma)\text{-good kernel}}$$

Maria-Florina Balcan

Kernels & Large Margins, Summary

- Powerful combination in ML in recent years!
  - A kernel implicitly allows mapping data into a high dimensional space and performing certain operations there without paying a high price computationally.
  - If data indeed has a large margin linear separator in that space, then one can avoid paying a high price in terms of sample size as well.

Maria-Florina Balcan

Outline

- Kernels & Large Margin Classifiers

Hot topic in recent years

- Kernels as Features [BBV04]

- General Similarity Functions [BB06]

Hot topic in the near future

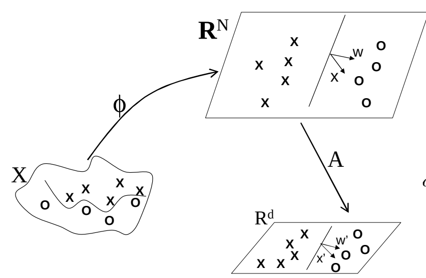
Kernels as Features [BBV04]

- **Main Idea:**
  - Designing a kernel function is much like designing a feature space.
  - Given a **good** kernel K, we can reinterpret K as defining a new set of $\tilde{O}(1/\gamma^2)$ features.

Kernels as Features [BBV04]

- If indeed large margin under K, then a random linear projection of the ϕ-space down to a low dimensional space approximately preserves linear separability.
  - by Johnson-Lindenstrauss lemma!

Main Idea - Johnson-Lindenstrauss lemma



$d = O\left(\frac{1}{\gamma^2}\log\frac{1}{\epsilon\delta}\right)$

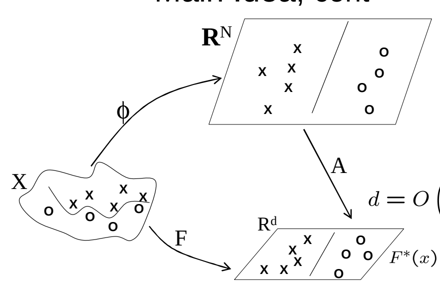
$w' = \frac{1}{\sqrt{d}}Aw$

$x' = \frac{1}{\sqrt{d}}Ax$

Main Idea - Johnson-Lindenstrauss lemma, cont

- For any vectors u,v with prob. (1-δ), $\measuredangle$(u,v) is preserved up to $\pm\,\gamma/2$, if we use $d = O\left(\frac{1}{\gamma^2}\log\frac{1}{\delta}\right)$
- Usual use in algorithms: m points, set δ=O(1/m^2)
- In our case, if we want w.h.p. $\exists$ separator of error ϵ, use $d = O\left(\frac{1}{\gamma^2}\log\frac{1}{\epsilon\delta}\right)$

Main Idea, cont



$d = O\left(\frac{1}{\gamma^2}\log\frac{1}{\epsilon\delta}\right)$

$F^*(x) = \frac{1}{\sqrt{d}}A\phi(x)$

- If c has margin γ in the ϕ-space, then F*(D,c), will w.h.p. have a linear separator of error at most ϵ.

Problem Statement

- For a given kernel K, the dimensionality and description of $\phi(x)$ might be large, or even unknown.
  - Do not want to explicitly compute $\phi(x)$.
- Given kernel K - produce such a mapping F efficiently:
  - running time that depends polynomially only on $1/\gamma$ and the time to compute K.
  - no dependence on the dimension of the "ϕ-space".

Main Result [BBV04]

- Positive answer - if our procedure for computing the mapping F is also given black-box access to the distribution D (unlabeled data).

Formally.....

- Given black-box access to $K(\cdot,\cdot)$, given access to D and γ, ϵ, δ, construct, in poly time, $F: X \to R^d$, where $d = O\left(\frac{1}{\gamma^2}\log\frac{1}{\epsilon\delta}\right)$ s. t. if c has margin γ in the ϕ-space, then with prob. $1-\delta$, the induced distribution in R^d is separable with error $\leq \epsilon$.

3 methods (from simplest to best)

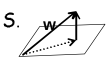
1. Draw d examples $x_1, \ldots, x_d$ from D. Use:
   $$F_0(x) = (K(x,x_1), \ldots, K(x,x_d)).$$
   For $d = (8/\epsilon)[1/\gamma^2 + \ln 1/\delta]$, if P was separable with margin γ in ϕ-space, then w.h.p. this will be separable with error ϵ. (but this method doesn't preserve margin).
2. Same d, but a little more complicated. Separable with error ϵ at margin $\gamma/2$.
3. Combine (2) with further projection as in JL lemma. Get d with log dependence on $1/\epsilon$, rather than linear. So, can set $\epsilon \ll 1/d$.

A Key Fact

Claim**:** If $\exists$ unit-length w of margin γ in ϕ-space, then if draw $x_1, \ldots, x_d \in D$ for $d \geq (8/\epsilon)[1/\gamma^2 + \ln 1/\delta]$, w.h.p. $(1-\delta)$ exists w' in $\mathrm{span}(\Phi(x_1),\ldots,\Phi(x_d))$ of error $\leq \epsilon$ at margin $\gamma/2$.

Proof: Let $S = \{\Phi(x)\}$ for examples x drawn so far.

- $w_{in} = \mathrm{proj}(w, \mathrm{span}(S))$, $w_{out} = w - w_{in}$.
- Say w_{out} is **large** if $\Pr_x\big(|w_{out}\cdot\Phi(x)| \geq \gamma/2\big) \geq \epsilon$; else **small**.
- If small, then done: $w' = w_{in}$.
- Else, next x has at least ϵ prob of improving S.
  $$|w_{out}|^2 \leftarrow |w_{out}|^2 - (\gamma/2)^2$$
- Can happen at most $4/\gamma^2$ times. □



A First Attempt

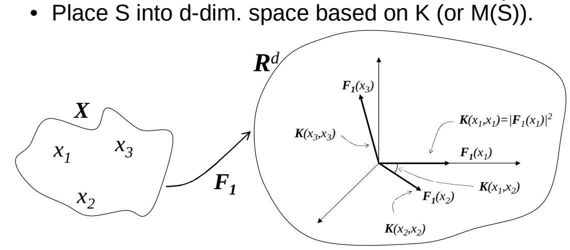
- If draw $x_1,\ldots,x_d \in D$ for $d = (8/\epsilon)[1/\gamma^2 + \ln 1/\delta]$, then whp exists w' in $\mathrm{span}(\phi(x_1),\ldots,\phi(x_d))$ of error $\leq \epsilon$ at margin $\gamma/2$.
- So, for some $w' = \alpha_1\phi(x_1) + \ldots + \alpha_d\phi(x_d)$,
  $$\Pr_{(x,l) \in P}[\mathrm{sign}(w' \cdot \phi(x)) \neq l] \leq \epsilon.$$
- But notice that $w'\cdot\phi(x) = \alpha_1 K(x, x_1) + \ldots + \alpha_d K(x, x_d)$.

$\Rightarrow$ vector $(\alpha_1,\ldots\alpha_d)$ is a separator in the feature space $(K(x,x_1), \ldots, K(x,x_d))$ with error $\leq \epsilon$.

- But margin not preserved because of length of target, examples.

A First Mapping

- Draw a set $S=\{x_1, \ldots, x_d\}$ of $d = O\left(\frac{1}{\epsilon}\left[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}\right]\right)$ unlabeled examples from D.
- Run K(x,y) for all $x,y \in S$, get $M(S)=(K(x_i,x_j))_{x_i,x_j \in S}$.
- Place S into d-dim. space based on K (or M(S)).

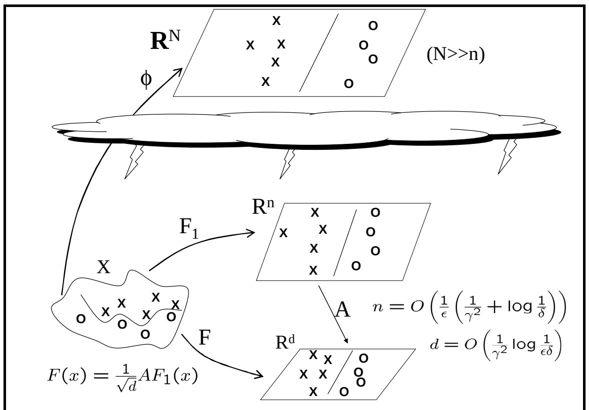


A First Mapping, cont

- What to do with new points?
- Extend the embedding F_1 to all of X:
  - consider F_1: $X \to R^d$ defined as follows: for $x \in X$, let $F_1(x) \in R^d$ be the point such that $F_1(x) \cdot F_1(x_i) = K(x,x_i)$, for all $i \in \{1, ..., d\}$.
- The mapping is equivalent to orthogonally projecting $\phi(x)$ down to span($\phi(x_1),\cdots, \phi(x_d)$).

An improved mapping

- A two-stage process, compose the first mapping, F_1, with a random linear projection.
- Combine two types of random projection:
  - a projection based on points chosen at random from D.
  - a projection based on choosing points uniformly at random in the intermediate space.



Improved Mapping - Properties

- Given ε, δ, γ <1, $d = O\left(\frac{1}{\gamma^2}\log(\frac{1}{\epsilon\delta})\right)$, if P has margin γ in the φ-space, then with probability ≥1-δ, our mapping into R^d, has the property that F(D,c) is linearly separable with error at most ε, at margin at most $\gamma/4$, given that we use $n = O\left(\frac{1}{\epsilon}\left[\frac{1}{\gamma^2}+\ln\frac{1}{\delta}\right]\right)$ unlabeled examples.

Improved Mapping, Consequences

- If P has margin γ in the φ-space then we can use $n = \tilde{O}(1/\gamma^4)$ unlabeled examples to produce a mapping into R^d for $d = O(\frac{1}{\gamma^2}\log\frac{1}{\epsilon'\gamma\delta})$, such that w.h.p. data is linearly separable with error $\ll \epsilon'/d$.
- The error rate of the induced target function in R^d is so small that a set S of $\tilde{O}(d/\epsilon')$ labeled examples will, w.h.p., be **perfectly** separable in R^d.
  - Can use any generic, zero-noise linear separator algorithm in R^d.

Implications, Open Problems

- Designing a kernel function -- designing a feature space.
- Alternative to "kernelizing" a learning algorithm:
  - rather than modifying the alg. to use kernels, construct instead a mapping into a low-diml space using the kernel and D; then run any un-kernelized alg over examples drawn from the mapped distribution.
- Open problem: Produce the desired mappings $F: X \to R^d$ in an oblivious manner (without access to D) for natural/standard kernels.

Outline

- Kernels & Large Margin Classifiers

Hot topic in recent years

- Kernels as Features [BBV04]

- General Similarity Functions [BB06]

Hot topic in the near future

General Similarity Functions

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
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).

A first Attempt: Definition satisfying properties (1) and (2)

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

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

How can we use it?

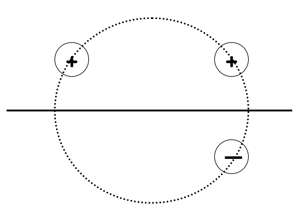
How to use it

At least a $1-\varepsilon$ prob mass of x satisfy:

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

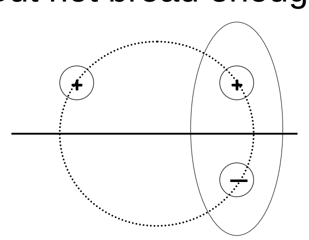
- Draw S^+ of $O(\gamma^{2}\ln(1/\delta^2))$ positive examples.
- Draw S^- of $O(\gamma^{2}\ln(1/\delta^2))$ negative examples.
- Classify x based on which gives better score.
- Hoeffding: for any given "good x", prob of error over draw of S^+,S^- at most δ^2.
- So, at most δ chance our draw is bad on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

But not broad enough



- $K(x,y)=x\cdot y$ has good (large margin) separator but doesn't satisfy the previous definition:
  - half of positives are more similar to negatives that to typical positives

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 Definition

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

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

- How to use it:
  - 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 feature space
    $w = [w(y_1),\ldots,w(y_n),-w(z_1),\ldots,-w(z_n)]$

Maria-Florina Balcan

And furthermore

Good Kernels are Good Similarity Functions

- 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.
  - $\varepsilon' = \varepsilon + \varepsilon_{extra}$, $\gamma' = \gamma^3\varepsilon_{extra}$.

Maria-Florina Balcan

And furthermore

Good Kernels are Good Similarity Functions

- An (ε,γ)-good kernel is an (ε',γ')-good similarity function under this definition.
  - $\varepsilon= \varepsilon + \varepsilon_{extra}$, $\gamma' = \gamma^3\varepsilon_{extra}$.

Proof (very rough 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 $w(y) \propto \ell(y)\cdot E[\text{weight}(y) \mid y \in S]$

Maria-Florina Balcan

Implications

- Provide the first rigorous explanation showing why a kernel is a good similarity function.
- Our algorithms do not require positive semidefinite functions!

Maria-Florina Balcan