blank

```
viewing in gv:
orientation: portrait
media: user defined width = 860; height = 680
use zoom in/out to achieve desired screen size

printing:
landscape shrink 90%
```

Kernel Methods for Learning Languages

Leonid Kontorovich
Carnegie Mellon University
joint work with
Corinna Cortes (Google Research) and Mehryar Mohri (Courant Institute)

February 20, 2007

The challenge

How do you learn a regular language from a labeled sample?

- teacher:
  - gives strings $x_i \in \Sigma^*$ and labels $y_i \in \{-1, +1\}$
  - tells you these are consistent with some regular language L

- you: produce a "reasonable" hypothesis $\hat{L}$

What do you do?

“Obvious” response

- construct minimal DFA $\hat{A}$ consistent with the sample:

$$x_i \in L(\hat{A}) \quad \text{iff} \quad y_i = +1$$

- Occam guarantee: $\mathrm{size}(\hat{A}) \le \mathrm{size}(A) \Rightarrow$ will achieve ϵ accuracy at δ confidence in

$$\mathrm{poly}(1/\epsilon, 1/\delta, \mathrm{size}(A))$$

  examples

- aside: is this PAC?

The catch

Finding the smallest DFA consistent with a set of accepted and rejected strings is **hard**

- Angluin, 1978 [1]; Gold, 1978 [3]: NP-complete
- Pitt and Warmuth, 1993 [5]: NP-hard even to *approximate* by $\tilde{A}$, with $\mathrm{size}(\tilde{A}) = \mathrm{poly}(\mathrm{size}(A))$
- Rivest and Schapire, 1987: "The problem of learning an automaton by passively observing [a labeled sample] is now well established to be a hard computational problem"

Our approach

- bottom line: want to learn a language, in some representation
- if automata are so hard to learn, maybe can find a representation better suited for learning?
- idea: embed the sample strings in a vector space and learn separating hyperplanes

Note about infinite-dimensional vectors

- if S is any countable set, $\mathbb{R}^{\mathbb{N}}$ and $\mathbb{R}^{S}$ are isomorphic
- vectors $\mathbf{w} \in \mathbb{R}^{\mathbb{N}}$ can be indexed by
  - number: $[\mathbf{w}]_i$, $i = 1, 2, \ldots$
  - set element: $[\mathbf{w}]_s$, $s \in S$
- the support of a vector, denoted $\|\mathbf{w}\|_0$, is the number of non-zero components of $\mathbf{w}$

Linear separability of general concept classes

- $\mathcal{C}$ is a concept class defined over a sample space $\mathcal{X}$
- fix some embedding $\phi : \mathcal{X} \to \{0,1\}^{\mathbb{N}}$
- require ϕ to have finite support: $\|\phi(x)\|_0 < \infty$ for all $x \in \mathcal{X}$
- concept $c \in \mathcal{C}$ is linearly separable (under ϕ) if there is a $\mathbf{w} \in \mathbb{R}^{\mathbb{N}}$ with $\|\mathbf{w}\|_0 < \infty$ s.t.

$$c = \{x \in \mathcal{X} : \langle \mathbf{w}, \phi(x) \rangle > 0\}$$

- concept class $\mathcal{C}$ is linearly separable if there is some ϕ with finite support s.t. all $c \in \mathcal{C}$ are linearly separable under ϕ

Important to use same ϕ for all $c \in \mathcal{C}$

- any $\mathcal{C} = \{c\}$ is trivially linearly separable!
- define $\phi : \mathcal{X} \to \{0,1\}$ by

$$\phi(x) \quad = \quad \mathbb{1}_{\{x \in c\}}$$

- it is only interesting to speak of linear separability for a collection of concepts
- we will always take $\mathcal{X}$ and $\mathcal{C}$ to be countable

Finite support of ϕ is important

Suppose we allow $\|\phi(x)\|_0 = \infty$. Then "embed by concept":

- define $\phi : \mathcal{X} \to \{0,1\}^{\mathcal{C}}$ by

$$[\phi(x)]_c \doteq \mathbb{1}_{\{x \in c\}}, \qquad c \in \mathcal{C}$$

- for any target $\hat{c} \in \mathcal{C}$, define the separating hyperplane $\mathbf{w} \in \mathbb{R}^{\mathcal{C}}$ by

$$[\mathbf{w}]_c \doteq \mathbb{1}_{\{c = \hat{c}\}}$$

  (note that $\|\mathbf{w}\|_0 = 1$)

- construction trivially ensures that

$$\langle \mathbf{w}, \phi(x) \rangle = \mathbb{1}_{\{x \in \hat{c}\}}, \qquad x \in \mathcal{X}$$

- still trivial even if we require $\|\phi(x)\|_p < \infty$, $p \geq 1$

Finite support of $\mathbf{w}$ is important

Suppose we allow $\|\mathbf{w}\|_0 = \infty$. Then "embed by instance":

- define $\phi : \mathcal{X} \to \{0,1\}^{\mathcal{X}}$ by

$$[\phi(x)]_u \doteq \mathbb{1}_{\{x=u\}}, \qquad u \in \mathcal{X}$$

(note that $\|\phi(x)\|_0 = 1$)

- for any target $\hat{c} \in \mathcal{C}$, define the separating hyperplane $\mathbf{w} \in \mathbb{R}^{\mathcal{X}}$ by

$$[\mathbf{w}]_u \doteq \mathbb{1}_{\{u \in \hat{c}\}}$$

- construction ensures

$$\langle \mathbf{w}, \phi(x) \rangle = \mathbb{1}_{\{x \in \hat{c}\}}$$

- again, $\|\mathbf{w}\|_p < \infty$ does not remove triviality

Other reasons to require $\|\phi\|_0, \|\mathbf{w}\|_0 < \infty$

- representation/storage: can't store infinitely many numbers!

- kernel

$$K(x, y) \quad = \quad \langle \phi(x), \phi(y) \rangle$$

  not necessarily well-defined for $\|\phi\|_0 = \infty$

- $\|\phi\|_0, \|\mathbf{w}\|_0 < \infty$ implies that for each $c \in \mathcal{C}$ there is a "separability dimension" $N(c) < \infty$ such that c is linearly separable in $\{0, 1\}^{N(c)}$ under ϕ

- linear separability guarantees positive margin in finite dimensions but not infinite

Generalization guarantee via margin

Theorem: Let $\mathcal{C}$ be a linearly separable concept class over $\mathcal{X}$ w.r.t. $\phi : X \to \{0,1\}^{\mathbb{N}}$. Define the class $\mathcal{F}$ of real-valued functions

$$\mathcal{F} \;=\; \{x \mapsto \langle \mathbf{w}, \phi(x) \rangle : \|\mathbf{w}\|_2 \leq 1\} .$$

There is a constant α_0 such that, for any distribution D over $\mathcal{X}$, for any concept $c \in \mathcal{C}$, there is an $R = R(c) > 0$ such that with probability at least $1 - \delta$ over m independently generated examples according to D, there exists a classifier $\mathrm{sgn}(f)$, with $f \in \mathcal{F}$, with generalization error no more than

$$\frac{\alpha_0}{m}\left(R \log^2 m + \log(1/\delta)\right)$$

Proof: Easy consequence of Bartlett and Shawe-Taylor [2].

*Why is this not PAC?

Is every concept class linearly separable?

Recall: we say $\mathcal{C}$ over $\mathcal{X}$ is linearly separable if there is a $\phi : \mathcal{X} \to \{0,1\}^{\mathbb{N}}$ with $\|\phi(\cdot)\|_0 < \infty$ s.t. for each $c \in \mathcal{C}$ there is a $\mathbf{w} \in \mathbb{R}^{\mathbb{N}}$ with $\|\mathbf{w}\|_0 < \infty$ s.t.

$$c = \{x \in \mathcal{X} : \langle \mathbf{w}, \phi(x) \rangle > 0\}$$

Two possible scenarios:

1. Every countable $\mathcal{C}$ over a countable $\mathcal{X}$ is linearly separable.

2. There exists some countable $\mathcal{C}$ over a countable $\mathcal{X}$ that is not linearly separable.

Which is it?

Deep/Trivial theorem

Theorem Every countable $\mathcal{C}$ over a countable $\mathcal{X}$ is linearly separable.

Proof Define two size functions:

$$|\cdot| : \mathcal{X} \to \mathbb{N}, \qquad \|\cdot\| : \mathcal{C} \to \mathbb{N}$$

with the property that each has finite level sets ($\#f^{-1}(n) < \infty$ for each $n \in \mathbb{N}$); in words, there are at most finitely many elements of a fixed size. Any countable set has such a size function.

We will define two auxiliary embeddings, χ and α, and will construct the requisite ϕ as their direct sum.

For intuition, it is helpful to keep in mind the dual roles of $\mathcal{X}$ and $\mathcal{C}$.

Fix a target $\hat{c} \in \mathcal{C}$.

Embedding by instance

- define $\chi : \mathcal{X} \to \{0,1\}^{\mathcal{X}}$ by

$$[\chi(x)]_u \doteq \mathbb{1}_{\{x=u\}}, \qquad u \in \mathcal{X}$$

- obviously, $\|\chi(x)\|_0 = 1$ for all $x \in \mathcal{X}$
- define $\mathbf{w}^\chi \in \mathbb{R}^{\mathcal{X}}$ by

$$[\mathbf{w}^\chi]_u \doteq \mathbb{1}_{\{u\in\hat{c}\}}\mathbb{1}_{\{|u|<\|\hat{c}\|\}}, \qquad u \in \mathcal{X}$$

- $\|\mathbf{w}^\chi\|_0 < \infty$ since size functions have finite level sets
-

$$\begin{aligned}
\langle \mathbf{w}^\chi, \chi(x)\rangle &= \sum_{u\in\mathcal{X}} [\mathbf{w}^\chi]_u[\chi(x)]_u \\
&= \sum_{u\in\mathcal{X}} \mathbb{1}_{\{u\in\hat{c}\}}\mathbb{1}_{\{|u|<\|\hat{c}\|\}}\mathbb{1}_{\{x=u\}} \\
&= \mathbb{1}_{\{x\in\hat{c}\}}\mathbb{1}_{\{|x|<\|\hat{c}\|\}}
\end{aligned}$$

Embedding by concept

- define $\alpha : \mathcal{X} \to \{0,1\}^{\mathcal{C}}$ by

$$[\alpha(x)]_c \doteq \mathbb{1}_{\{x \in c\}} \mathbb{1}_{\{\|c\| \leq |x|\}}, \qquad c \in \mathcal{C}$$

- since size functions have finite level sets, $\|\alpha(x)\|_0 < \infty$
- define $\mathbf{w}^\alpha \in \mathbb{R}^{\mathcal{C}}$ by

$$[\mathbf{w}^\alpha]_c \doteq \mathbb{1}_{\{c = \hat{c}\}}, \qquad c \in \mathcal{C}$$

- now

$$\begin{aligned}
\langle \mathbf{w}^\alpha, \alpha(x) \rangle &= \sum_{c \in \mathcal{C}} [\mathbf{w}^\alpha]_c [\alpha(x)]_c \\
&= \sum_{c \in \mathcal{C}} \mathbb{1}_{\{c = \hat{c}\}} \mathbb{1}_{\{x \in c\}} \mathbb{1}_{\{\|c\| \leq |x|\}} \\
&= \mathbb{1}_{\{x \in \hat{c}\}} \mathbb{1}_{\{|x| \geq \|\hat{c}\|\}}
\end{aligned}$$

Combining the two embeddings

- define the canonical embedding $\phi : \mathcal{X} \to \{0,1\}^{\mathbb{N}}$ by

$$\phi(x) = \chi(x) \oplus \alpha(x);$$

  note that

$$\|\phi(x)\|_0 = \|\chi(x)\|_0 + \|\alpha(x)\|_0 < \infty$$

- define

$$\mathbf{w} = \mathbf{w}^{\chi} \oplus \mathbf{w}^{\alpha};$$

  again,

$$\|\mathbf{w}\|_0 = \|\mathbf{w}^{\chi}\|_0 + \|\mathbf{w}^{\alpha}\|_0 < \infty$$

- by above calculations,

$$\begin{aligned} \langle \mathbf{w}, \phi(x) \rangle &= \langle \mathbf{w}^{\chi}, \chi(x) \rangle + \langle \mathbf{w}^{\alpha}, \alpha(x) \rangle \\ &= \mathbb{1}_{\{x \in \hat{c}\}} \mathbb{1}_{\{|x| < \|\hat{c}\|\}} + \mathbb{1}_{\{x \in \hat{c}\}} \mathbb{1}_{\{|x| \geq \|\hat{c}\|\}} \\ &= \mathbb{1}_{\{x \in \hat{c}\}}. \end{aligned}$$

$\square$

Deterministic finite-state automata (DFA)

- a DFA is a tuple $A = (\Sigma, Q, F, \delta)$ where
  - Σ is a finite alphabet
  - $Q = \{1, 2, \ldots\}$ is a finite set of states
  - $F \subset Q$ is the set of the accepting states
  - $\delta : Q \times \Sigma \to Q$ is the transition function
  - starting state $q_0 \equiv 1$

- $\mathrm{DFA}(n, \Sigma)$ is the set of all DFAs over Σ with $|Q| = n$

- sanity check:

$$|\mathrm{DFA}(n, \Sigma)| = 2^n n^{n|\Sigma|}$$

huge, but easy to compute

Universal kernel for regular languages

- take $\mathcal{X} = \Sigma^*$ as the instance space; $|x|$ is string length of $x \in \Sigma^*$
- take $\mathcal{C} = \bigcup_{n=1}^{\infty} \mathrm{DFA}(n, \Sigma)$ as the concept space; $\|A\|$ is number of states of $A \in \mathcal{C}$
- for $A \in \mathcal{C}$, let $L(A)$ be the (regular) language accepted by A
- let $\phi : \Sigma^* \to \{0,1\}^{\mathbb{N}}$ be the canonical embedding for $\mathcal{X}$, $\mathcal{C}$
- define the universal regular kernel $K(\cdot,\cdot) : \Sigma^* \times \Sigma^* \to \mathbb{R}$ by

$$\begin{aligned} K(x,y) &= \langle \phi(x), \phi(y) \rangle \\ &= \mathbb{1}_{\{x=y\}} + \sum_{n=1}^{\min\{|x|,|y|\}} K_n(x,y) \end{aligned}$$

where

$$K_n(x,y) = \sum_{A \in \mathrm{DFA}(n,\Sigma)} \mathbb{1}_{\{x \in L(A)\}} \mathbb{1}_{\{y \in L(A)\}}$$

Regular languages are linearly separable

It follows from the main theorem that

- any regular language L, as a Boolean function on Σ^*, is linearly separable under the canonical embedding ϕ in some finite-dimensional space $\mathbb{R}^N$, where $N = N(L)$

- furthermore, any regular language is linearly separable with a strictly positive margin

- any regular language L can be represented by some "support strings" $\{s_i \in \Sigma^* : 1 \leq i \leq m\}$ with weights $\alpha \in \mathbb{R}^m$:

$$L = \left\{ x \in \Sigma^* : \sum_{i=1}^{m} \alpha_i K(s_i, x) > 0 \right\}$$

- additionally, any language separated by K is regular (follows from LK, Cortes, Mohri [4, Thm. 5]

A complication

- don't know how to compute $K_n(x, y)$ efficiently
- a likely candidate for #P-complete (I thought)
- but a famous computer scientist thinks otherwise!
- anybody care to settle this?

Efficient ϵ-approximation for K_n

- easy to draw an automaton A uniformly at random from $\mathrm{DFA}(n, \Sigma)$
- for $x, y \in \Sigma^*$, define $S_{x,y}^{(n)} \subset \mathrm{DFA}(n, \Sigma)$ by

$$S_{x,y}^{(n)} = \{A \in \mathrm{DFA}(n, \Sigma) : \{x, y\} \subset L(A)\};$$

  this is the set of the n-state DFAs that accept both x and y
- note that $K_n(x, y) = \left| S_{x,y}^{(n)} \right|$
- let P be the uniform distribution on $\mathrm{DFA}(n, \Sigma)$
-

$$K_n(x, y) = P(S_{x,y}^{(n)}) \left| \mathrm{DFA}(n, \Sigma) \right|$$

Approximating $P(S_{x,y}^{(n)})$

- generate m automata $\{A_i : 1 \leq i \leq m\}$ independently and uniformly from $\mathrm{DFA}(n, \Sigma)$
- define

$$\hat{P}(S_{x,y}^{(n)}) \quad = \quad \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{x \in L(A_i)\}} \mathbb{1}_{\{y \in L(A_i)\}}$$

- by multiplicative Chernoff's bound,

$$\mathbf{P}\left\{ \left| \hat{P}(S_{x,y}^{(n)}) - P(S_{x,y}^{(n)}) \right| > \epsilon P(S_{x,y}^{(n)}) \right\} \quad \leq \quad 2 \exp(-\epsilon^2 m P(S_{x,y}^{(n)})/3)$$

Approximating K_n

- with probability at least $1 - 2\exp(-\epsilon^2 m P(S_{x,y}^{(n)})/3)$, we have

$$(1-\epsilon)\hat{K}(x,y) \le K(x,y) \le (1+\epsilon)\hat{K}(x,y),$$

  where $\hat{K}_n(x,y) \doteq \hat{P}(S_{x,y}^{(n)})\,|\mathrm{DFA}(n,\Sigma)|$

- need

$$m \quad \ge \quad \frac{3\log(2/\alpha)}{\epsilon^2 P(S_{x,y}^{(n)})}$$

  sampling steps to have an ϵ-approximation to $K(x,y)$ with probability at least $1-\alpha$.

- potential problem: what if $P(S_{x,y}^{(n)})$ is really small?

Fortunately, not a real problem

Claim: For all $n \geq 1$, for all $x, y \in \Sigma^*$, we have

$$\frac{1}{4} \leq P(S_{x,y}^{(n)}) \leq \frac{1}{2}.$$

Proof: Essentially, by symmetry.

The upper bound is simple – it follows from the fact that $K_n(x,x) = \frac{1}{2}\,|\mathrm{DFA}(n,\Sigma)|$. Indeed, for any $x \in \Sigma^*$, for every $A^+ \in \mathrm{DFA}(n,\Sigma)$ that accepts x there is exactly one $A^- \in \mathrm{DFA}(n,\Sigma)$ that does not (obtained by changing the state in which A^+ ends up after reading x from accepting to non-accepting). The upper bound follows from the obvious relation $K_n(x,y) \leq K_n(x,x)$ for all $x, y \in \Sigma^*$.

To prove the lower bound, take the "worst" case where $x, y \in \Sigma^*$ are such that every $A \in \mathrm{DFA}(n,\Sigma)$ has $\delta(1,x) \neq \delta(1,y)$. In other words, no automaton ends up in the same state after reading x and as it does after reading y (recall that all automata start in state $q_0 = 1$). Since every state is independently chosen to be accepting or not with equal probability, exactly one-fourth of all $A \in \mathrm{DFA}(n,\Sigma)$ will accept both x and y. Clearly, this fraction will be higher if we allow some automata to end up in the same state upon reading x and y. $\square$

Open problems

- Can $K(x, y)$ be efficiently computed? Is there any universal regular kernel that's efficiently computable?

- Running SVM with kernel K on a labeled sample S yields a maximum-margin hyperplane which corresponds to some automaton A. Is A the smallest automaton consistent with S?

- Is there any universal kernel for which we can prove a result of type "margin is at least $1/\operatorname{poly}(\operatorname{size}(\hat{A}))$"?

- If such a kernel were efficiently computable, this would pretty much solve the problem of learning regular languages!

References

[1] D. Angluin. "On the complexity of minimum inference of regular sets." Information and Control. 39(3):337–350, 1978.

[2] Peter Bartlett and John Shawe-Taylor, "Generalization performance of support vector machines and other pattern classifiers." Advances in kernel methods: support vector learning, 43–54, 1999.

[3] E. M. Gold. "Complexity of automaton identification from given data." Information and Control. 37(3):302–420, 1978.

[4] Leonid Kontorovich, Corinna Cortes, Mehryar Mohri, "Learning Linearly Separable Languages." ALT 2006.

[5] L. Pitt and M. Warmuth. "The minimum consistent DFA problem cannot be approximated within any polynomial." Journal of the Association for Computing Machinery, 40(1):95–142, 1993.