Joachims '99: discussion guide

Prepared by Lillian Lee
CS 6740 / 4758 / 6300
Sept 4, 2012

Plan:

2-5 ~~mins~~: general remarks; further background (since this was a majorly lauded paper)

5-10 mins: ~~remarks on what~~ discuss what the results tell us

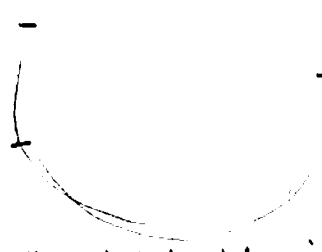
10 mins: discuss the transduction setting

10 mins: how can we make our own transductive classifier, and become as famous as Thorsten?
(spoiler: no complete answer yet – really want to show that the structural risk minimization framework opens up interesting possibilities)

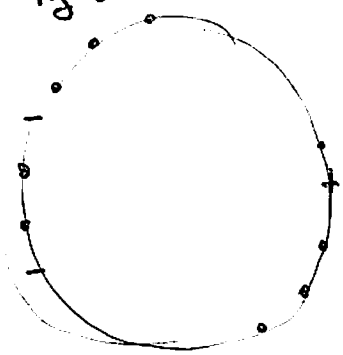
- 10-years-later award for most lasting impact of an ICML '99 paper
- retrospective slides available from Thorsten's homepage
- code for a different (more efficient) transductive learner available from Thorsten's website; reference: Joachims 2005, spectral graph transduction

⊛ Thorsten was a grad student when he wrote this paper!
(Dream big!)

Main intuition: (see also Fig 2, Vapnik '98 fig 8.2)



in this little-labeled data situation, best pick seems to be $\vec{w} = (1,0)$ or 0°.



But if the test data looks like this, a better choice, intuitively, is $\vec{w} = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})$ or −45°, say. There wasn't enough training data to "see" this pattern.

— about the results.

Goal: show that use of unlabelled data's structure can overcome lack of labelled data, all else being equal.

"Punchline" data: Fig 5 ; Fig 6

↙ fixed |train|=17

↘ vary |train|, TSVM almost always better than SVM, version w/out test structure.

These are arguably minor points; remember our focus on the positive!

(Trick?) question 1: aren't these results just the result of overfitting the test set?

question 2: what influenced the x-scale?

question 3 (maybe just my misunderstanding): why a diff. defn of breakeven point for TSVM's?

[It's not: if $\frac{f_{++}}{f_{++}+f_{+-}} = \frac{f_{++}}{f_{++}+f_{-+}}$, then $f_{+-}=f_{-+}$.]

question 4: what influenced diff. choices of features for fig 8, 9?

Nice to include the discussion of the "project" class. What do we learn from it?

Thoughts re: transductive setting

(★) use the structure of the "entirety" of test set as extra info

(1) → do we really need to label all the (test) points? (what if only precision matters?) [estimating proportions]

(3) → can we request (which) items to label? ("verify the margins")

→ is this really "learning"?

(should evaluate over several test sets, to defuse claims of overfitting)

(2) → does this relate to personalization?

(offhand: why would each user be getting their own distribution of points? An example is my email)

(5) (4) → can we use the transductive idea for non-classification tasks, like simplification or translation?

(4) → is there a way to "tell" if transduction is going to be not a useful way to go?

(cf. e.g. Joachims 2003)

(again)

V98 = Vapnik, Statistical Learning Theory, 1998

Q: What from the theory/concept of transductive inference can be useful for general NLP tasks?

V98, §0.8, "The Structural Risk Minimization Principle", in plain English: eq 6 in J99

Test set error is bounded above for all ~~functions~~ hypothesis functions in a given class by a function of the # of training errors, # size of sample, and [this is new] VC-dimension ("expressivity"/"complexity"/"capacity") of the class of hypothesis functions.

=>(!) picking the class of hypotheses is important too (~~lower comp~~ in fact, lower "complexity" ~~reduces~~ reduces the confidence interval that is part of the bound. But, more "complex" hypothesis classes allow us to lower training error, remember.

(skip next if not interested in the theoretical argument & and its origins)
But note typo in thm 1 as stated in paper.

V98 ~~§8~~ Thm 8.2: prob that some rule among a set w/ N equivalence classes on train + test exceeds ε is bounded by a term Nx (some other function) ← has deviation that

V98 Thm 8.4: for linear decision rules, we can set up a nested sequence of "structures" embedded (each a group of equivalence classes on rules) ~~and bound N_r for each H_r~~. $H_r \in H_{r+1} \subset \ldots$

~~Corresponds to J98 Thm 1.~~

~~(note typo: it should be $N_r < \exp(d(\ln \frac{n+k}{d} + 1))$)~~

V98 §8.5: "complexity" relates to "margin" separating the points, in the way all linear decision rules in a given ~~equivalence~~ class do.

- V98, §10.9: how to find the optimal hyperplane in the transductive setting.
Corresponds to J98's OP1.

This is still true; it's crossed out only because it's not germane to the argument at hand.

"~~Main~~ Main principle of inference from small sample size" (§0.9): try to solve problem directly, not solve a more general problem as an intermediate step, b/c you might not have enough data for the general problem.

To understand if we could apply these ideas to other settings ...

V98 § 8.5, how do we get to the "convex hull" ~~interpret~~ interpretation?

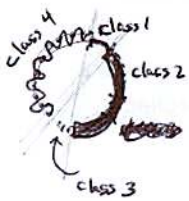
- the available data gives us the entire set (and finite) of equivalence classes. [pg. 350]
- ~~important idea: w/out test data, this isn't true.~~
- any subset of the entire set is a class of possible hypotheses, so one "line of attack" is to pick a "low-capacity" such subset. [SRM principle, § 0.8]

- if we could think of the ~~entire set~~ subsets as "structured", forming a nested sequence,

$$H_1' \subset H_2' \subset \cdots \subset H'$$

↖ ~~set~~ set of possible hypotheses, as a union of equivalence classes

then the VC dimensions ~~etc~~ of the H_i''s are ~~nondec~~ monotonically increasing in i. [pg. 222]



- in the linear-separator case, (let's assume, again, constraint to pass through origin), each equivalence class = region on unit hypersphere. (see fig 8.1 of Vapnik)

We can make H_i' contain only those regions (classes) that are bigger than a threshold c_i, ~~$c_1 > c_2 > \cdots$~~. This makes the H_i''s nested, as desired. But the size of the regions may not be easy to compute, so we don't know which to pick.

Let's change our definition of "size" of region ~~to~~ (= an equivalence class of hypotheses, so they all give the same labels to all the available points) to: (normalized) distance btwn the convex hull of the points labeled positive, and the convex hull of the points labeled negative by those fns.

Then we set H_i' = regions (classes) with "size" bigger than ~~the~~ margin for that class

~~"We can then show that (Thm 8.4) that~~

It then can be shown that for any classifier in H_i', the error bound is, roughly, dependent on i and the empirical risk (with probability of being exceeded a parameter)

So, our "best" choice is: among all classifiers with training error 0, go with the one that ~~(is in the class of fns~~ yields biggest "convex hull distance" between the separated classes. (Since it's the best among all fns in the hypothesis class imposing that margin.)

See Thorsten's spectral graph transduction paper (2003) for a different nested structure; its relation to k-nn.

Q: sequential data "geometric" representations?