Transductive Learning for Text Categorization using Support Vector Machines

Thorsten Joachims

ICML 1999

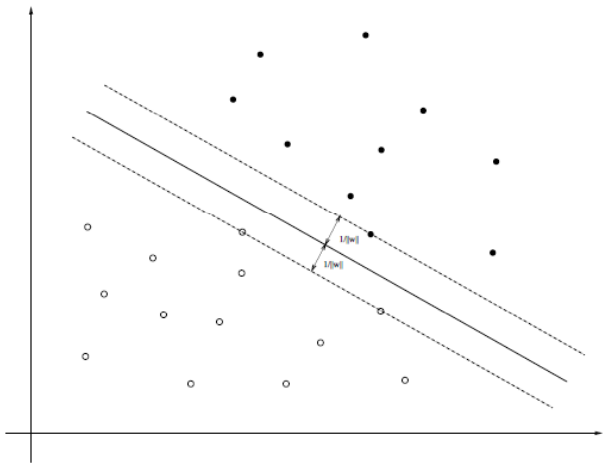
Some slides from Thorsten's "Best Paper from ICML 1999 Retrospective"

SVMs

- Find the *optimal* linear separator
- *Optimal* = the largest **margin** between it and the positive examples on one side and the negative examples on the other
- **Margin** refers to the separation between the positive and negative examples
- The points closest to the separator are called the **support vectors**

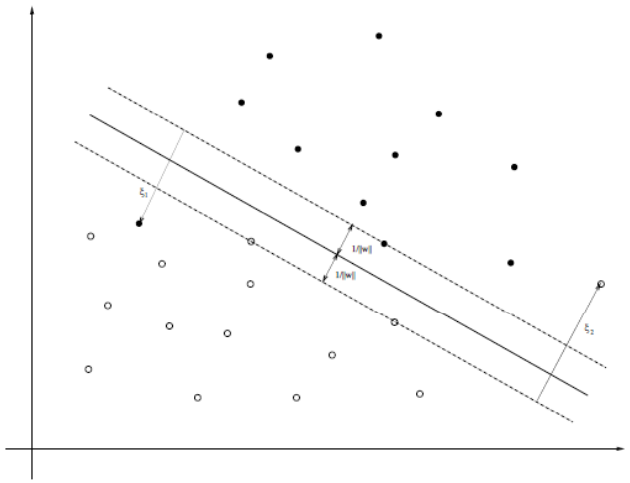
SVMs

optimal
margin
classifier



SVMs

soft
margin
classifier

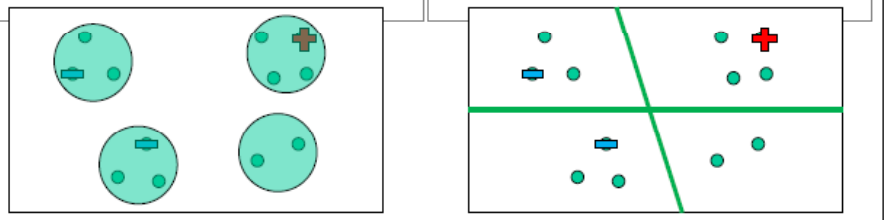


Definitions

- i.i.d. sample
  - Independent and identically distributed
  - Presumed for both training and test sets
- H refers to the space of possible decision functions
- VC-dimension: don't worry about this for this course
  - It refers to the capacity/complexity of H

Input

Tom Mitchell	Vladimir Vapnik
"What can we do with all the text data on the web?"	**Transduction: Predicting only at known locations is easier**
– [Blum/Mitchell] Co-training • Exploit redundant representations – [Nigam/McCallum/Thrun/Mitchell] Semi-supervised Naïve Bayes • Generatively model clusters in P(X) • Mixture model	– Finite number of predictions vs. continuous function – Define margin w.r.t. test points – Generalization error bounds



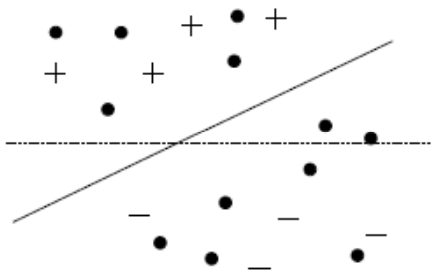
Induction vs. Transduction

- Induction
  - Induce a decision function with a low error rate on the whole distribution of examples for the learning task
  - Unnecessarily complex
    - Don't always care about the decision function
- Transduction
  - When we care **most** that we classify a particular *test set* of examples with as few errors as possible.
  - Setting: small training set; large test set.

Benefit of studying the test set?

- Training and test sets split H into a finite number of equivalence classes
  - Two functions from H belong to the same equivalence class if they both classify the training and test sample in the same way
  - Simplifies the learning problem
- Can examine the layout of the test examples when constructing the classifier

TSVM Learning

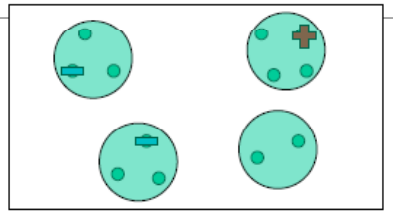


Input

Tom Mitchell

"What can we do with all the text data on the web?"

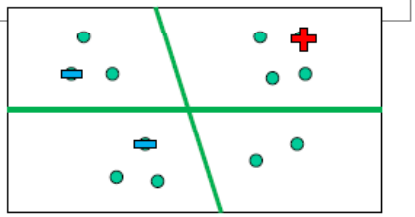
- [Blum/Mitchell] Co-training
  - Exploit redundant representations
- [Nigam/McCallum/Thrun/Mitchell] Semi-supervised Naïve Bayes
  - Generatively model clusters in P(X)
  - Mixture model



Vladimir Vapnik

Transduction: Predicting only at known locations is easier

- Finite number of predictions vs. continuous function
- Define margin w.r.t. test points
- Generalization error bounds



TSVMs for text categorization

- Why?
  - SVMs have been shown to be good for text categorization
  - Can exploit the strong *co-occurrence* patterns that appear in text

Altavista (1999)
- hits(pepper & salt) → 327K
- hits(pepper & physics) → 4.2K
- hits(physics) > hits(salt)

Google (2009)
- hits(pepper & salt) → 159M
- hits(pepper & physics) → 1.3M
- hits(physics) = 107M > hits(salt) = 56M

Text categorization with co-occurrence patterns

	nuclear	physics	atom	pepper	basil	salt	and
D1	**1**						1
D2	1	1	1				1
D3			1				1
D4				1	1		1
D5				1		1	1
D6					**1**	**1**	1

Learning TSVMs

- Basic idea?

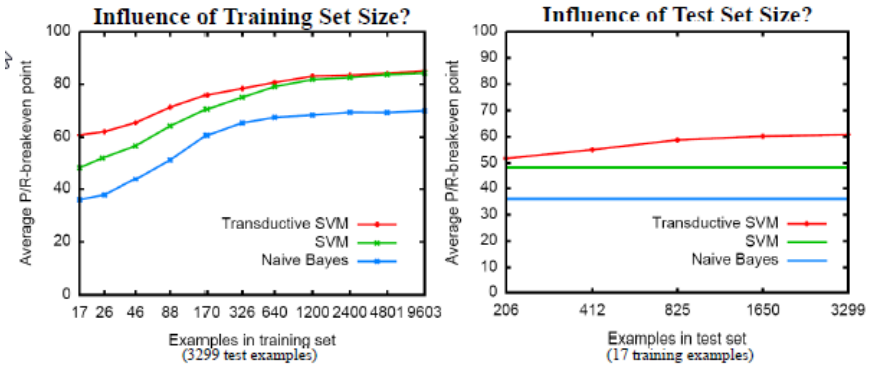
Experiment: Reuters-21587

- Top 10 categories
- ~12000 features
  - after stemming and stopword removal
- Macro-averaged precision/recall break-even point

	Bayes	SVM	TSVM
earn	78.8	91.3	95.4
acq	57.4	67.8	76.6
money-fx	43.9	41.3	60.0
grain	40.1	56.2	68.5
crude	24.8	40.9	83.6
trade	22.1	29.5	34.0
interest	24.5	35.6	50.8
ship	33.2	32.5	46.3
wheat	19.5	47.9	54.4
corn	14.5	41.3	43.7
average	35.9	48.4	60.8

Figure 5: P/R-breakeven point for the ten most frequent Reuters categories using 17 training and 3,299 test examples. Naive Bayes uses feature selection by empirical mutual information with local dictionaries of size 1,000. No feature selection was done for SVM and TSVM.

Experiment: Reuters-21587



Experiment: WebKB

- 4 classes
- 9 training examples, 3957 test examples
- P/R break-even point per class (and average)

	Bayes	SVM	TSVM
course	57.2	68.7	93.8
faculty	42.4	52.5	53.7
project	21.4	37.5	18.4
student	63.5	70.0	83.8
macro-average	46.1	57.2	62.4

For Projects: Experimental Methodology

- Train/validation/test set division
- Performance measures
- Proper comparisons
  - Alternative algorithms
  - Multiple data sets
  - Proper baseline
  - Variations of key parameters