

CSCI 5622 Machine Learning

ML Semi-Supervised Learning

DATE	READ	DUE
Today, Nov 9	<u><i>Semisupervised Learning</i></u>	Exprmnt 1 Write-up
Mon, Nov 11	Active Learning	
Wed, Nov 16	Active Learning cont...	Exp 1 Peer Feedback

www.RodneyNielsen.com/teaching/CSCI5622-F09/

Instructor: Rodney Nielsen

Assistant Professor Adjunct, CU Dept. of Computer Science

Research Assistant Professor, DU, Dept. of Electrical & Computer Engr.

Research Scientist, Boulder Language Technologies

ML Semi-Supervised Learning (SSL)

- Supervised ML often requires a significant amount of data to achieve reasonable results
 - SRL: 1M words text \rightarrow \sim 2M training examples
 - \sim 80% F-measure on out-of-domain data
- But Google indexes trillions of words
- Can we use all of that unlabeled data to improve ML performance

ML Applications

- Virtually all NLP tasks and applications:
 - POS tagging, Parsing, Semantic Role Labeling, Entity and event detection, ...
 - Information Extraction, Information Retrieval, Document Classification, Educational Assessment,...
- Virtually all image processing tasks & apps:
 - Edge detection, shape & texture recognition,...
 - Face recognition, object recognition,...
- Vertical profiling: waveform identification,...

ML Document Classification

Text Classification from Labeled and Unlabeled Documents using EM

KAMAL NIGAM[†] knigam@cs.cmu.edu
ANDREW KACHITES MCCALLUM[‡] mccallum@justresearch.com
SEBASTIAN THRUN[†] thrun@cs.cmu.edu
TOM MITCHELL[†] tom.mitchell@cmu.edu

[†]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
[‡]Just Research, 4616 Henry Street, Pittsburgh, PA 15213

Received March 15, 1998; Revised February 20, 1999

Editor: William W. Cohen

Abstract. This paper shows that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled documents. This is important because in many text classification problems obtaining training labels is expensive, while large quantities of unlabeled documents are readily available.

We introduce an algorithm for learning from labeled and unlabeled documents based on the combination of Expectation-Maximization (EM) and a naive Bayes classifier. The algorithm first trains a classifier using the available labeled documents, and probabilistically labels the unlabeled documents. It then trains a new classifier using the labels for all the documents, and iterates to convergence. This basic EM procedure works well when the data conform to the generative assumptions of the model. However these assumptions are often violated in practice, and poor performance can result. We present two extensions to the algorithm that improve classification accuracy under these conditions: (1) a weighting factor to modulate the contribution of the unlabeled data, and (2) the use of multiple mixture components per class. Experimental results, obtained using text from three different real-world tasks, show that the use of unlabeled data reduces classification error by up to 30%.

Keywords: text classification, Expectation-Maximization, integrating supervised and unsupervised learning, combining labeled and unlabeled data, Bayesian learning

1. Introduction

Consider the problem of automatically classifying text documents. This problem is of great practical importance given the massive volume of online text available through the World Wide Web, Internet news feeds, electronic mail, corporate databases, medical patient records and digital libraries. Existing statistical text learning algorithms can be trained to approximately classify documents, given a sufficient set of labeled training examples. These text classification algorithms have been used to automatically catalog news articles (Lewis & Gale, 1994; Joachims, 1998) and web pages (Craven, DiPasquo, Freitag, McCallum, Mitchell, Nigam, & Slattery, 1998; Shavlik & Eliassi-Rad, 1998), automatically learn the reading interests of users (Pazzani, Muramatsu, & Billsus, 1996; Lang, 1995), and automati-

0 aardvark
1 abstract
4 computer
0 dumpling
178 labeled
92 learning
17 machine
7 science
212 unlabeled
0 zephyr

ML If we don't have all the labels?

	CLS	dumpling	labeled	learning	machine	science	zephyr
	LBL						
Doc ₁	1	0	178	92	17	7	0
Doc ₂	0	15	1	1	0	0	1
Doc ₃	1	0	28	32	10	3	1
Doc ₄	?	0	0	7	5	5	0
Doc ₅	?	0	0	5	0	0	5

- Use Semi-Supervised Learning
- One option is EM

ML Apply EM to Naïve Bayes

- Initialize
 - $P(y=c_k)$ or $P(c_k)$ and $P(w_v|c_k)$
 - based on the labeled data
- E-Step
 - Compute *expected* value for doc. class labels
- M-Step
 - *Maximize* the likelihood of data given the labels

ML Ex: One Attribute & Two Classes

- E-Step (Expectation):

$$\begin{aligned} E[y_k^{(n)}] &= P(c_k | doc_n; \theta) \\ &= \frac{P(c_k | \theta) P(doc_n | c_k; \theta)}{P(doc_n | \theta)} \\ &= \frac{P(c_k | \theta) \prod_{w_i \in doc_n} P(w_i | c_k; \theta)}{\sum_{h=1}^K \left(P(c_h; \theta) \prod_{w_i \in doc_n} P(w_i | c_h; \theta) \right)} \end{aligned}$$

ML Ex: One Attribute & Two Classes

- (Maximization) M-Step:
 - Recompute θ

$$P(c_k | \theta_{t-1}) = \frac{1 + \sum_{i=1}^N P(c_k | doc_n; \theta_{t-1})}{N + K}$$

$$P(w_l | c_k; \theta_{t-1}) = \frac{1 + \sum_{n=1}^N N(w_l, doc_n) P(c_k | doc_n; \theta_{t-1})}{|V| + \sum_{s=1}^{|V|} \sum_{n=1}^N N(w_s, doc_n) P(c_k | doc_n; \theta_{t-1})}$$

ML When will this work?

- If and only if the documents in a given class are relatively consistent as if generated by a single process

ML Questions

- Questions???

ML Co-Training

- Allow two classifiers given independent views to learn from one another
- Each attribute set (view) must be sufficient
- Each view must be independent of the other given the class labels

ML Ex: Document Classification

- Classify computer science department web pages as course home pages
- Two views:
 - The words on the page to be classified
 - The displayed text of the hyperlinks that connect to the page to be classified

ML *The Co-Training Algorithm*

- Given
 - Labeled data L
 - Unlabeled data U
- Loop
 - Train h_1 (hyperlink classifier) using instances in L
 - Train h_2 (page classifier) using instances in L
 - Have h_1 and h_2 each classify instances in U
 - Have h_1 and h_2 each add p positively- and n negatively-labeled instances from U to L

ML Ex: Document Classification

- Classify course home pages
- Two views: pages versus hyperlinks
- Started with just 3 positive examples and 9 negative examples
- 1000 additional unlabeled examples
- Achieved an error rate of 5%
- Reduced the error rate by over 50% over just using the labeled data

ML Theorem

- If
 - X and X' (the two views) are conditionally independent given the labels, and
 - f is PAC learnable using either view
- Then
 - f is PAC learnable from weak initial classifiers plus unlabeled data

ML Questions

- Questions???

ML One View of the Space of SSL

Method	Use feature split?	
	Yes	No
Incremental		
Iterative		

Nigam and Ghani, 2000

ML One View of the Space of SSL

Method	Use feature split?	
	Yes	No
Incremental	Co-training	
Iterative		

ML One View of the Space of SSL

Method	Use feature split?	
	Yes	No
Incremental	Co-training	
Iterative		EM

- Like Co-training: 2 separate views of the data
- Like EM: Iteratively update unobserved values
 - Relabel all the data using one classifier
 - Train the second classifier on the new labels and relabel the data for the first classifier
 - Iterate until convergence

ML One View of the Space of SSL

Method	Use feature split?	
	Yes	No
Incremental	Co-training	
Iterative	Co-EM	EM

- Start with a small labeled dataset
- Generate a class probability distribution for all the unlabeled data
- Add the unlabeled example that the classifier is most confident about to the labeled data pool
- Iterate until all examples have been added

ML One View of the Space of SSL

Method	Use feature split?	
	Yes	No
Incremental	Co-training	Self-training
Iterative	Co-EM	EM

ML One View of Data Space of SSL

Natural feature split

No natural feature split

ML Performance of SSL Methods

Natural feature split

No natural feature split

Method	Use natural feature split?	
	Yes	No
Incremental	Co-training 3.7	Self-training 5.8
Iterative	Co-EM 3.3	EM 8.9

Method	Use <i>random</i> feature split?	
	Yes	No
Incremental	Co-training 5.5	Self-training 5.8
Iterative	Co-EM 5.1	EM 8.9

Method	Use <i>random</i> feature split?	
	Yes	No
Incremental	Co-training 28.0	Self-training 27.0
Iterative	Co-EM 29.9	EM 31.2

ML Questions

- Questions???

ML Bootstrapping

- Start with a (small) set of seed instances of the concept you want to learn
- Until stopping criteria
 - Learn something about the attribute patterns that co-occur with the instances
 - Find other instances of the concept using these patterns
- And/or start with a (small) set of attribute patterns indicative of the concept

ML Bootstrapping NE Tagging

- Named Entity (NE) tagging involves detecting references to people, locations, times, etc.
 - Bootstrapping here is similar to Co-training
 - Example:
 - View 1: Reference text of names (e.g., a gazetteer)
 - View 2: Context of NEs in text to be tagged
- Denver The conference was in Denver
- Seattle Seattle, Washington

ML Bootstrapping Integration

* The class II transactivator **CIITA** interacts with the TBP-associated factor **TAFII32**. *

ASAP1 binds to **Smad7** and becomes tyrosine phosphorylated in mammalian cells.

Interaction Pairs:

Smad7, Smurf2
TrkC, BMPRII
CIITA, TAFII32
ASAP1, Src
...

Smurf2

...
not with

an E3
GF beta

TrkC directly
or (**BMPRII**),
acting

ML Bootstrapping Integration

* During CD30 signal transduction, we found that binding of TRAF2 to the cytoplasmic domain of CD30 results in the rapid depletion of TRAF2 and the associated protein TRAF1 by proteolysis. *

...

Interaction Pairs:
Smad7, Smurf2
TrkC, BMPRII
CIITA, TAFII32
ASAP1, Src
TRAF2, CD30

interacts
to domains of

interacts

eral
32. *
proteins and
FAK family

domains of Src
phosphorylated in

ML Bootstrapping Integration

However, only the MEC1(2-2368) component established the *Mec1-Rfa1* interaction in a modified two-hybrid system.

Requirement of DDC2 overexpression for the *Mec1-Rfa1* interaction:

...

Interaction Pairs:
Smad7, Smurf2
TrkC, BMPRII
CIITA, TAFII32
ASAP1, Src
TRAF2, CD30
Mec1, Rfa1

interacts
nit of Y
to domains of Y
interaction

iated HIV
n this study
ents that
raction or
ul in the

nowled a
ture, as
interaction

ML Bootstrapping Integration

* *Egr-1* and the *serum response factor* were found to interact ...

...

Interaction Pairs:

CIITA, TAFII32

ASAP1, Src

TRAF2, CD30

Mec1, Rfa1

Egr-1, SRF

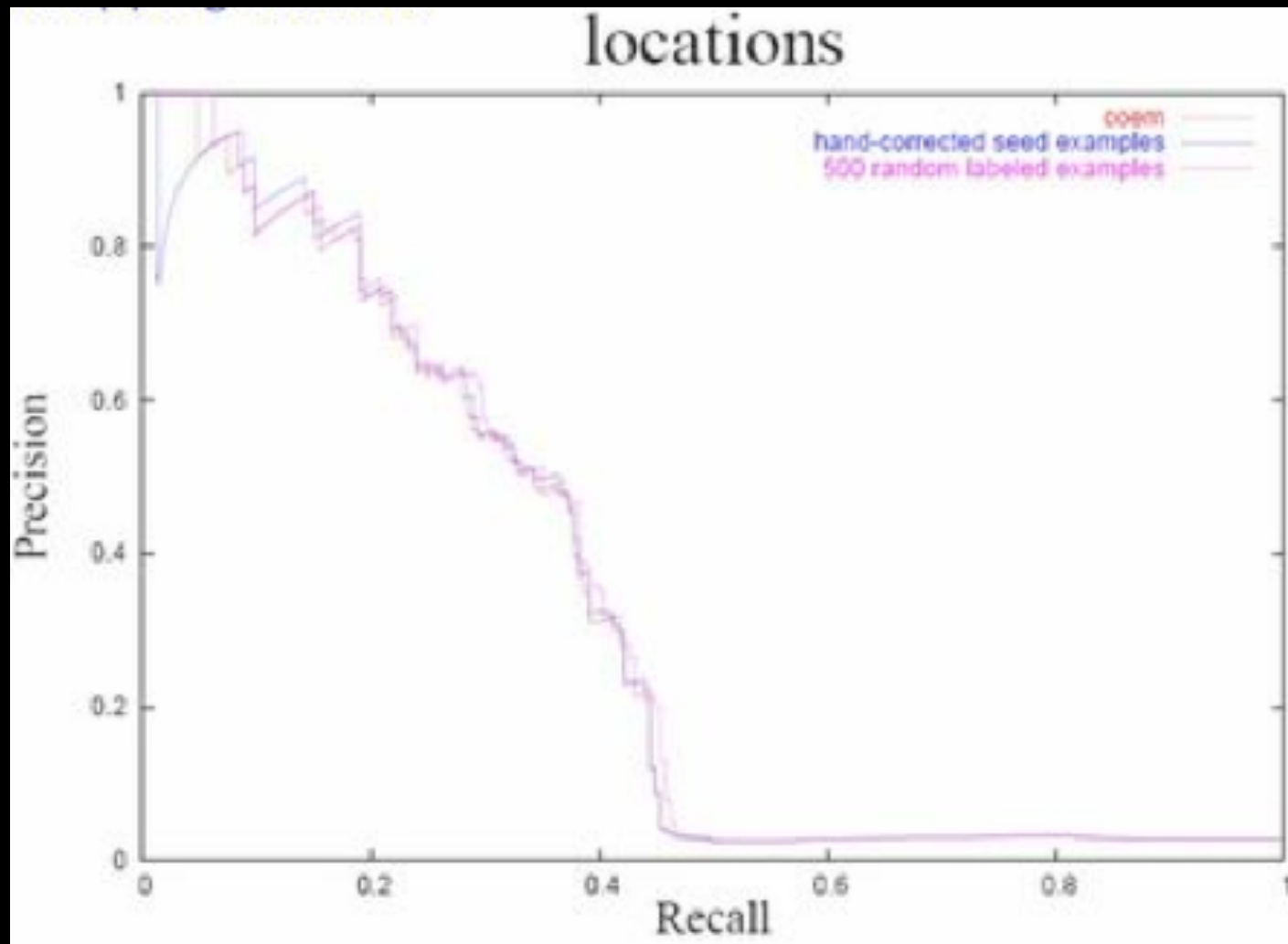
...

... and to

... interact

... in a previous study.

ML Bootstrapping: Precision Degradation



ML Bootstrapping: Precision Degradation

England

Sweden

Saudi Arabia

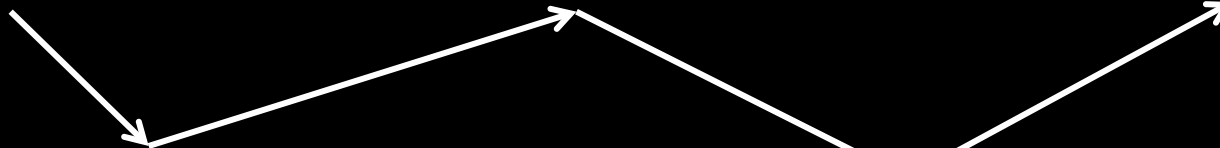
India

Colorado

Denial

You (*Lion King*)

Airport



king of X
lives in X

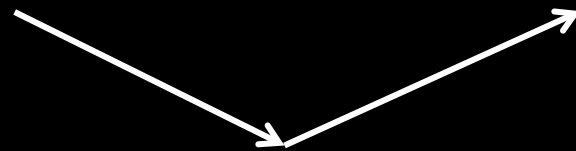
X of service attack
confront X
state of X

ML

Filtering Patterns

- Some patterns are more precise than others
- How do we identify these and how do we use them?

England
Sweden
Saudi Arabia



X's ambassador Y
the king of X, Y,

~~Turkey
Indian
Brussels
Canaries
Hungary
Israel~~

~~X manager
X truck jobs~~

ML Coupled Bootstrap Learning

- Some patterns are more precise than others
- How do we identify these and how do we use them?
 - Co-training: two separate feature sets
 - X1: word = Sweden
 - X2: {rt1 = ', rt2 = s, rt3 = ambassador}

Sweden 's ambassador

ML Coupled Bootstrap Learning

- Some patterns are more precise than others
- How do we identify these and how do we use them?
 - Concurrent multiple-class learning
 - Ontological constraints
 - Country's Diplomat
 - Country is-a Location, Diplomat is-a Person
 - Location \neq Person (Mutual Exclusion)

Sweden 's ambassador

ML Coupled Bootstrap Learning

- Some patterns are more precise than others
- How do we identify these and how do we use them?
 - Concurrent entity and relation learning
 - AmbassadorOf(X, Y)

Sweden 's ambassador to the *United States*, Jonas Hafström,

ML Coupled Bootstrap Learning

- Key to better performance:
 - More constraints: additional learning functions
- How can we couple more learners?
 - Prefer additional learners with independent errors
 - Learn from other data sources (eg, html, semantic web)
 - Learn new constraints

ML Coupled Bootstrap Learning

- Input: An ontology O , and text corpus C
- Output: Instances & patterns for each predicate
- SHARE initial instances and patterns among predicates
- forever do
 - for each predicate p in O do
 - EXTRACT candidate instances and patterns
 - FILTER candidates
 - TRAIN instance and pattern classifiers
 - ASSESS candidates using classifiers
 - PROMOTE highest-confidence candidates
 - SHARE promoted items among predicates

ML Coupled Bootstrap Learning

- Learn probabilistic Horn clauses
- 40 learned rules for teamPlaySport & playSport
- Inferred 124 new beliefs, e.g.,
 - teamPlaysSport(Caps,hockey),
 - playSport(JasonGiambi,baseball)

0.84 playSport(?x,?y) \leftarrow playsFor(?x,?z), teamPlaysSport(?z,?y)

0.70 playSport(?x,baseball) \leftarrow playsFor(?x,cubs)

...

0.81 teamPlaysSport(?x,?y) \leftarrow playsForTeam(?x,?z), playSport(?z,?y)

0.70 teamPlaysSport(?x,basketball) \leftarrow playsAgainst(?x,pistons)

0.64 teamPlaysSport(?x,?y) \leftarrow playsAgainst(?x,?z), teamPlaysSport(?z,?y)

...

ML Never-ending Language Learning

- Tom Mitchell et al.
 - Requires accurately constrained SSL
 - Constrain learning by *coupling* several SSL tasks
 - Learning should improve over time, not degrade

ML Project Applications & Coupling

- Peter B: Detecting new housing dev from satellite imagery
 - New bridges, streets, trees, vehicles, fences, sidewalks, loss of ...
- Patrick: Id instruments in a piece of music
 - Other instruments, genre, lyrics, ...
- Chenyu: Id genre of a piece of music ...
- Xinyu: Recommend movies for a group
 - Recommend actors, directors, genres, writers, ...
- Jena: Id & classify linguistic constructions
 - NEs, location types, object types, ...
- Brian: Id search spam web pages on the Internet
 - Co-training: X1=words, X2=meta info, X3=linked pp
 - Couple with learning spammers, (il)legitimate ads, ...

ML Questions

- Questions???

ML Presentations

- Wed, Nov 11: Chenyu Zheng
- Volunteers
 - Mon, Nov 16:
 - Wed, Nov 18:

ML Projects

- Office hours: will stay until all questions are addressed
- Or make an appointment