Bag of Words Classification

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries. At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.
Naïve Bayes Learner

Train:

For each class $c_j$ of documents

1. Estimate $P(c_j)$

2. For each word $w_i$ estimate $P(w_i \mid c_j)$

Classify (doc):

Assign doc to most probable class

$$\text{arg max } P(c_j) \prod_{w_i \in \text{doc}} P(w_i \mid c_j)$$

* assuming words are conditionally independent, given class
The Problem

- Want higher accuracy from fewer labeled examples

Opportunity 1:

- Use all that unlabeled data
Redundantly Sufficient Features

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto); B.Sc. (Nat. Tech. U. Ath)

Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu
Redundantly Sufficient Features

Professor Faloutsos

my advisor
Redundantly Sufficient Features

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto); B.Sc. (Nat. Tech. U. Ath)

Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu
Redundantly Sufficient Features

Professor Faloutsos

my advisor

U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science, (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto); B.Sc. (Nat. Tech. U. Ath

Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;
CoTraining Algorithm #1
[Blum&Mitchell, 1998]

Given: labeled data L, unlabeled data U

Loop:
    Train g1 (hyperlink classifier) using L
    Train g2 (page classifier) using L
    Allow g1 to label $p$ positive, $n$ negative examps from U
    Allow g2 to label $p$ positive, $n$ negative examps from U
    Add these self-labeled examples to L
CoTraining: Experimental Results

- begin with 12 labeled web pages (academic course)
- provide 1,000 additional unlabeled web pages
- average error: using labeled data only 11.1%;
- average error: cotraining 5.0%

Typical run:
Example 2: Learning semantic lexicons

[Riloff and Jones, 1999]

Learn which noun phrases represent locations:

“We have operations in bustling Kuwait City”

- Classifier for x1:
  - “operations in … NP”

- Classifier for x2:
  - rote dictionary: New York, Paris, Germany, ...
Learning semantic lexicons
[Riloff and Jones, 1999]

“We are headquartered in Pittsburgh.”
“We are headquartered in sunny Tehran.”
“Our offices are located in downtown Tehran.”

10 seed words
Example: Learning Locations

• 10 seed words:
  – United_States Germany England Switzerland France Canada Mexico Japan China Australia

• Top words added by bootstrapping:
  – Europe Greece Italy Singapore Finland UK North_America States de_Benelux Deutschland de_Benelux_seminars Asia/Pacific Middle_East/Africa U.S. Hong_Kong Spain Portugal World Philippines Countries Oregon...
Example: Learning Locations

- Top rules learned by bootstrapping:
  - offices in ?x
  - facilities in ?x
  - operations in ?x
  - loans in ?x
  - operates in ?x
  - locations in ?x
  - producer in ?x
CoTraining Setting

\[ f : X \rightarrow Y \]

where \( X = X_1 \times X_2 \)

where \( x \) drawn from unknown distribution

and \( \exists g_1, g_2 \quad (\forall x) g_1(x_1) = g_2(x_2) = f(x) \)

- If
  - \( x_1, x_2 \) conditionally independent given \( y \)
  - \( f \) is PAC learnable from noisy labeled data
- Then
  - \( f \) is PAC learnable from weak initial classifier plus unlabeled data
Co-Training Rote Learner
Co-Training Rote Learner

My advisor

hyperlinks

pages
Co-Training Rote Learner
Co-Training Rote Learner
Co-Training Rote Learner
Rote CoTraining error given $m$ examples

**CoTraining setting:**

learn $f : X \rightarrow Y$

where $X = X_1 \times X_2$

where $x$ drawn from unknown distribution

and $\exists g_1, g_2 \ (\forall x) g_1(x_1) = g_2(x_2) = f(x)$

$$E[error] \leq \sum_j p_j (1 - p_j)^m$$

Where $p_j$ is probability that a randomly drawn example will fall into the $j$th connected component of the graph of $U+L$
How many *unlabeled* examples suffice?

Want to assure that connected components in the underlying distribution, $G_D$, are connected components in the observed sample, $G_S$

$O(\log(N)/\alpha)$ examples assure that with high probability, $G_S$ has same connected components as $G_D$ [Karger, 94]

$N$ is size of $G_D$, $\alpha$ is min cut over all connected components of $G_D$
What if CoTraining Assumption Not Perfectly Satisfied?
What if CoTraining Assumption Not Perfectly Satisfied?
What if CoTraining Assumption Not Perfectly Satisfied?

- Idea: Want classifiers that produce a \textit{maximally consistent} labeling of the data
- If learning is an optimization problem, what function should we optimize?
What Objective Function?

\[ E = E1 + E2 \]

\[ E1 = \sum_{<x,y> \in L} (y - \hat{g}_1(x_1))^2 \]

\[ E2 = \sum_{<x,y> \in L} (y - \hat{g}_2(x_2))^2 \]
What Objective Function?

$$E = E_1 + E_2 + c_3 E_3$$

$$E_1 = \sum_{<x,y> \in L} (y - \hat{g}_1(x_1))^2$$

$$E_2 = \sum_{<x,y> \in L} (y - \hat{g}_2(x_2))^2$$

$$E_3 = \sum_{x \in U} (\hat{g}_1(x_1) - \hat{g}_2(x_2))^2$$

- Error on labeled examples
- Disagreement over unlabeled
What Objective Function?

\[ E = E_1 + E_2 + c_3 E_3 + c_4 E_4 \]

\[ E_1 = \sum_{(x, y) \in L} (y - \hat{g}_1(x))^2 \]

\[ E_2 = \sum_{(x, y) \in L} (y - \hat{g}_2(x))^2 \]

\[ E_3 = \sum_{x \in U} (\hat{g}_1(x) - \hat{g}_2(x))^2 \]

\[ E_4 = \left( \frac{1}{|L|} \sum_{(x, y) \in L} y - \frac{1}{|L| + |U|} \sum_{x \in L \cup U} \frac{\hat{g}_1(x) + \hat{g}_2(x)}{2} \right)^2 \]

- Error on labeled examples
- Disagreement over unlabeled
- Misfit to estimated class priors
What Function Approximators?
What Function Approximators?

\[
\hat{g}_1(x) = \frac{1}{1 + e^{-\sum w_{j,1} x_j}}
\]

\[
\hat{g}_2(x) = \frac{1}{1 + e^{-\sum w_{j,2} x_j}}
\]

- Move away from rote learning
- Same fn form as Naïve Bayes, Max Entropy
- Use gradient descent to simultaneously learn \(g_1\) and \(g_2\), directly minimizing 
  \[E = E_1 + E_2 + E_3 + E_4\]
- No word independence assumption
Gradient CoTraining

\[ \hat{g}_1(x) = \frac{1}{1 + e^{\sum_j w_{j,1} x_j}} \]

\[ \hat{g}_2(x) = \frac{1}{1 + e^{\sum_j w_{j,2} x_j}} \]

Gradient

\[ \nabla E[\vec{w}] = \begin{bmatrix} \frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n} \end{bmatrix} \]

Training rule:

\[ \Delta \vec{w} = -\eta \nabla E[\vec{w}] \]
Classifying Jobs for FlipDog

X1: job title
X2: job description

C++/Java Consultants at Elite Placement Services
Job Number: C1 Salary Range: $80K Job Description: Functions of this position include the consulting, development and implementation of EAI solutions supporting e-commerce and B2B initiatives for...

Chief Software Architect at Elite Placement Services
Job Number: CSA1 Salary Range: to $150K Job Description: Responsible for the end-to-end architecture of all n-tiered web-based applications and complementary products. Provide design direction for the...

Web Application Developers at MI Systems, Inc.
Location: Houston, TX Last Updated: 10/04/00 Job Type: Full-Time Contract Length: 0 Salary: open Hourly Pay: See on Synopsis: Permanent Opportunities (2) Application Developers with...

Sales Consulting Engineer at Visual Numerics, Inc.
Job Code 00-022-H Back to Top WHAT'S THE JOB? Performs pre-sales technical support for products to customers and non-customers. Technical support includes providing verbal and written response...

Peoplesoft Software Analyst (Systems Analyst III) at I.T. Staffing, Inc.
Date Posted: 10/12/00 Location: Houston, TX (Some international travel required) Job Description: CLIENT/SERVER APPLICATION ADMINISTRATION. SETTING UP USERS AND SECURITY FOR DATABASE AND APPLICATION....

Peoplesoft Software Analyst (Systems Analyst III) at I.T. Staffing, Inc.
Date Posted: 10/12/00 Location: Houston, TX (Some international travel required) Job Description: CLIENT/SERVER APPLICATION ADMINISTRATION. SETTING UP USERS AND SECURITY FOR DATABASE AND APPLICATION....
Gradient CoTraining

Classifying FlipDog job descriptions: SysAdmin vs. WebProgrammer

Final Accuracy
Labeled data alone: 86%
CoTraining: 96%
Gradient CoTraining
Classifying Upper Case sequences as Person Names

<table>
<thead>
<tr>
<th>Using labeled data only</th>
<th>25 labeled</th>
<th>5000 unlabeled</th>
<th>2300 labeled</th>
<th>5000 unlabeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cotraining</td>
<td>.76</td>
<td>.87</td>
<td>.85</td>
<td>.89</td>
</tr>
<tr>
<td>Cotraining without fitting class priors (E4)</td>
<td>.73</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

* sensitive to weights of error terms E3 and E4
Potential CoTraining Domains

- Web page classification [Blum, Mitchell 98]
- Semantic lexicon generation [Riloff, Jones 99], [Collins, Singer 99]
- Word sense disambiguation [Yarowsky 95]
- Speech recognition [de Sa, Ballard 98]
- Multimedia classification ??
- Robotic perception ??
- Models of human learning ??