Active Learning

Burr Settles

Machine Learning 10-701 / 15-781
Nov 14, 2012
Let’s Play 20 Questions!

- I’m thinking of something; ask me yes/no questions to figure out what it is...
How Do We *Automate* Inquiry?

A Though Experiment
A Thought Experiment

• suppose you are on an Earth convoy sent to colonize planet Zelgon

people who ate the round Zelgian fruits found them *tasty*!

people who ate the rough Zelgian fruits found them *gross*!
Poisonous vs. Yummy Alien Fruits

• there is a continuous range of round-to-rough fruit shapes on Zelgon:

![fruit shapes]

you need to learn how to classify fruits as **safe** or **noxious**

and you need to do this while risking as little as possible (i.e., colonist health)
Supervised Learning Approach

Problem:

PAC theory tells us we need $O(1/\varepsilon)$ tests to obtain an error rate of $\varepsilon$...

...a lot of people might get sick in the process!
Can We Do Better?

this is just a **binary search**...

requiring $O(1/\varepsilon)$ fruits (e.g., samples)
but only $O(\log_2 1/\varepsilon)$ tests (e.g., queries)

our first “active learning” algorithm!
Supervised Learning

raw unlabeled data
\( x_1, x_2, x_3, \ldots \)

labeled training instances
\( \langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \langle x_3, y_3 \rangle, \ldots \)

expert / oracle
analyzes experiments to determine labels

random sample
Active Learning

- **Active learner** induces a classifier
- **Expert/oracle** analyzes experiments to determine labels

- Inspect the unlabeled data
- Request labels for selected data

Raw unlabeled data: \( x_1, x_2, x_3, \ldots \)
Learning Curves

text classification: baseball vs. hockey

active learning

passive learning

cost (e.g., number of instance queries)
Who Uses Active Learning?

Sentiment analysis for blogs; Noisy relabeling
– Prem Melville

Biomedical NLP & IR; Computer-aided diagnosis
– Balaji Krishnapuram

MS Outlook voicemail plug-in [Kapoor et al., IJCAI'07];
“A variety of prototypes that are in use throughout the company.” – Eric Horvitz

“While I can confirm that we're using active learning in earnest on many problem areas… I really can't provide any more details than that. Sorry to be so opaque!”
– David Cohn
Active Learning Scenarios
Problems with Query Synthesis

an early real-world application: neural-net queries synthesized for handwritten digits
[Lang & Baum, 1992]

problem: humans couldn’t interpret the queries!

ideally, we can ensure that the queries come from the underlying “natural” distribution
Active Learning Scenarios

- **Membership query synthesis**: model generates a query de novo.

  - **Stream-based selective sampling**: sample an instance, model decides to query or discard.
    - More common in *theory* papers.

  - **Pool-based active learning**: sample a large pool of instances, model selects the best query.
    - More common in *application* papers.

  - **Instance space or input distribution**

  - **Query is labeled by the oracle**
Active Learning Approaches
(1) Uncertainty Sampling
Zelgian Fruits Revisited

- let’s interpret our Zelgian fruit binary search in terms of a *probabilistic* classifier:

\[ P(Y = \text{😊} | X) \]
Uncertainty Sampling

- query instances the learner is *most uncertain* about

400 instances sampled from 2 class Gaussians

random sampling
30 labeled instances
(accuracy=0.7)

uncertainty sampling
30 labeled instances
(accuracy=0.9)

[Lewis & Gale, SIGIR’94]
Common Uncertainty Measures

least confident

$$\phi_{LC}(x) = 1 - P_\theta(y^*|x)$$

margin

$$\phi_M(x) = P_\theta(y_1^*|x) - P_\theta(y_2^*|x)$$

total entropy

$$\phi_{ENT}(x) = -\sum_{y} P_\theta(y|x) \log_2 P_\theta(y|x)$$
Common Uncertainty Measures

\[ (a) \text{ least confident – binary} \quad (b) \text{ margin – binary} \quad (c) \text{ entropy – binary} \]

\textit{note:} for binary tasks, these are functionally equivalent!
Common Uncertainty Measures

Illustration of preferred (dark red) posterior distributions in a 3-label classification task

**Note:** for multi-class tasks, these are *not* equivalent!
Information-Theoretic Interpretation

• the “surprisal” $\mathcal{I}$ is a measure (in bits, nats, etc.) of the information content for outcome $y$ of variable $Y$:

$$
\mathcal{I}(y) = \log \frac{1}{P(y)} = -\log P(y)
$$

• so this is how “informative” the oracle’s label $y$ will be

• but the learner doesn’t know the oracle’s answer yet! we can estimate it as an *expectation* over all possible labels:

$$
E_y \left[ -\log P_\theta(y|x) \right] = -\sum_y P_\theta(y|x) \log P_\theta(y|x)
$$

• which is *entropy*-based uncertainty sampling
Uncertainty Sampling in Practice

• pool-based active learning:
  – evaluate each $x$ in $U$
  – rank and query the top $K$ instances
  – retrain, repeat

• selective sampling:
  – threshold a “region of uncertainty,” e.g., [0.2, 0.8]
  – observe new instances, but only query those that fall within the region
  – retrain, repeat
Uncertainty Sampling: Example

target function

neural net trained from 100 random pixels

active neural net (stream-based uncertainty sampling)
Simple and Widely-Used

- **text classification**
  - Lewis & Gale ICML’94;

- **POS tagging**
  - Dagan & Engelson, ICML’95; Ringger et al., ACL’07

- **disambiguation**
  - Fujii et al., CL’98;

- **parsing**
  - Hwa, CL’04

- **information extraction**
  - Scheffer et al., CAIDA’01; Settles & Craven, EMNLP’08

- **word segmentation**
  - Sassano, ACL’02

- **speech recognition**
  - Tur et al., SC’05

- **transliteration**
  - Kuo et al., ACL’06

- **translation**
  - Haffari et al., NAACL’09
Uncertainty Sampling: Failure?!
What To Do?

• uncertainty sampling only uses the confidence of *one single* classifier
  – e.g., a “point estimate” for parametric models
  – this classifier can become overly confident about instances is really knows nothing about!

• instead, let’s consider a different notion of “uncertainty”... about the *classifier itself*
Active Learning Approaches

(2) Hypothesis Space Search
Remember Version Spaces?

- the set of all classifiers that are consistent with the labeled training data
- the larger the version space $\mathcal{V}$, the less likely each possible classifier is... we want queries to reduce $|\mathcal{V}|$
Alien Fruits Revisited

- let’s try interpreting our binary search in terms of a version space search:

possible classifiers (thresholds): 1
Version Space Search

• in general, the version space $\mathcal{V}$ may be too large to enumerate, or to measure the size $|\mathcal{V}|$ through analytical trickery

• observation: for the Zelgian fruits example, uncertainty sampling and version space search gave us the same queries!

• how far can uncertainty sampling get us?
Version Spaces for SVMs

\( \mathcal{F} \) (feature space)

"version space duality" (Vapnik, 1998)

points in \( \mathcal{F} \) correspond to hyperplanes in \( \mathcal{H} \) and vice versa

\( \mathcal{H} \) (hypothesis space)

SVM with largest margin is the center of the largest hypersphere in \( \mathcal{V} \)
Bisecting the SVM Version Space

- hence, uncertainty sampling is a special case of version space search for SVMs (and other so-called “max-margin” classifiers)
Query By Disagreement (QBD)

- in general, uncertainty doesn’t cut it

- **idea**: we wish to quickly *eliminate bad hypotheses*; train two classifiers $G$ and $S$ which represent the two “extremes” of the version space

- if these two models disagree, the instances falls within the “region of uncertainty”

[Cohn et al., ML'94]
Neural Network Triangles Revisited

target function

QBD:

uncertainty sampling:
Query By Committee (QBC)

• simpler, more general approach

• train a committee of classifiers $\mathcal{C}$
  – no need to maintain $G$ and $S$
  – committee can be any size

• query instances for which committee members disagree

[Seung et al., COLT'92]
QBC in Practice

• selective sampling:
  – train a committee $C$
  – observe new instances, but only query those for which there is disagreement (or a lot of disagreement)
  – retrain, repeat

• pool-based active learning:
  – train a committee $C$
  – measure disagreement for each $x$ in $\mathcal{U}$
  – rank and query the top $K$ instances
  – retrain, repeat
QBC Design Decisions

• how to build a committee:
  – “sample” models from $P(\theta|\mathcal{L})$
    • [Dagan & Engelson, ICML’95; McCallum & Nigam, ICML’98]
  – standard ensembles (e.g., boosting, bagging)
    • [Abe & Mamitsuka, ICML’98]

• how to measure disagreement (many):
  – “XOR” committee classifications
  – view vote distributions as probabilities, use uncertainty measures...
QBC Disagreement Measures

• “soft” vote entropy:

\[ x^*_{SVE} = \arg\max_x - \sum_y P_c(y|x) \log P_c(y|x) \]

• average Kullback-Liebler (KL) divergence:

\[ x^*_{KL} = \arg\max_x \frac{1}{|C|} \sum_{\theta \in C} KL( P_\theta(Y|x) \parallel P_c(Y|x) ) \]
QBC Disagreement Measures

heatmaps illustrating query heuristics for a 3-label classification task using multinomial logistic regression (e.g., a MaxEnt model)
QBC Disagreement Measures

\[ P_{\theta(1)} \quad P_{\theta(2)} \quad P_{\theta(3)} \]

uncertain hypotheses; but in agreement

\[ P_C \]

SVE cannot tell either of these apart

\[ P_{\theta(1)} \quad P_{\theta(2)} \quad P_{\theta(3)} \]

confident hypotheses; but in \textit{dis}agreement

\[ P_C \]

KL divergence will query this
Information-Theoretic Interpretation

- we want to query the instance whose label contains maximal mutual information about the version space: \(I(Y; \mathcal{V})\)

- consider the identity:

\[
I(Y; \mathcal{V}) = H(\mathcal{V}) - H(\mathcal{V}|Y)
\]

\[
= H(\mathcal{V}) - \mathbb{E}_Y [H(\mathcal{V}|y)]
\]

- this justifies querying instances which will reduce \(|\mathcal{V}| \approx H(\mathcal{V})\) in expectation
Information-Theoretic Interpretation

• an alternate, equivalent identity:

\[
I(Y; \mathcal{V}) = KL\left( P(Y, \mathcal{V}) \parallel P(Y)P(\mathcal{V}) \right)
\]

\[
= \mathbb{E}_{\theta \in \mathcal{V}} \left[ KL\left( P_{\theta}(Y) \parallel P(Y) \right) \right]
\]

• which, under a few simple assumptions, reduces to the KL-divergence heuristic for QBC
imagine Zelgon has both grey and red fruits, with different thresholds?

there are two queries A and B both bisect \( \mathcal{V} \)

which query will result in the lowest classification error?
Active Learning Approaches

(3) Using the Data Distribution
Expected Error Reduction

- minimize the expected 1/0 loss of a query $x$

$$x^*_{ER} = \arg\min_x \mathbb{E}_{Y|\theta,x} \left[ \sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta^+,x'} [y \neq \hat{y}] \right]$$

$$= \arg\min_x \sum_y P_\theta(y|x) \left[ \sum_{x' \in \mathcal{U}} 1 - p_{\theta^+}(\hat{y}|x') \right]$$

- expectation over possible labelings of $x$
- sum over unlabeled instances
- 0/1 error of $x'$ after retraining with $x$
Expected Error Reduction

- minimize the expected log loss of a query $x$

\[
x_{LL}^* = \underset{x}{\arg\min} \sum_{y} P_\theta(y|x) \left[ \sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|x',x'} \left[ -\log p_{\theta+}(y|x') \right] \right]
\]

\[
= \underset{x}{\arg\min} \sum_{y} P_\theta(y|x) \left[ \sum_{x' \in \mathcal{U}} \sum_{y'} - p_{\theta+}(y'|x') \log p_{\theta+}(y'|x') \right]
\]

\[
= \underset{x}{\arg\min} \sum_{y} P_\theta(y|x) \sum_{x' \in \mathcal{U}} H_{\theta+}(Y|x')
\]

- expectation over labelings of $x$
- sum over unlabeled instances
- entropy of $x'$ after retraining with $x$
Text Classification Examples

comp.graphics vs. comp.windows.x
Text Classification Examples

comp.sys.ibm.pc.hardware vs. comp.os.ms-windows.misc

[Roy & McCallum, ICML’01]
Information-Theoretic Interpretation

- aim to maximize the *information gain* over $\mathcal{U}$

\[
x^* = \arg\max_x \sum_{x' \in \mathcal{U}} \left( H_\theta(Y|x') - \mathbb{E}_{Y|\theta,x} [H_\theta+(Y|x')] \right)
\]

\[
= \arg\max_x \sum_{x' \in \mathcal{U}} H_\theta(Y|x') - \sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta,x} [H_\theta+(Y|x')]
\]

\[
= \arg\min_x \sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta,x} [H_\theta+(Y|x')].
\]
Poor Scalability

• expected error reduction tries to directly optimize the loss of interest, but…

• quickly becomes intractible
  – logistic regression requires $O(ULG)$ time
  – MaxEnt would require $O(M^2ULG)$ time
Approximation: Density-Weighting

- assume that the information gained per unlabeled instance \( x' \) is proportional to its similarity to the query \( x \):

\[
x^* = \arg \max_x \sum_{x' \in \mathcal{U}} \left( H_\theta(Y|x') - \mathbb{E}_{Y|x',x}[H_{\theta+}(Y|x')] \right)
\]

\[
\approx \arg \max_x \sum_{x' \in \mathcal{U}} \left( \text{sim}(x, x') \times H_\theta(Y|x) \right).
\]
Active Learning++
Beyond Instance Queries
Beyond Instance Queries

• most research in active learning has been based on a few simple assumptions:
  – “cost” is proportional to training set size
  – queries must be unlabeled instances
  – there is only a single classifier to train
1. Real Annotation Costs

empirical study of *time* as labeling cost for four data sets:

- **CKB**
  - \( \mu = 492.2 \)
  - \( \sigma = 593.5 \)

- **SIVAL**
  - \( \mu = 31.9 \)
  - \( \sigma = 17.3 \)

- **Spec**
  - \( \mu = 7.6 \)
  - \( \sigma = 6.1 \)

- **SigLE**
  - \( \mu = 27.0 \)
  - \( \sigma = 14.7 \)

[Results supported by Aurora et al., ALNLP’09; Vijayanarasimhan & Grauman, CVPR’09]

[Settles et al., 2008]
Strategies for Variable Annotation Costs

• use the current trained model assist with automatic pre-annotation
  – some successes [Baldridge & Osbourne ’04; Culotta & McCallum ’05; Baldridge & Palmer ’09; Felt et al. ’12]

• train a regression cost model in parallel (i.e., to predict time or $$) and incorporate that into the query selection heuristic
  – mixed results [Settles et al. ’08; Haertel et al. ’08; Tomanek and Hahn ’10]
2. New Query Types

• in many NLP applications, “features” are discrete variables with semantic meaning:
  – words
  – affixes
  – capitalization
  – other orthographic patterns

• what if active learning systems could ask about “feature labels,” too?

[Druck et al., EMNLP’09; Settles, EMNLP’11]
DUALIST

settles, EMNLP’11
Results: Movie Reviews

![Graphs showing accuracy over time for DUALIST active and passive methods.](image)
Results: WebKB

![Graphs showing accuracy and time for DUALIST active and passive documents over time](image-url)
Results: Science

Accuracy over time (seconds) for DUALIST active and passive documents.
3. Multi-Task, Multi-View Active Learning

• CMU’s NELL (Never Ending Language Learner)

• **given:** an ontology (schema), access to the Web, and a few seed examples per predicate, and periodic access to humans

• **task:** run 24x7 each day, populating a knowledge base with new facts
  – learning to read and reading to learn ...

[Carlson et al., AAAI’10]
NELL’s Architecture

- multiple tasks/views constrain each other, helping to prevent concept drift (“checks and balances”)
- to date: >1.5 million beliefs at 80% precision
One View: CPL
(contextual patterns)

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>emotion</td>
<td>hearts full of $X$</td>
</tr>
<tr>
<td>beverage</td>
<td>cup of aromatic $X$</td>
</tr>
<tr>
<td>newspaper</td>
<td>op-ed page of $X$</td>
</tr>
<tr>
<td>teamPlaysInLeague</td>
<td>$X$ ranks second in $Y$</td>
</tr>
<tr>
<td>bookAuthor</td>
<td>$Y$ classic $X$</td>
</tr>
</tbody>
</table>
### Another View: CMC
(orthographic features)

<table>
<thead>
<tr>
<th>Predicate</th>
<th>Feature</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>mountain</td>
<td>LAST=peak</td>
<td>1.791</td>
</tr>
<tr>
<td>mountain</td>
<td>LAST=mountain</td>
<td>1.093</td>
</tr>
<tr>
<td>mountain</td>
<td>FIRST=mountain</td>
<td>-0.875</td>
</tr>
<tr>
<td>musicArtist</td>
<td>LAST=band</td>
<td>1.853</td>
</tr>
<tr>
<td>musicArtist</td>
<td>POS=DT_NNS</td>
<td>1.412</td>
</tr>
<tr>
<td>musicArtist</td>
<td>POS=DT_JJ_NN</td>
<td>-0.807</td>
</tr>
<tr>
<td>newspaper</td>
<td>LAST=sun</td>
<td>1.330</td>
</tr>
<tr>
<td>newspaper</td>
<td>LAST=university</td>
<td>-0.318</td>
</tr>
<tr>
<td>newspaper</td>
<td>POS=NN_NNS</td>
<td>-0.798</td>
</tr>
<tr>
<td>university</td>
<td>LAST=college</td>
<td>2.076</td>
</tr>
<tr>
<td>university</td>
<td>PREFIX=uc</td>
<td>1.999</td>
</tr>
<tr>
<td>university</td>
<td>LAST=state</td>
<td>1.992</td>
</tr>
<tr>
<td>university</td>
<td>LAST=university</td>
<td>1.745</td>
</tr>
<tr>
<td>university</td>
<td>FIRST=college</td>
<td>-1.381</td>
</tr>
<tr>
<td>visualArtMovement</td>
<td>SUFFIX=ism</td>
<td>1.282</td>
</tr>
<tr>
<td>visualArtMovement</td>
<td>PREFIX=journ</td>
<td>-0.234</td>
</tr>
<tr>
<td>visualArtMovement</td>
<td>PREFIX=budd</td>
<td>-0.253</td>
</tr>
</tbody>
</table>
Gender Issues

@cmunell
NELL

I think "sarah palin" is a #Male
(http://bit.ly/dz11Wc)

3 Nov via NELLbot

Retweeted by _stephie_c and 71 others
I proudly voted for _
_ is still the governor
_ is the Republican nominee
_ signed the legislation
_ signed this bill

impeachment proceedings of _
_ 's inaugural
_ signs bill
endorsed _
vice presidential candidates like _

• these CPL patterns are generally correlated with *males* across the Web

• even though CMC learned that “Sarah” is a *female* name, these patterns initially overwhelmed all other evidence, and NELL predicted *male*

• these days, NELL uses multi-task/view active learning algorithms to identify beliefs with “conflicting” evidence, and query them
Interesting Open Issues

- better cost-sensitive approaches
- “crowdsourced” labels (noisy oracles)
- batch active learning (many queries at once)
- HCI / user interface issues
- data reusability
For Further Reading...

new book published by Morgan & Claypool

free to download from the CMU campus network

[active-learning.net]