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:

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  \langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \langle x_3, y_3 \rangle, \ldots
```

random sample

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

**supervised learner** induces a classifier

**expert / oracle** analyzes experiments to determine labels
Active Learning

Active learner induces a classifier

Inspect the unlabeled data

Raw unlabeled data:
\[ x_1, x_2, x_3, \ldots \]

Request labels for selected data:
\[ \langle x_1, ? \rangle \]
\[ \langle x_2, ? \rangle \]
\[ \langle x_1, y_1 \rangle \]
\[ \langle x_2, y_2 \rangle \]

Expert / oracle analyzes experiments to determine labels

Learning Curves

Text classification: baseball vs. hockey

Active learning
Passive learning

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

IBM

Sentiment analysis for blogs; Noisy relabeling
– Prem Melville

SIEMENS

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

Microsoft

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

Google

“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

more common in theory papers

more common in application papers
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)

Common Uncertainty Measures

**least confident**

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

**margin**

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

**entropy**

\[ \phi_{ENT}(x) = - \sum_y P_\theta(y|x) \log_2 P_\theta(y|x) \]
Common Uncertainty Measures

Note: for binary tasks, these are functionally equivalent!

Note: for multi-class tasks, these are not equivalent!
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?!

![target function](image1)
![initial sample](image2)

active neural net (stream-based uncertainty sampling)
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 $V$, the less likely each possible classifier is... we want queries to reduce $|V|$

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 $V$ may be too large to enumerate, or to measure the size $|V|$ 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?

[Seung et al., COLT’ 92]

Query By Committee (QBC)

• simpler, more general approach

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

• query instances for which committee members disagree
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 $U$
  – rank and query the top $K$ instances
  – retrain, repeat

QBC Design Decisions

• how to build a committee:
  – “sample” models from $P(q|L)$
    • [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:

\[ x_{KL}^* = \arg\max_x \frac{1}{|C|} \sum_{\theta \in C} KL( P_{\theta}(Y|x) \parallel P_C(Y|x) ) \]

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

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

*uncertain hypotheses; but in agreement*

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

*confident hypotheses; but in disagreement*

SVE cannot tell either of these apart

KL divergence will query this

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:

[Settles et al., 2008]

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

DUALIST & [Settles, EMNLP’ 11]
Interesting Open Issues

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