Theory, Methods and Applications of Active Learning
Learning to Discover

**Sequential approach:** select new samples/experiments that are predicted to be maximally informative in discriminating hypotheses.
Laplace

Decided to make new astronomical measurements when “the discrepancy between prediction and observation [was] large enough to give a high probability that there is something new to be found.”

Jaynes (1986)
What is Active Learning?

Examples come in pairs, a feature and a label, denoted \((x, y)\).

Select unlabeled examples \((x, ?)\) for labeling if the predicted label \(\hat{y}\) is highly uncertain. These examples may be especially informative.

In contrast, passive learning is based on labeling random examples.
Does active learning always help?

Two problems:

1. active learning is **greedy** and usually **myopic**, and therefore can converge to a suboptimal hypothesis

2. uncertainty sampling is ‘**noise-seeking**’, and thus may dwell unnecessarily long on highly noisy cases

Maybe the chance of heart disease is 50/50 in this region.
Why Active Learning? Understanding the Mind
Abstract—The present article is an attempt to give—in the frame of the theory of the organism-environment system (Jarvillehto, 1998a)—a new interpretation to the role of efferent influences on receptor activity and to the functions of senses in the formation of knowledge. It is argued, on the basis of experimental evidence and theoretical considerations, that the senses are not transmitters of environmental information, but create a direct connection between the organism and the environment, which makes the development of a dynamic living system, the organism-environment system, possible. In this connection process, the efferent influences on receptor activity are of particular significance because, with their help, the receptors may be adjusted in relation to the parts of the environment that are most important in achieving behavioral results. Perception is the process of joining of new parts of the environment to the organism-environment system; thus, the formation of knowledge by perception is based on reorganization (widening and differentiation) of the organism-environment system, and not on transmission of information from the environment. With the help of the efferent influences on receptors, each organism creates its own peculiar world that is simultaneously subjective and objective. The present considerations have far-reaching influences as well on experimental work in neurophysiology and psychology of perception as on philosophical considerations of knowledge formation.
Visual Perception

Attentional mechanisms probably limit our capacity to about 44 bits per-glimpse (Verghese and Pelli (1992))

So how to we perceive ‘reality’ from so few bits of information?

Churchland, Ramachandran, & Sejnowksi ’94: “Interactive vision is exploratory and predictive. Visual learning allows an animal to predict what will happen in the future; behavior, such as eye movements, aids in updating and upgrading the predictive representations.”

Clark 2002: “we establish the required visual contact with our world by an ongoing process of active exploration, in which the world acts as a kind of reliable, interrogable, external memory”
Seven records of eye movements by the same subject. Each record lasted 3 minutes. 1) Free examination. Before subsequent recordings, the subject was asked to: 2) estimate the material circumstances of the family; 3) give the ages of the people; 4) surmise what the family had been doing before the arrival of the "unexpected visitor;" 5) remember the clothes worn by the people; 6) remember the position of the people and objects in the room; 7) estimate how long the "unexpected visitor" had been away from the family (from Yarbus 1967).
Why Active Learning? Understanding Complex Systems
National Ecological Observation Network (NEON)
Where, When and How to collect information?

Clustering and Topology Discovery

Network Monitoring

Learning by Queries

wireless sensor networks/mobile sensing

remote sensing
Why Active Learning? Automating Science

Background Knowledge

Hypothesis

Analysis

Experiment Selection

Scientist

Experiment Outcome

Hypothesis

There is a need for “autonomous experimentation”

“Towards 2020 Science” – 40 eminent scientists’ visions of the future of science
Machine Learning (Passive)

Raw unlabeled data

$X_1, X_2, X_3, \ldots$

Labeled data

$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots$

passive learner

automatic classifier

expert/oracle analyzes/experiments to determine labels
Active Learning

Raw unlabeled data

Learner requests labels for selected data

(X₁, Y₁)
(X₃, Y₃)
(X₅, Y₅)

expert/oracle analyzes/experiments to determine labels

X₁, X₂, X₃, ...
For the first time, a robotic system has made a novel scientific discovery with virtually no human intellectual input.

Scientists designed "Adam" to carry out the entire scientific process on its own: formulating hypotheses, designing and running experiments, analyzing data, and deciding which experiments to run next. "It’s a major advance," says David Waltz of the Center for Computational Learning Systems at Columbia University. "Science is being done here in a way that incorporates artificial intelligence. It’s automating a part of the scientific process that hasn’t been automated in the past."

Adam is the first automated system to complete the cycle from hypothesis, to experiment, to reformulated hypothesis without human intervention.
Hypothesis and Query/Feature Spaces

\[ \mathcal{H} = \text{space of hypotheses or models} \]
\[ \mathcal{X} = \text{space of queries or unlabeled features} \]

\( h^* \) is the true model (might not belong to \( \mathcal{H} \)).

**Noiseless Learning**: \( x \in \mathcal{X} \rightarrow y = h^*(x) \)

**Noisy Learning**: \( x \in \mathcal{X} \rightarrow y = h^*(x) + \text{noise} \)

**Active Learning**: Sequentially select most informative queries/examples based on past queries/examples and responses.
A Simple Algorithm for Noiseless Active Learning

Cohn, Atlas and Ladner '92  \( h : \mathcal{X} \rightarrow \{-1, +1\}, \ h^* \in \mathcal{H} \)

initialize: \( i = 1, \mathcal{H}_1 = \mathcal{H} \)

while \( |\mathcal{H}_i| > 1 \)

1. Select \( x_i \in \{\text{any } x \in \mathcal{X} \text{ where } h \in \mathcal{H}_i \text{ disagree}\} \)

2. Query with \( x_i \) to obtain \( y_i = h^*(x_i) \)

3. Set \( \mathcal{H}_{i+1} = \{h \in \mathcal{H}_i : h(x_i) = y_i\}, \ i = i + 1 \)

Version Space

Region of Disagreement

CAL algorithm may also be operated in an online fashion
Generalized Binary Search / Splitting Algorithm
(circa 1970)

initialize: $i = 1, \mathcal{H}_1 = \mathcal{H}$

while $|\mathcal{H}_i| > 1$

1. Select $x_i = \arg\min_{x \in \mathcal{X}} |\sum_{h \in \mathcal{H}_i} h(x)|$

2. Selects a query for which disagreement among viable hypotheses is maximal

3. Set $\mathcal{H}_{i+1} = \{h \in \mathcal{H}_i : h(x_i) = y_i\}, i = i + 1$
Flavors of Active Learning and Analysis

Query synthesis: construct desired queries/questions

Pool-based: select unlabeled examples from a large collection of data

Stream-based: monitor a stream of unlabeled examples for uncertain cases

How many queries or labeled examples are required?

Extended Teaching Dimension a combinatorial parameter of $\mathcal{H}$ and $\mathcal{X}$ (Hegedüs '95, Hellerstein et al '96)

Disagreement Coefficient a measure of the growth of the region of disagreement (Hanneke '07)

Neighborly Condition geometric relationship between $\mathcal{X}$ and $\mathcal{H}$ (Nowak '08)
What if there is noise or mismatch?

Noise-tolerance:

1. stochastic version space (all hypotheses with errors that could be explained by noise alone)
2. repeated querying (collect several labels for uncertain examples until highly confident in probably correct labeling)
3. hypothesis weighting (weight each hypothesis according to its prediction performance)

Agnostic active learning: If $h^*$ is not in $H$, then can we at least guarantee performance equal to that of passive learning? Yes

Split sample budget into three equal parts:
- active learning with 1/3 of sample budget $\rightarrow \hat{h}_n$
- passive learning with 1/3 of sample budget $\rightarrow \tilde{h}_n$
- remaining 1/3 of samples are collected from region of disagreement between $\hat{h}_n$ and $\tilde{h}_n$, best hypothesis wins!
Active Learning for Classification

$\mathcal{H}$ is a collection of functions mapping from $\mathcal{X} \rightarrow \{-1, +1\}$

Active learning can very effectively “narrow down” the location of the optimal decision boundary.
Example

Suppose we have a sensor network observing a binary activation pattern with a linear boundary. How many sensors must be queried to determine the pattern?

Correct boundary determined after querying 12 sensors
Active Learning for Regression

$\mathcal{H}$ is a collection of functions mapping from $\mathcal{X} \rightarrow \mathbb{R}$

Active learning can focus sampling in locations where the target function has the most unpredictable behavior.
Active Learning for Image Processing

8192 non-adaptive samples

+ 8192 adaptive samples

16384 non-adaptive samples
Active Learning for Fun!

"Is the person wearing a hat?"

"Does the person have blue eyes?"

"Binary Search" works very well in simple conditions.
\( \mathcal{X} = [0,1]^d \) or \( \mathbb{R}^d \) — The feature space

\( \mathcal{Y} = \{0,1\} \) or \( \mathbb{R} \) — The label space

\((x,y) \in \mathcal{X} \times \mathcal{Y} \sim P_{XY}\)

**Goal:** Construct a predictor \( h : \mathcal{X} \rightarrow \mathcal{Y} \) to minimize

\[ R(f) \equiv E [\text{loss}(Y, h(X))] \]
Passive Learning

Optimal prediction rule

\[ h^* = \arg \min_{h \in \mathcal{H}} \mathbb{E} [\text{loss}(Y, h(X))] \]

depends on \( P_{XY} \) (usually unknown)

However, we can learn a good prediction rule from a training sample

\[ \{(X_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} P_{XY} \]

Learning:

\[ \{(X_i, Y_i)\}_{i=1}^n \Rightarrow \hat{h}_n \]
Semi-Supervised and Active Learning

Unlabeled data are abundant, labels are expensive!

**SSL:** Design a predictor based on iid labeled and unlabeled examples:

\[
\{X_j\}_{j=1}^m \overset{iid}{\sim} P_X, \quad \{(X_i, Y_i)\}_{i=1}^n \overset{iid}{\sim} P_{XY} \quad \overset{\text{design}}{\Rightarrow} \quad \hat{h}_{m,n}
\]

(implicit assumption that \(P_{Y|X}\) is ‘linked’ to \(P_X\))

**AL:** Design a predictor based on iid unlabeled and select labeled examples:

\[
\{X_j\}_{j=1}^m \overset{iid}{\sim} P_X \quad \overset{\text{select}}{\Rightarrow} \quad \{(X_i, Y_i)\}_{i=1}^n \overset{\text{design}}{\Rightarrow} \quad \hat{h}_{m,n}
\]

(no need for a link between \(P_{Y|X}\) and \(P_X\))
Learning Rates and Sample Complexity

**Excess Error:** Error of learning algorithm compared to $h^*$.

$$\mathcal{E}(\hat{h}_n) := \mathbb{E}_{XY}[\text{loss}(Y, \hat{h}_n(X))] - \mathbb{E}_{XY}[\text{loss}(Y, h^*(X))]$$

**Consistency:** Learning algorithm error tends to zero.

$$\mathbb{P}(\hat{h}_n \neq h^*) \to 0$$

$$\mathbb{P}(\mathcal{E}(\hat{h}_n) > \varepsilon) \to 0$$  \hspace{1cm} (\mathbb{P} \text{ is wrt training data})

**Error Decay Rate:**

$$\mathbb{P}(\hat{h}_n \neq h^*) \leq \delta(n)$$

$$\mathbb{P}(\mathcal{E}(\hat{h}_n) > \varepsilon) \leq \delta(n, \varepsilon)$$

**Sample Complexity:** $n(\varepsilon, \delta) =$ minimum number of samples required to achieve $(\varepsilon, \delta)$ performance.
Research Questions

1. Is active learning always consistent?

2. Can active learning outperform passive learning?

3. If so, can we quantify the improvements of active learning in terms of error decay rates or sample complexity?

4. What conditions are especially favorable for active learning?

5. Are there fundamental limits to active learning that no algorithm can exceed?

6. How can we determine when an active learning algorithm is optimal?

7. Can practical active learning algorithms achieve optimal performance?
A few success stories…
Learning a decision hyperplane in $\mathbb{R}^d$


**Passive sampling:** uniformly at random $\Rightarrow$ err $\sim \frac{d}{n}$

**Selective sampling:** selected to be informative $\Rightarrow$ err $\sim e^{-n/d}$

Selective sampling yields exponential speed-up in learning!
Learning Rates for Multidimensional Thresholds

Compare with passive learning

Active Learning: Theorem (R. Castro and RN'07)

polynomial decay

exponential decay

learning curves
Now you see it, now you don’t!

\[ X = \text{sparse signal} + \text{noise} \]

An \( n \times 1 \) vector with \( n^{1-\beta}, 0 < \beta < 1 \), non-zero entries of magnitude \( \mu > 0 \). **Sparsistency**: Can the sparsity pattern be reliably perceived in presence of noise?

**Passive sensing**: Yes, if \( \mu > \sqrt{2\beta \log n} \), otherwise no.

**Selective sensing**: Yes, if \( \mu > a_n \), for any \( a_n \to \infty \).

Weak signals/patterns are imperceptible without selective sensing!