Active Learning

Literature Survey

Burr Settles
Presented by: Lovedeep Gondara

Technical report, 2010

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Outline

1. Introduction
    - What is active learning?
    - Example
2. Scenarios
    - Membership query synthesis
    - Stream-based selective sampling
    - Pool based sampling
3. Query Strategy Frameworks
    - Uncertainty sampling
    - Query-by-committee
    - Expected model change
    - Expected Error reduction
    - Variance reduction
    - Density weighted methods
4. Analysis of Active Learning
    - Empirical Analysis
    - Theoretical Analysis

What is active learning?

Definition

- Sub-field of machine learning, based on idea of letting model choose its own data.
- If it can, it should perform better.

What is active learning?

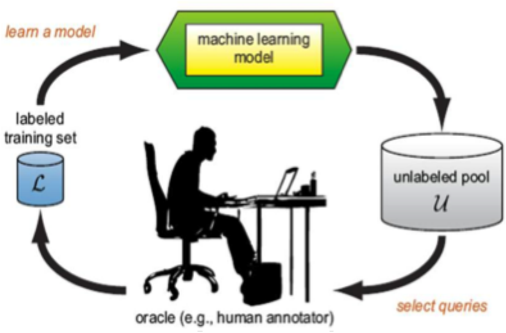
Definition

- Gathering labels for all data can be challenging.
- Active learning circumvents this issue by asking an oracle to label instances.
- Aims to achieve high accuracy using little data.

What is active learning?

Example

Figure: Pool based active learning cycle



Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Example

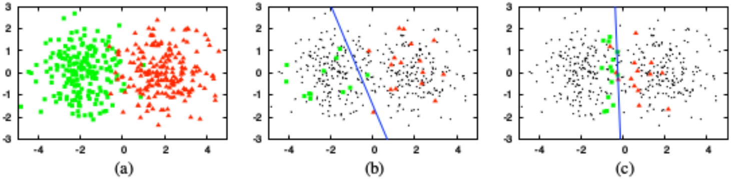
Pool based

- Queries are selected from a pool of unlabelled instances using uncertainty sampling.
- Selects the instance in the pool about which model is least certain.

Example

Pool based

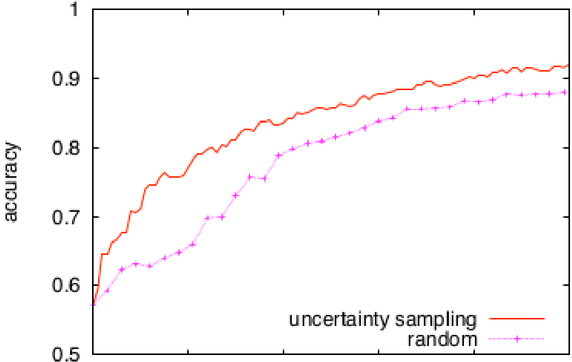
Figure: An illustrative example of pool-based active learning. (a) A toy data set of 400 instances, evenly sampled from two class Gaussians. The instances are represented as points in a 2D feature space. (b) A logistic regression model trained with 30 labeled instances randomly drawn from the problem domain. The line represents the decision boundary of the classifier (70% accuracy). (c) A logistic regression model trained with 30 actively queried instances using uncertainty sampling (90%).



Example

Pool based

Figure: Learning curves for text classification: baseball vs. hockey. Curves plot classification accuracy as a function of the number of documents queried for two selection strategies: uncertainty sampling (active learning) and random sampling (passive learning). We can see that the active learning approach is superior here because its learning curve dominates that of random sampling.



Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Membership query synthesis

- Learner requests labels for any unlabelled instance
- Assuming generated queries are de novo

Membership query synthesis

- Query synthesis can be awkward for human annotators
- Examples: Image annotation and NLP
- Works better when annotators are non-humans

Outline

1 Introduction
- What is active learning?
- Example

2 Scenarios
- Membership query synthesis
- Stream-based selective sampling
- Pool based sampling

3 Query Strategy Frameworks
- Uncertainty sampling
- Query-by-committee
- Expected model change
- Expected Error reduction
- Variance reduction
- Density weighted methods

4 Analysis of Active Learning
- Empirical Analysis
- Theoretical Analysis

Stream-based selective sampling

- Assumption: Unlabelled instance comes at no or minimal cost
- Sample first, then learner decided whether to ask for label or not

Stream-based selective sampling

- Uniform distribution: Similar to membership query synthesis
- Non uniform or unknown distribution: Still sensible queries

Selective sampling

What to query?

- Use some informative measure, such that more informative instances are more likely to be queried
- Region of uncertainty: Only query instances that fall within it

Outline

1. Introduction
    - What is active learning?
    - Example
2. Scenarios
    - Membership query synthesis
    - Stream-based selective sampling
    - Pool based sampling
3. Query Strategy Frameworks
    - Uncertainty sampling
    - Query-by-committee
    - Expected model change
    - Expected Error reduction
    - Variance reduction
    - Density weighted methods
4. Analysis of Active Learning
    - Empirical Analysis
    - Theoretical Analysis

Pool based sampling

- Assumption: Large amount of unlabelled instances are available
- Assuming a closed pool, queries are drawn from it

Pool based sampling

- Instances are queried in a greedy fashion
- Evaluate all instances in a pool using some informative measure

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Uncertainty sampling

- Query instances that active learner is least certain about
- Straightforward for probabilistic models

Uncertainty sampling

- Only considers information about most probable model
- Throws away information about remaining label distribution
- Margin sampling aims to correct for this bias

Uncertainty sampling

Margin sampling

- Margin sampling incorporates the posterior of second most likely label
- For very large label sets, still ignores much of output distribution
- Use entropy for a more general approach

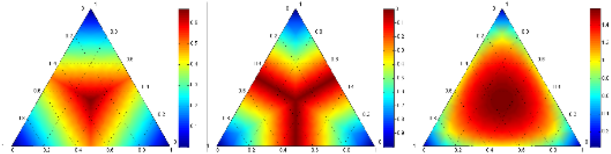
Uncertainty sampling

Entropy

- Entropy is information theoretic measure representing amount of information needed to encode a distribution
- For binary classification it reduces to margin and least confident strategies

Uncertainty sampling

Figure: Heatmaps illustrating the query behavior of common uncertainty measures in a three-label classification problem. Simplex corners indicate where one label has very high probability, with the opposite edge showing the probability range for the other two classes when that label has very low probability. Simplex centers represent a uniform posterior distribution. The most informative query region for each strategy is shown in dark red, radiating from the centers.



(a) least confident (b) margin (c) entropy

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

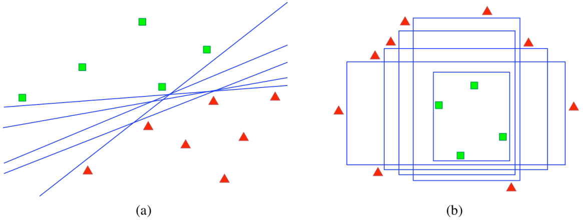
Query-by-committee

- Committee of models trained on same labelled set representing competing hypothesis
- Each member is allowed to vote on labelling of query candidates.
- Instance about which they most disagree is most informative query

Query-by-committee

- Minimize the version space, i.e. set of hypothesis that are consistent with current labelled training data

Figure: Version space examples for (a) linear and (b) axis-parallel box classifiers. All hypotheses are consistent with the labeled training data in L (as indicated by shaded polygons), but each represents a different model in the version space.



Query-by-committee

Disagreement

- Vote entropy (QBC generalization of entropy based uncertainty sampling)
- Average Kullback-Leibler divergence

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Expected model change

- Select the instance that would impart greatest change to the model if we knew its label
- Expected gradient length approach for discriminative probabilistic models
- Learner should query the instance which would result in new training gradient of largest magnitude

Expected model change

- Prefers instances that are likely to most influence the model
- Computationally expensive if feature space and and label set is large
- Non scaled features may cause issues

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Expected Error reduction

- Amount of generalization error reduction
- Query instance with minimal expected future error

Expected Error reduction

- Can be very computationally expensive
- Applications have only been considered in simple binary classification tasks

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Variance reduction

- Indirectly minimize generalization error by minimizing output variance
- Computationally expensive

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Density weighted methods

- Informative instances: not only uncertain, but also representative
- Inhabit dense regions of input space

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Does Active learning works?

- Literature suggests that it does
- Companies like Google, IBM, Microsoft use it
- All this indicates that Active Learning has matured to the age of practical use

Does Active learning works?

- Training set built in cooperation with an active learner is tied to the model that was used to generate it.
- Labelled instances are a biased distribution

Outline

1. Introduction
   - What is active learning?
   - Example
2. Scenarios
   - Membership query synthesis
   - Stream-based selective sampling
   - Pool based sampling
3. Query Strategy Frameworks
   - Uncertainty sampling
   - Query-by-committee
   - Expected model change
   - Expected Error reduction
   - Variance reduction
   - Density weighted methods
4. Analysis of Active Learning
   - Empirical Analysis
   - Theoretical Analysis

Does Active learning works?

- Remains elusive irrespective of recent advances
- Bound on number of queries required to learn a sufficiently accurate model?

Summary

- Active learning is an interesting concept
- There is still a lot of ground that need to be covered (Theoretically and Empirically)

For Further Reading I

Settles, Burr.
Active learning.
Morgan & Claypool Publishers, 2012.

Uncertainty Sampling

For problems with three or more labels

$$x^*_{LC} = argmax_x 1 - P_\theta(\hat{y}|x) \tag{1}$$

where $\hat{y} = argmax_x P_\theta(y|x)$ is the class label with highest posterior probability under the model θ.

Margin Sampling

$$x_M^* = argmin_x P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x) \tag{2}$$

where $\hat{y}_1$ and $\hat{y}_2$ are first and second most probable class levels.

Entropy

$$x_H^* = argmax_x - \sum_i P_\theta(\hat{y}_i|x) \log P_\theta(\hat{y}_i|x) \tag{3}$$

where $\hat{y}_i$ ranges over all possible class levels.

Query by committee

Vote entropy

$$x^*_{VE} = argmax_x - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C} \tag{4}$$

where $\hat{y}_i$ ranges over all possible class levels and $V(y_i)$ is number of votes a label receives, C being committee size.

Query by committee

KL divergence

$$x^*_{KL} = argmax_x \frac{1}{C} \sum_{c=1}^{C} D(P_{\theta(c)} || P_C) \qquad (5)$$

where

$$D(P_{\theta(c)} || P_C) = \sum_i P_{\theta(c)}(y_i|x) \log \frac{P_{\theta(c)}}{P_C(y_i|x)} \qquad (6)$$

$\theta(c)$ is a particular model in the committee, and C is committee as a whole, thus

$$P_C(y_i|x) = \frac{1}{C} \sum_{c=1}^{C} P_{\theta(c)}(y_i|x) \qquad (7)$$

is the consensus probability that y_i is correct label.

Expected model change

Let $\Delta l_\theta(L)$ be the gradient of the objective function l with respect to the model parameters θ.
Let $\Delta l_\theta(L \cup < x, y >)$ be the new gradient obtained by adding the training tuple $< x, y >$ to L.
AS query algorithm does not know the true label in advance, we instead calculate the length as an expectation over the possible labellings

$$x^*_{EGL} = argmax_x \sum_i P_\theta(y_i|x)||\Delta l_\theta(L \cup < x, y_i >)|| \tag{8}$$

where $||.||$ is Euclidean norm of each resulting gradient vector

Expected error reduction

Estimate expected future error of a model trained using $L \cup < x, y >$ on remaining unlabeled instances, query the instance with minimal expected future error, we can minimise 0/1 loss:

$$x^*_{0/1} = argmin_x \sum_i P_\theta(y_i|x)(\sum_{u=1}^{U} 1 - P_{\theta^{+<x,y_i>}}(\hat{y}|x^{(u)})) \tag{9}$$

where $\theta^{+<x,y_i>}$ is new model after retraining with $< x, y_i >$.
We can also use expected log-loss:

$$x^*_{log} = argmin_x \sum_i P_\theta(y_i|x)(-\sum_{u=1}^{U} \sum_j P_{\theta^{+<x,y_i>}}(\hat{y}_j|x^{(u)}) \log P_{\theta^{+<x,y_i>}}(\hat{y}_j|x^{(u)})) \tag{10}$$

Density Weighted Method

We wish to query instances as follows:

$$x^*_{ID} = argmax_x \phi_A(x) \times (\frac{1}{U}\sum_{u=1}^{U} sim(x, x^{(u)}))^{\beta} \tag{11}$$

$\phi_A(x)$ is informativeness of x according to some base query strategy A, such as QBC etc. Second term weights it by its average similarity to all other instances in input distribution. β controls the relative importance of density term.