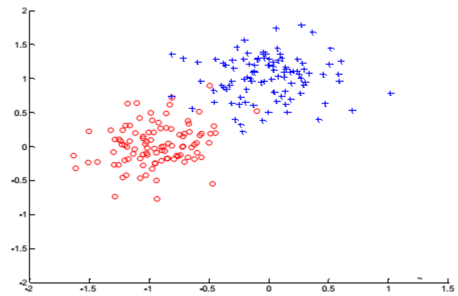
Active Learning

Nils Murrugarra Llerena
University of Pittsburgh

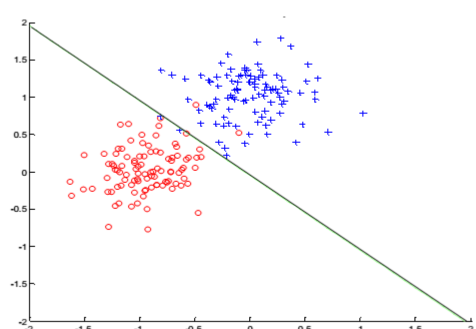
Outline

- Introduction
- Why to use active learning?
- Scenarios
- Query Strategies
- Analysis
- Extensions
- Practical Considerations
- Related Research Areas
- Conclusion

Introduction



Introduction



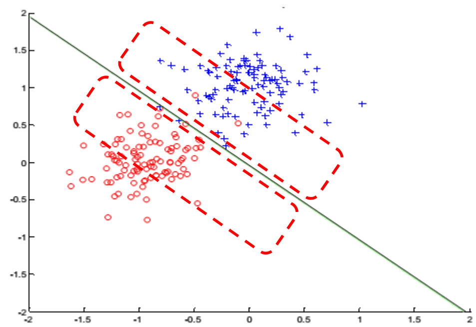
Classifier learned

Introduction

Active Learning

•If a learning algorithm is allowed to choose data from which to learn, it will perform better with less training data.

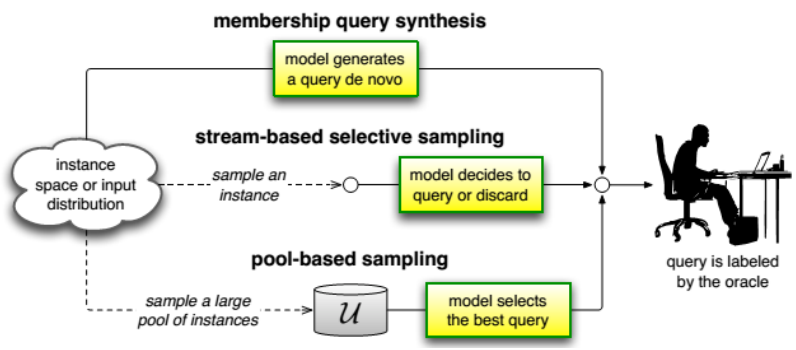
•This means that if the classifier learns the instances that are more “hard” to classify that will be a good classifier using less data.



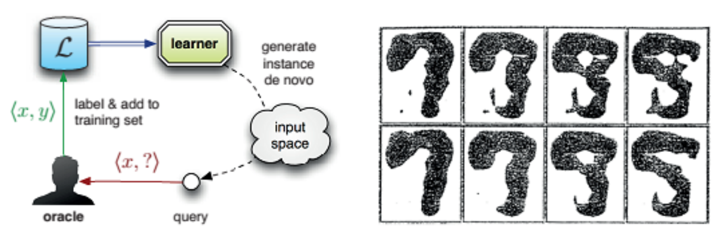
Why active learning?

- There are many tasks where labels are: **time-consuming** and/or **expensive** to obtain.
  - **Speech Recognition**
    - Trained Linguistics needed
    - Annotation at word level takes longer time than the audio length
  - **Information Extraction**
    - Finding entities and relations in a news text can take half-hour or more
    - Need some expertise in medical domains
  - **Classification and Filtering**
    - Annotating thousands of data examples can be tedious and redundant

Scenarios – Active Learning



Scenarios: Membership Query Synthesis



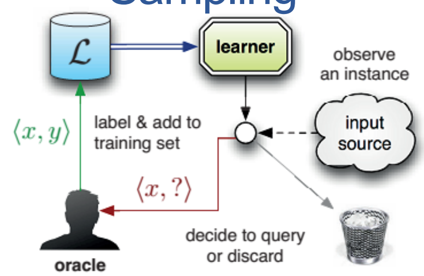
(a) query synthesis

(b) an example from handwriting recognition

- The learner has a definition of the input space
  - Feature dimensions and ranges are known.
- **Problem:**
  - Many generated images doesn't contain recognizable numbers (neural network).

9

Scenarios: Stream-Based Selective Sampling



- Obtain an unlabeled instance is free.

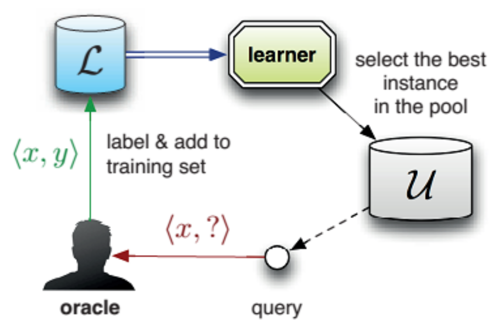
- **How to query?**
  - Use an " informativeness measure " or "query strategy" (threshold).
  - Compute a region of uncertainty and pick instances that fall in that region.

- **Applications:**
  - news on the web, computer network traffic, phone conversations, ATM transactions, web searches, sensor data, …

- **Advantage:**
  - Suitable for mobile and embedded devices (memory and power is limited)

10

Scenarios: Pool-based active learning



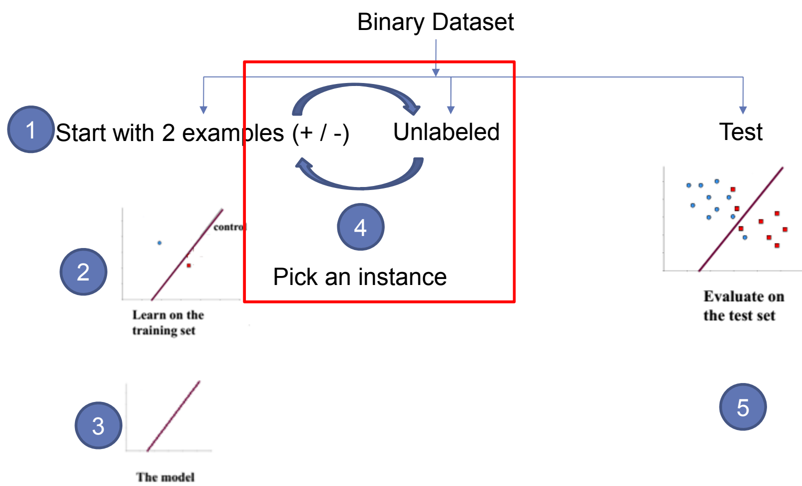
- Large amount of data can be collected at once.

- **How to select?**
  - Use an informativeness measure to evaluate all instances.

- **Comparison:**
  - Pool-based: ranks the entire collection.
  - Stream-based: scans the data sequentially and make individual decisions.

Scenarios: Example

Steps in an Active Learning Approach

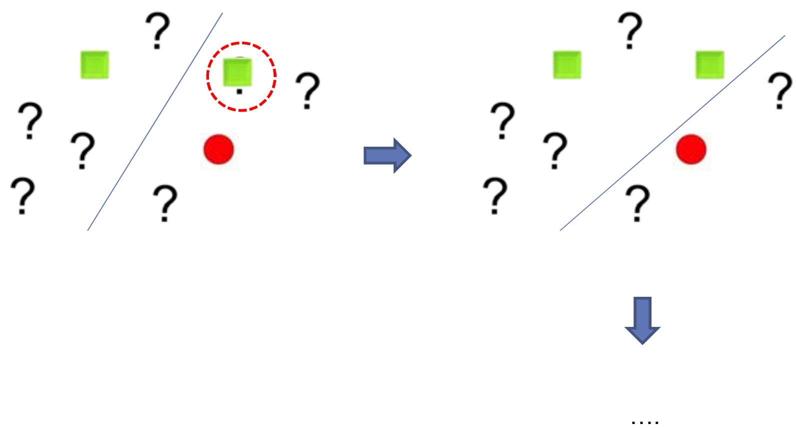


Query Strategies

How we evaluate the informativeness of unlabeled instances?

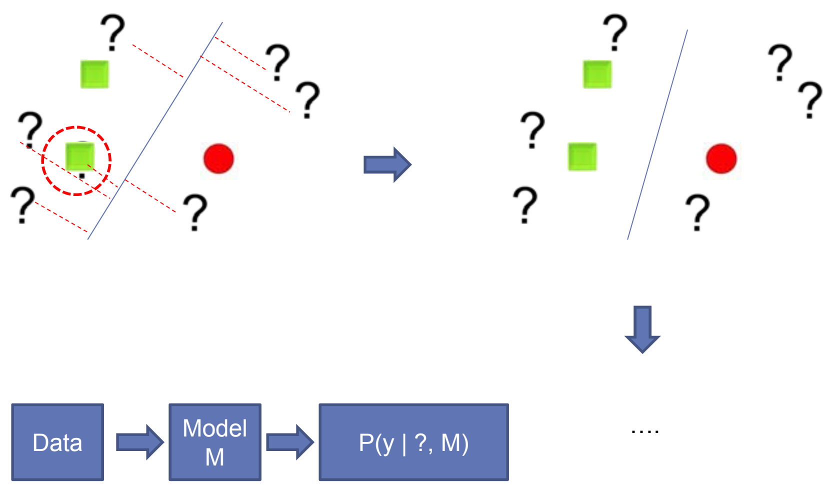
13

Query Strategy: Random



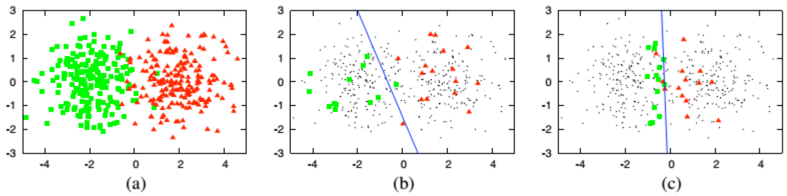
14

Query Strategy: Uncertainty



Uncertainty based to distance to margin

Query Strategies - Potential of AL: Random versus Uncertainty Sampling



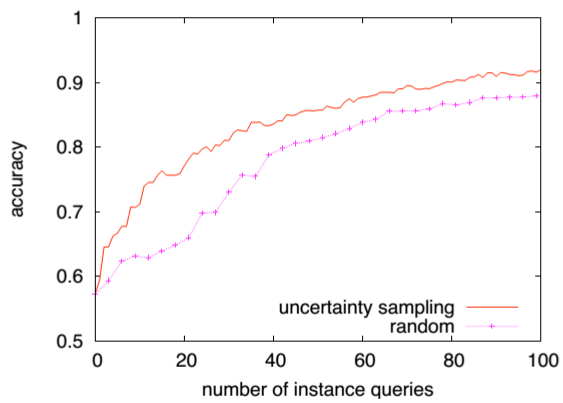
Two class Gaussians

30 instances selected randomly [70% accuracy]

30 instances selected using uncertainty sampling [90% accuracy]

How we compare Query Strategies?

Query Strategies – AL Evaluation



Learning Curves

Learning curves are employed to compare query strategies

Query Strategy: Uncertainty Sampling

Least Confident

•Query an instance for which the learner is least certain how to label it.

- Two classes: Select the instance whose positive posterior probability is near 0.5
- Three or more: Select the instance whose prediction is the least confident.

$$x^*_{LC} = \arg\max_x (1 - P_\theta(\hat{y} \mid x))$$

ŷ : class label with the highest probability

$$\hat{y} = \arg\max_y P_\theta(y \mid x)$$

Query Strategy: Uncertainty Sampling

$$x^*_{LC} = \arg\max_x (1 - P_\theta(\hat{y} \mid x))$$

- **P ≈ 0**, produce a **higher value** (1) => Pick least certain classifier
- **P ≈ 1**, produce a **lower value** (0)

The model's belief that it will mislabel x.

Drawback

- It only considers information about the most probable label.
  - Throws away information about the remaining label distribution.

Query Strategy: Uncertainty Sampling

Margin Sampling

$$x_M^* = \arg\min_x (P_\theta(\hat{y}_1 | x) - P_\theta(\hat{y}_2 | x))$$

$\hat{y}_1$ and $\hat{y}_2$: first and second most probable class labels under the model θ

- **Large margin**, instances easy to differentiate
- **Small margin**, more ambiguous to differentiate

Drawback

- For very large label sets, the margin approach still ignores the output distribution of the remaining classes.

How to incorporate all labels distribution?

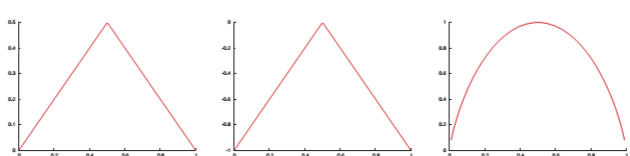
Query Strategy: Uncertainty Sampling

Entropy

$$x_H^* = \arg\max_x (H_\theta(Y | x))$$

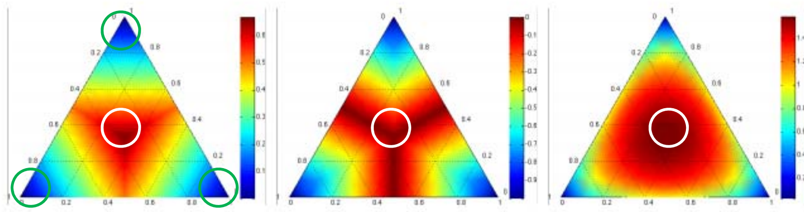
$$= \arg\max_x (-\sum_y P_\theta(y | x) * \log P_\theta(y | x))$$

- Is a measure of variable's average information content.
  - Impurity measure
    - **Worst case**, (2 classes), probability 0.5
  - Measure if all labels have very similar classification probabilities



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

Query Strategies: Uncertainty Sampling



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

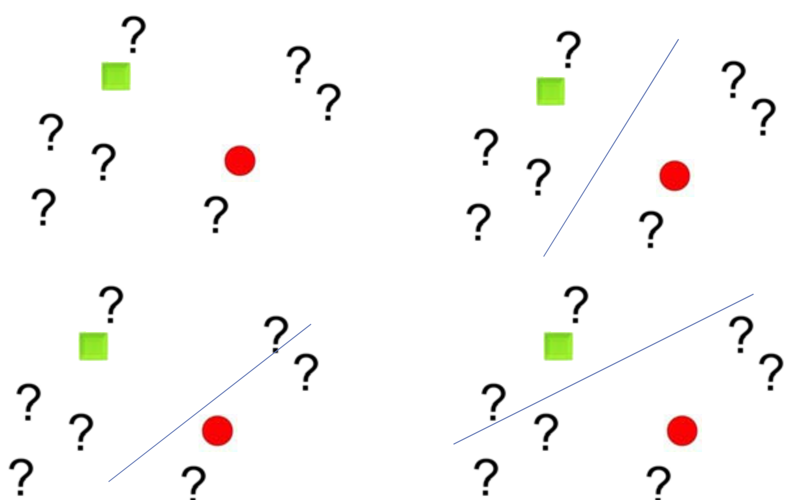
3 Classifiers: c1 vs (c2, c3); c2 vs (c1,c3) and c3 vs (c1,c2)

Advantage (entropy)

- Doesn't favor instances where only one of the labels is highly unlikely.
  - **Entropy**, minimize log-loss
  - **Least confident and margin**, reduce classification error.

Why if we incorporate more than one model ?

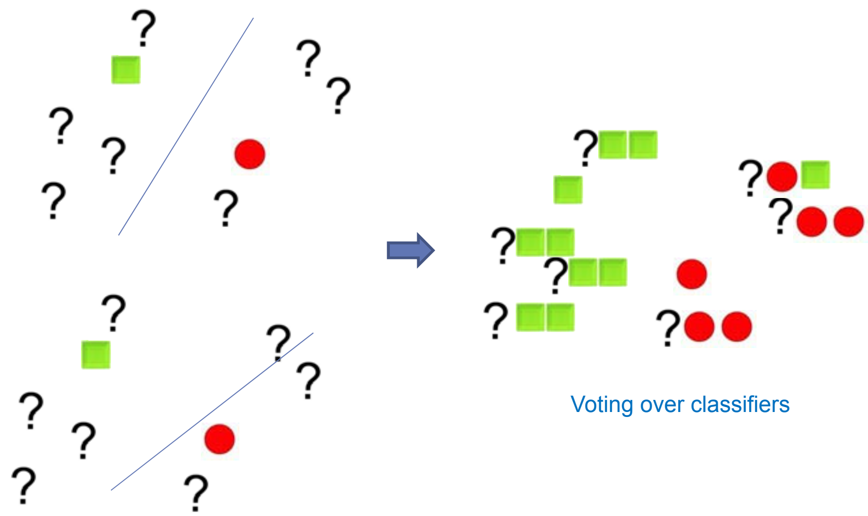
Query Strategy: Query-By-Committee



Different model initialized randomly by a Perceptron approach

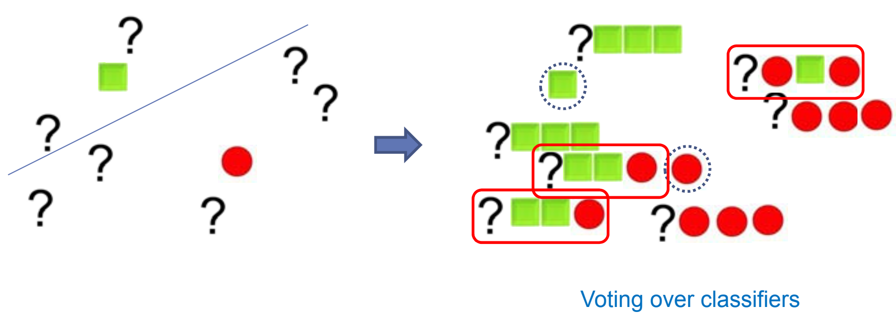
23

Query Strategy: Query-By-Committee

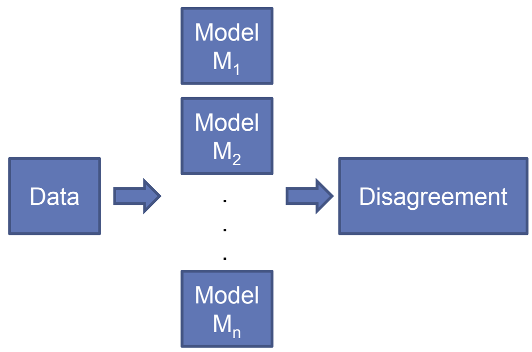


24

Query Strategy: Query-By-Committee



Query Strategy: Query-By-Committee



How to generate different models?

Query Strategy: Query-By-Committee

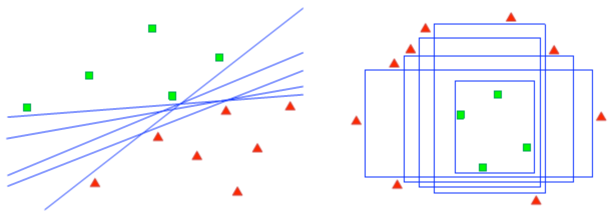
How to generate different models?

- Use a bootstrap procedure (e.g. bagging) to subsample the labeled dataset.
- Try different parameters in the classifier
  - Radial SVM, change gamma and cost parameters
  - Decision trees, try different pruning algorithms.

Query Strategy: Query-By-Committee

Maintain a committee of learners C = {θ$^{(1)}$, . . . , θ$^{(C)}$}, which are all trained in the labeled set **L** (or subsets).

- Each learner **vote** on the label of the query candidate
- Pick the instance where they most **disagree**.



Considerations

- Consider learners that represent different regions
- Have a measure of disagreement among the learners

How measure disagreement for more than 2 classes?

Query Strategy: Query-By-Committee

Disagreement measures

- **Vote entropy**

$$x^*_{VE} = \arg\max_x \left(-\sum_i \frac{V(y_i)}{C} * \log \frac{V(y_i)}{C}\right)$$

Measure Impurity

- **V(y$_i$)**, number of votes the label y_i receives
- **C**, committee size

- **KL - Divergence**

$$x^*_{KL} = \arg\max_x \left(\frac{1}{C} * \sum_{c=1}^{C} D(P_{\theta(c)} \| P_C)\right)$$

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

- **θ$^{(C)}$**, a model in the committee
- **C**, all the committee

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

Query Strategy: Query-By-Committee

Disagreement measures

- **KL – Divergence**
  - It measures the difference between two probabilities
  - **Most informative query**: Instance that has the largest average difference between:
    - any one committee member
    - and the consensus (all learners)

- **KL - Divergence**

$$x^*_{KL} = \arg\max_x \left(\frac{1}{C} * \sum_{c=1}^{C} D(P_{\theta(c)} \| P_C)\right)$$

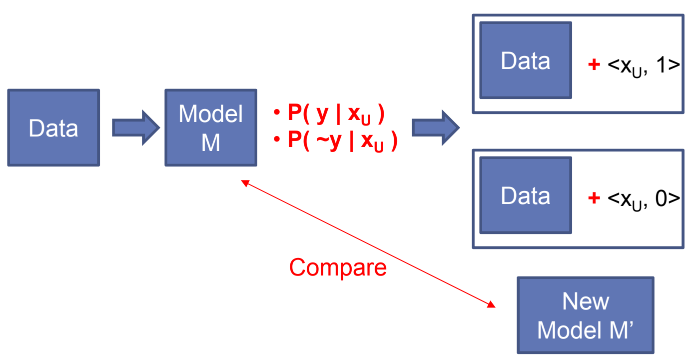
$$D(P_{\theta(c)} \| P_C) = \sum_i P_{\theta(c)}(y_i \mid x) * \log \frac{P_{\theta(c)}(y_i \mid x)}{P_C(y_i \mid x)}$$

- $\theta^{(C)}$, a model in the committee
- **C**, all the committee

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

Query Strategy: Expected Model Change

- Select the instance that would impact the greatest change to the current model



Compare and quantify the change due to the point inclusion in the labeled set

Query Strategy: Expected Model Change

- **Expected Gradient Length (EGL)**
  - Can be applied to any learning algorithm that uses gradient based parameter training
  - It determines the importance of the data point with respect to its influence on the model parameters (their change)

$\nabla E(\theta)$: Gradient of error E with respect to the current model θ (M)

$$\nabla E(\theta) = [\frac{\partial E}{\partial \theta_1}, \frac{\partial E}{\partial \theta_2}, ..., \frac{\partial E}{\partial \theta_m}]$$

- instance <x_i, y> is selected

$\nabla E_i^+(\theta)$: new gradient by adding <x_i, 1>

$\nabla E_i^-(\theta)$: new gradient by adding <x_i, 0>

Combine

$$= o_i \, \| \nabla E_i^+(\theta) \| + (1 - o_i) \, \| \nabla E_i^-(\theta) \| \qquad \| x \| = \sqrt{x_1^2 + ... + x_n^2}$$

Query Strategy: Expected Model Change

- **Expected Gradient Length (EGL)**
  - **How to measure the impact/change?**: Consider the norm of the training gradient (i.e. vector used to re-estimate parameter values).

$$x^*_{EGL} = \arg\max_x (\sum_i P_\theta(y_i \mid x) * \| \nabla \ell_\theta(T \cup \langle x, y_i \rangle) \|)$$

$$\nabla \ell_\theta(T \cup \langle x, y_i \rangle) \approx \nabla \ell_\theta(\langle x, y_i \rangle)$$

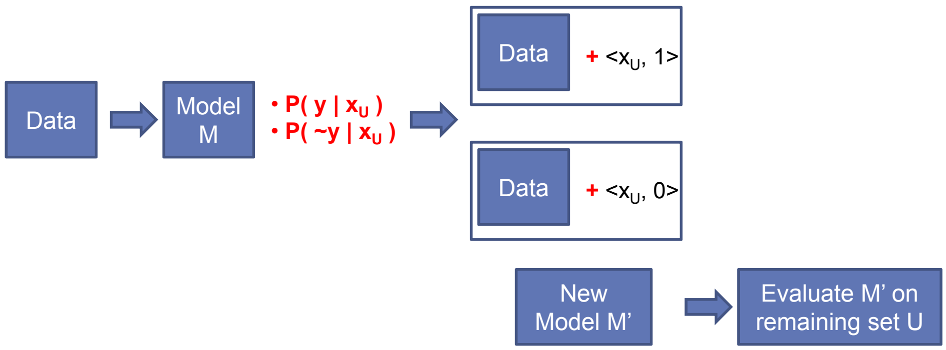
We don't know the correct label y, for that we consider an expectation over all possible labels.

Drawback

- Computational expensive, if both the feature space and set of labels are very large

Query Strategy: Expected Error Reduction

- Select the instance that **reduce** the **generalization error**.



Does it reduce the error? (incorrect prediction)

Is model M' more confident?

Query Strategy: Expected Error Reduction

- Select the instance that **reduce** the **generalization error**.

- **Minimize the Expected 0/1-loss function**

$$x^*_{0/1} = \arg\min_x \Big(\sum_i P_\theta(y_i \mid x) * \Big(\overbrace{\sum_{u=1}^{U} (1 - \underbrace{P_{\theta+<x,y_i>}}_{\text{New model after train with <x,y}_i\text{>}}(\hat{y} \mid x^{(u)}))}^{\text{Loss function on unlabeled data (\# of incorrect predictions)}}\Big)\Big)$$

Goal: Reduce the expected total number of incorrect predictions.

Query Strategy: Expected Error Reduction

•Reduce expected entropy over U

$$x^*_{\log} = \underset{x}{\operatorname{argmin}}\Big(\sum_i P_\theta(y_i \mid x) * \overbrace{\Big(\sum_{u=1}^{U} \underbrace{-\sum_j P_{\theta+<x,y_i>}(y_j \mid x^{(u)}) * \log P_{\theta+<x,y_i>}(y_j \mid x^{(u)})}_{\text{Entropy}}\Big)}^{\text{Entropy over U}}\Big)$$

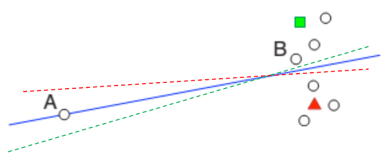
Goal: Increase confidence in prediction (minimize entropy).

Drawback
- Most computational expensive framework,
  - require estimate the future error over U for each query
  - a new model is retrained for each query (iterate over all the pool)
- Usually employed in binary classification tasks.

Query Strategy: Density-Weighted Methods

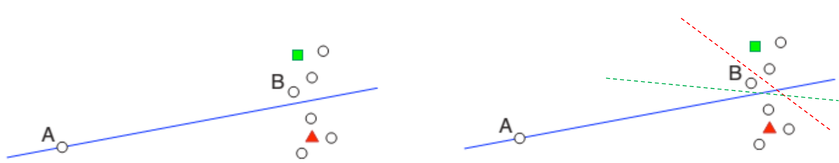
Previous Approaches
- Uncertainty, QBC and EGL are more likely to pick **outliers**
  - Uncertainty: See example
  - QBC and EGL could pick possible outliers
    - Controversial
    - Generate significant change in the model
- Expected error **avoid** the previous problems (less probable to pick outliers)
  - Because they focus on the entire input space than individual instances.

Query Strategy: Density-Weighted Methods



- **Margin**, select point A
- pointA haven't give too much information.
  - Decision boundary almost the same.
  - data distribution as a **whole**

Query Strategy: Density-Weighted Methods

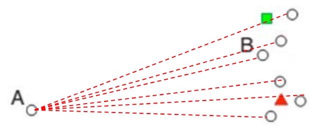


- point B is more informative about the data distribution

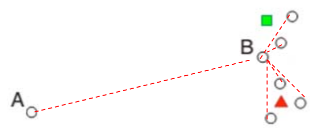
How to combine informativeness and data distribution information?

Query Strategy: Density-Weighted Methods

Data Distribution



Get average distance



- Distance ≈ 0, similar examples (Dissimilarity measure)

Similarity measure
- Similar examples, value ≈ 1

Query Strategy: Density-Weighted Methods

Model the input distribution during the query selection
- Define informative instances as:
  - uncertain
  - are “representative” of the data distribution

Average similarity to all other instances
- ≈ 1, more similar with all data

$$x^*_{ID} = \arg\max_x \phi_A(x) * \left(\frac{1}{U} * \sum_{u=1}^{U} sim(x, x^{(u)})\right)^{\beta}$$

Control parameter

Informativeness of query (e.g. uncertainty sampling)
- ≈ 1, more informative

Analysis of Active Learning

Empirical Analysis

• AL helps to reduce the number of labeled instances required to achieve a certain accuracy in the majority of reported results.

Theoretical Analysis

- **Would be Nice!!**
  - Sort of bound in the number of queries to learn a sufficient accurate model
  - This number should be less than passive learning.

• Let's consider instances in one-dimensional line and our model is:

$$g(x;\theta) = \begin{cases} 1 & if\,(x > \theta), and \\ 0 & otherwise \end{cases}$$

Analysis of Active Learning

Theoretical Analysis

• Let's consider instances in one-dimensional line and our model is:

$$g(x;\theta) = \begin{cases} 1 & if\,(x > \theta), and \\ 0 & otherwise \end{cases}$$

According to **PAC model**

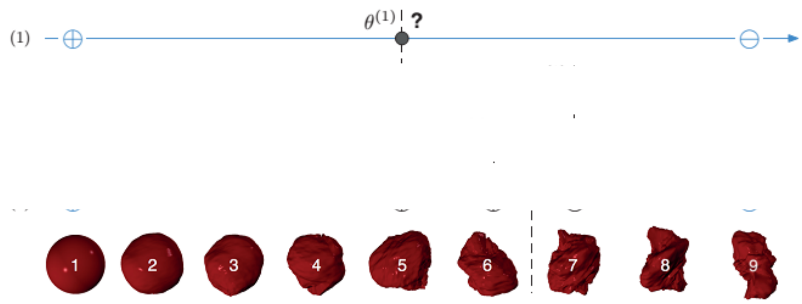
- The data distribution can be perfectly classified with O(1/e) random labeled instances.

Pool-based AL

- Consider the point on a real line: their labels are a sequence of 0's and 1's.
- **Goal:** Discover the location where the transition occurs

Analysis of Active Learning

Theoretical Analysis



Pool-based AL

•Using a binary search. A classifier with error less than e can be obtained with O(log 1/e)

Analysis of Active Learning

Theoretical Analysis

According to **Bayesian Assumption**

- It is possible to achieve generalization error e after seeing O(d/e) unlabeled instances (d is the VC dimension).

Stream-based and Pool-based AL (QBC)

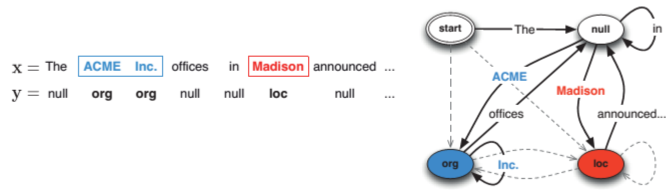
- It is possible to achieve generalization error e, requesting only O(d log 1/e)
- Exponential improvement

Extensions of Active Learning

AL for Structured Outputs

- Sequential models can produce a probability distribution for every possible label sequence **y**, the number of which can grow exponentially in the sequence length **T**.

- Least confident approach is famous in this setting, because the most likely output sequence ŷ and the associated $P_{\theta}(\hat{y} \mid x)$ can be efficiently computed with dynamic programming (Viterbi algorithm).



Extensions of Active Learning

Active Feature Acquisition

Instances may have incomplete feature descriptions

- Credit card company can have access to their clients information but not the transactions for other credit companies
- For medical diagnosis, can have access to some basic symptoms, but not all (complex, expensive or risky procedures)

Goal: Select most informative feature to obtain (request) [train time]

Solution:

- Impute the missing values and then acquire the ones that the model is less certain

Extensions of Active Learning

Active Classification
Missing feature values can be acquired at test time.

Active Class Selection
Query an instance of a given class label

Active Clustering
Subsample unlabeled instances in a way that they self-organize into groups:
- less overlap or noise

Practical Considerations

Batch-Mode Active Learning
Majority of active learning techniques consider that queries are selected one at a time.
- time to induce a model is **expensive**
- All process is **inefficient**

Goal: Query instances in groups.

How to select the optimal query set?
- k-best queries doesn't work properly
  - it fails to consider overlap information in k-best instances
- Most approaches use greedy heuristics that instances in the query are diverse and informative.
- e.g. query centroids of clusters that lie closes to the decision boundary

Practical Considerations

Noisy Oracles

Even if labels come from human experts, they might not be reliable:

- Some instances are really difficult to annotate
- People can be distracted or fatigued over time

How to use non-experts as oracles?

- Averaging labels of multiple non-experts

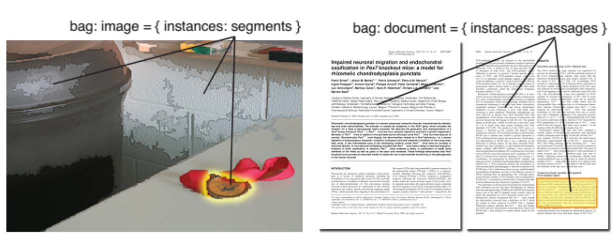
Practical Considerations

Alternative Query Types

• Multiple-instance Active Learning

Instances are grouped in bags:

- labeled **negative**, if all of its instances are negative
- labeled **positive**, if at least one instance is positive



(a) (b)

Advantages

- Coarse labels sometimes are available at low cost.
- Allowed to query for labels are finer granularity.
- Could consider approaches of mixed-granularity.

Practical Considerations

Alternative Query Types
• Tandem Learning
- Interleave instance-label queries with feature-salient queries.
- e.g. is the word "ball" a discriminative feature for sport documents ?

Multi-Task Active Learning
Same instances may be labeled in multiple ways for different subtasks.
- parsing and NER
  - Alternating
  - Rank-combination, each task rank the queries and select the highest combined rank
- Images for binary classification tasks.

Stopping Criteria
When accuracy has reached a non-change state?
- Use intrinsic measure of stability within the learner.
  - If the measure degrades, STOP active learning
- Real Stop, based on economic factors (before intrinsic measures)

Related Research Areas

Semi-supervised Learning (SSL)
In conjunction with AL, they try to get the most out of the unlabeled data

- Self training pick the most confident unlabeled instance. In contrast, AL uncertainty sampling pick the least confident instance.
- Co-training consider ensemble methods as QBC consider them for AL.

AL and SSL attack the problem from opposite directions

Reinforcement Learning
- In order to improve
  - the learner must take risks and try actions for which it is uncertain about the final result (as AL)

Equivalence Query Learning
- Similar to membership query learning
- It generates an hypothesis of the target concept class
  - The oracle confirm or deny the hypothesis

Related Research Areas

Model Parroting and Compression

- Neural Networks achieve better generalization accuracy than decision trees in many applications.
- Decision trees are more comprehensible by humans.

Proposal: Extract high accurate decision trees from neural networks.

AL

- Consider an "oracle model", trained using a small set of the available labeled data
- Consider a "parrot model", that can query using the "oracle model"
  - label of any unlabeled data (pool-based)
  - Synthesize new instances (membership-query)

Conclusions

- AL is a growing research area
  - Data is **easy** to **obtain**
  - **Difficult/costly** to **label**
- AL has been studied related to:
  - scenarios
  - query strategies
  - Extensions
  - Practical Considerations
  - Related Areas
- However there are still much work to do and open questions …

Questions