Active Learning

Review of: Active Learning Literature Survey by Burr Settles

Seyoon Ko

Seoul National University

2016/12/03

Contents

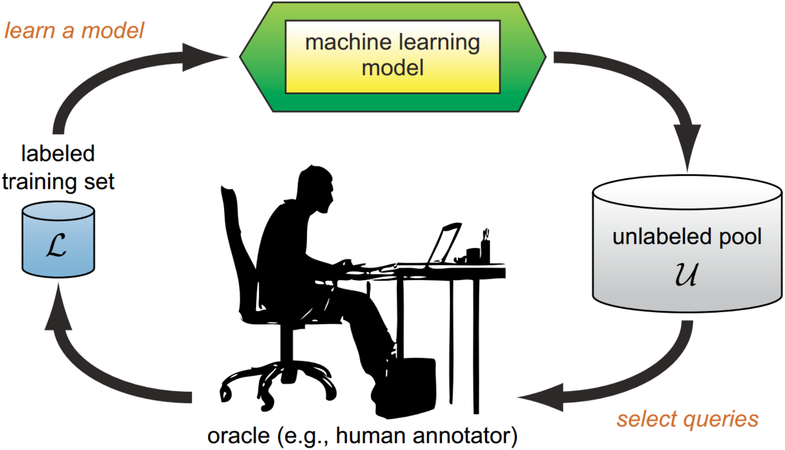
Section 1

Introduction

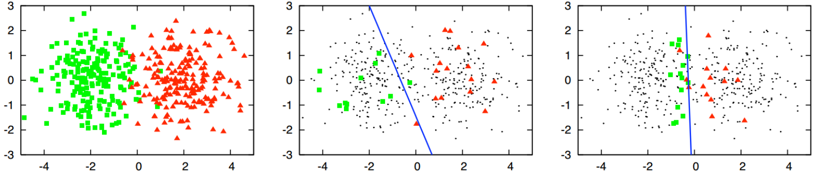
What is Active Learning?

- Subfield of ML and AI
- Learning algorithm is allowed to choose the data from which it learns
  $\rightarrow$ performs bettter with less training
- Especially when retrieving data is cheap, but labeling is not
  - Speech Recognition
  - Information Extraction
  - Classification and Filtering (of media)
- Overcomes the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle
  - Aims to achieve high accuracy using as few labeled instances as possible

Active Learning Cycle



Active Learning Examples: Logistic Regression

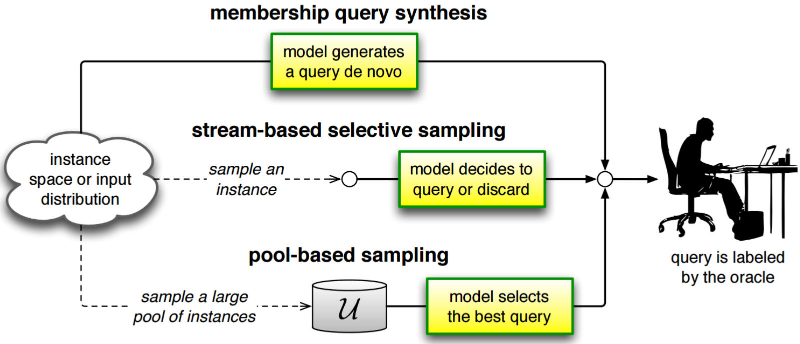


(a) (b) (c)

Section 2

Scenarios

Scenarios Overview



Membership Query Synthesis

- Learner may request labels for any unlabeled instance in the input space
  - includes queries that learner generates de novo, rather than those sampled from underlying natural distribution
- Tractable and efficient for finite for finite problem domains [Angluin, 2001]
- Regression learning tasks: learning to predict the abs. coord. of a robot hand given the joint angles of its mechanical arm as inputs [Cohn et al., 1996]
- Can be awkward if the oracle is human annotator
  - traing a NN to classify handwritten characters: synthetized image had no recognizable symbols, only artifical hybrid characters
  - What would happen for NLP?
- works well for "robot scientist" scenario
  - a laboratory robot autonomously synthesizes composition of mixture of chemicals, and physically performs experiment [King et al., 2004, 2009]

Stream-based Selective Sampling [Cohn et al., 1990, 1994]

- sample from the actual distribution, learner decides whether to request its label
- stream-based or sequential
- labeling by...
  - use query strategy to decide whether to label an example (to come)
  - explicitly set region of uncertainty [Cohn et al., 1994]
- part-of-speech tagging [Dagan and Engelson, 1995]
- Sensor scheduling [Krishnamurthy, 2002]

Pool-Based Sampling [Lewis and Gale, 1994]

- a small set of labeled data $\mathcal{L}$ and a large pool of unlabeled data $\mathcal{U}$ avalable
- queries selectively drawn from the pool (usually non-changing)
  - typically in greedy fashion according to certain informativeness measure used to evaluate all instances in the pool
- Text classification

  [Lewis and Gale, 1994; Callum and Nigam, 1998; Tong and Koller, 2000; Hoi, et al., 2006a]
- Information Extraction [Thompson et al., 1999; Settles and Craven, 2008]
- Image classification and retrieval [Tong and Chang, 2001; Zhang and Chen, 2002]
- Video classification and retrieval [Yan et al., 2003; Hauptmann et al., 2006]
- Speech recognition [Tur et al., 2005]
- Cancer Diagnosis [Liu, 2004]

Stream-based vs Pool-based

- Stream-based method:
  - scans through the data sequentially and makes query decisions individually
  - effective when memory or processing power is limited e.g. mobile and embedded system
- Pool-based method:
  - evaluates and ranks the entire collection before selecting the best query
  - much more common

Section 3

Query Strategy Frameworks

Query Strategy

- Criteria for choosing which sample to query
- Uncertainty sampling
- Committee-based
- Expected model change
- Expected error reduction
- Variance reduction
- Density-weighted

Uncertainty Sampling

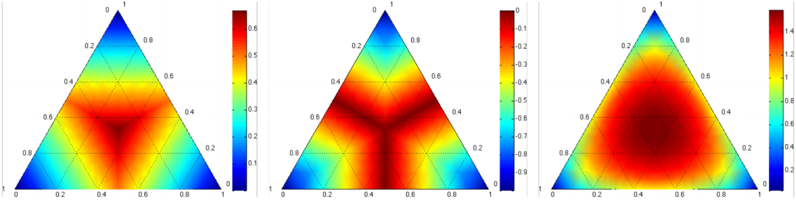
- queries instances which is least certain how to label
- straightforward for probabilistic learning models
- binary classification: instance whose posterior prob of being positive is nearest to 0.5

Uncertainty Sampling: Multi-class case

- Least confident, $\arg\max_x 1 - P_\theta(\hat{y}|x)$
  - $\hat{y} = \arg\max_y P_\theta(y|x)$
  - Natural in e.g. softmax models,
  - statistical sequence models in information extraction tasks: most likely sequence and likelihood can be computed using DP
  - [Culotta and McCallum, 2005; Settles and Craven, 2008]
- Margin sampling [Scheffer et al., 2001], $\arg\min_x P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)$
  - $\hat{y}_1$, $\hat{y}_2$: 1st and 2nd most probable class labels under the model
  - for large label sets, still ignores much of the output distribution for the other classes
- Shannon entropy $\arg\max_x - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$
  - well generalized for any number of class labels, or models for sequences[Settles and Craven, 2008]
  - or trees[Hwa, 2004]

Uncertainty Sampling: Multi-class Case

- Empirical comparisons showed mixed, application-dependent results (still better than baselines)
  - Author says: entropy if objective is to minimize log-loss, margin/LC to reduce classification error



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

Uncertainty Sampling in various cases

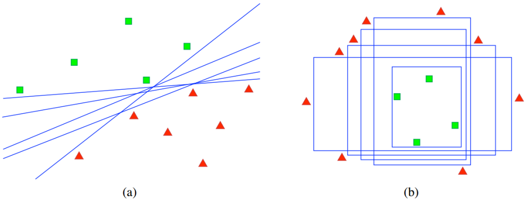
- Decision tree [Lewis and Catlett, 1994]: heuritics involving loss ratio, FP/FN
- Nearest-neighbor classifiers [Fujii et al., 1998; Lindenbaum et al., 2004]: voting
- SVMs [Tong and Koller, 2000]: distance to decision boundary
- Regression: unlabeled instance with highest output variance in its prediction
  - under gaussian assumption, equivalent to entropy
  - closed-form approximations of output variance can be computed for e.g. GRF, NN
  - Optimal experimental design [Federov, 1972]

Query-By-Committee [Seung et al., 1992]

- maintains a committee $\mathcal{C} = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$ of models trained on $\mathcal{L}$, representing competing hypotheses
- queries controversial regions of the input space

Query-By-Committee (2)

- Version space: set of hypotheses consistemt with the current labeled training set



- constrain the size of version space as much as possible, by using multiple hypotheses
- We need:
  - committee of models that represent different regions of the version space
  - measure of disgreement among committe members

Query-By-Committee (3)

- Hypothesis Selection
  - sampling a committee of two random hypotheses consistent with $\mathcal{L}$ [Seung et al., 1992]
  - Generative models: sampling from posterior
  - bagging and boosting [Abe and Mamitsuka, 1998]
  - ensemble encouraging diversity [Melville and Mooney, 2004]
- Disagreement Measure
  - vote entropy[Dagan and Engelson, 1995]
  - KL Divergence[McCallum and Nigam, 1998]
  - other divergences

Expected Model Change

- select the instance that would cause greatest change to the current model if we knew its label
- Expected Gradient Length [Settles et al., 2008b]
  - Useful for gradient-based training

$$\arg\max_x \sum_i P_\theta(y_i|x) \left\| \nabla l_\theta(\mathcal{L} \cup \langle x, y_i \rangle) \right\|$$

- can be computationally expensive if both feature space and set of labelings are very large
- should be well-normalized
  - can use regularization to control this effect

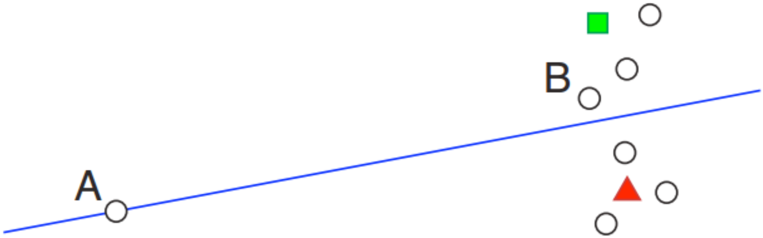
Methods Considering whole $\mathcal{U}$ or the entire input space

- Expected Error Reduction
  - estimate expected future error of a model trained using $\mathcal{L} \cup \langle x, y \rangle$ on the remaining unlabeled instances in $\mathcal{U}$ e.g. for expected log-loss:

$$\arg\min_x \sum_i P_\theta(y_i|x) \left(\sum_{u=1}^{U} \sum_j P_{\theta+\langle x, y_i \rangle}(y_j|x^{(u)}) \log P_{\theta+\langle x, y_i \rangle}(y_j|x^{(u)}) \right)$$

  - can be interpreted as maximizing expected information gain or mutual information of output
- Variance Reduction
  - minimizing trace, det, eigenvalues of inverse Fisher information matrix
  - From theories of optimal experimental design
- (+): Focuses on entire input space rather than individual instances, less prone to outliers
- (-): relatively slower than other methods
  - Expected error reduction: computation over the entire pool for each intance

Outlier problem



Density-Weighted Methods

- Utilize $\mathcal{U}$ when estimating future errors and output variances by weighting
- Information density framework [Settles and Craven, 2008]:

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

- ϕ_A: informativeness of x according to base strategy (uncertainty sampling or QBC)
- second term: average similarity to all other instances in the input distribution
  - can be pre-computed and cached
  - advantageous for interactive real-time oracles
  - β: controls relative importance of the density term
- can set similarity as clustering results
  - e.g. ∞ for other clusters, average similarity to instances for same clusters

Section 4

Analysis of Active Learning

Empirical Analysis

- Does it work?
  - YES. many publications, and the author's personal acquaintances at CiteSeer, Google, IBM, Microsoft, and Siemens say so.
- Caveats
  - training built in cooperation with an active learner is inherently tied to the model that was used to generate it: labeled instances are from biased distribution
  - can sometimes need more training samples even for same models[Schein and Ungar, 2007]
  - proficiency of annotator is correlated with how well active learning helps[Baldridge and Palmer, 2009]

Theoretical Analysis

- find a bound on the number of queries required to learn a sufficiently accurate model for a given task
- theoretical guarantees that this number is less than in the passive supervised learning
- simple case: 1D binary thresholding

$$g(x;\theta) = \begin{cases} 1 & \text{if } x > \theta \\ 0 & \text{o.w.} \end{cases}$$

- We need $O(1/\epsilon)$ samples to achieve error bound ϵ with high probability
- pool-based setting: we can perform binary search on unlabeled data, and $O(\log(1/\epsilon))$ samples are enough, exponential reduction

Theoretical Analysis (2)

- Query-by-committee: under Bayesian assumption, generalization error ϵ is achieved
  - with $O(d/\epsilon)$ samples
  - with $O(d \log(1/\epsilon))$ of them labeled
  - d: VC dimension
- A variant of perceptron update [Dasgupta et al., 2005]
  - Same asymptotic result
  - without Bayesian assumption
  - lightweight and efficient

Theoretical Analysis (3)

- general pool-based setting, if using linear classifiers:
  - $O(1/\epsilon)$ needed in worst case, not better, but also not worse than passive supervised learning [Dasgupta, 2004]
- certain active learning strategies should always be better than supervised learning at the limit [Balcan et al., 2008]
- Agonistic actie learning [Balcan et al., 2006]
  - only requires that unlabled instances are drawn i.i.d., without needing to know correct concept class in advance
  - Polynomial time reduction [Dasgupta et al., 2008]
  - explicitly use complexity bounds and queries can be assessed by how valuable they are in distinguishing among these simple hypotheses

Theoretical Analysis (4)

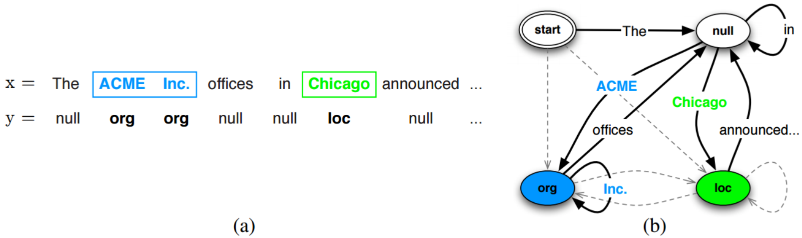
- most positive theoretical results are based on intractable algorithms, or too complex and particular to be used in practice
  - analyses on efficient algorithms are based on uniform or near-uniform input distributions [Balcan et al., 2006; Dasgupta et al., 2005], or severely restricted hypothesis spaces
  - usually only for simple classifications, minimizing 0-1 loss
  - some needs explicit enumeration of version spaces: usually intractable
- Some recent work has begun to address these issues
  - Hierarchical sampling [Dasgupta and Hsu, 2008]
  - Importance-weighted [Beygelzimer et al., 2009]

Section 5

Setting Variants

Active Learning for Structured Outputs

- Information extraction:
  - example: sequence labeling
  - input: structured sequence of feature vectors
  - output: structured e.g. sequences, trees



- many works on CRFs, HMMs, probabilistic context-free grammars, etc.

Active Feature Acquisition, Classification, and Class Selection

- When feature is expensive
- Incomplete feature descriptions:
  - incomplete customer data, client disclosure, medial diagnostics, etc.
- Active feature acquisition: allows the learner to request more complete feature information
  - features can be obtained at a cost, e.g. running additional diagnostics, etc.
  - goal: select more informative features to obtain during training
- Active classification: missing features may be available during test time rather than training time
- Active Class Selection[Lomasky et al., 2007]
  - learner to allowed to query a known class label, obtaining each instance incurs a cost

Active Clustering

- based on expected value of information criterion
- generate the unlabeled instances in such a way that they self-organize into groupings with less overlap or noise than for clusters induced using random sampling [Hofmann and Buhmann, 1988]
- Can work with constraints:
    - two instances must belong to the same cluster
    - two instances cannot belong to the same cluster
    - [Grira et al., 2005]
    - [Andrzejewski et al., 2009] for topic modeling

Section 6

Practical Considerations

Batch-mode

- usually queries are selected in serial
- allowing the learner to query instances in groups
  - distributed, parallel environment
  - models with slow learning procedure
- Q-best queries often does not work well: overlap in information content among them
  - encouraging diversity in batch [Brinker, 2003; Xu et al., 2007], usually using greedy heuristics
  - extension of Fisher information with sub-modular functions [Hoi et al., 2006b]
  - as a discriminative optimization, and try to make the most informative batch[Gou and Schuurmans, 2008]

More Practical Considerations

- Noisy Oracles:
  - "crowdsourcing" labeling: non-expert oracle
  - selective repeated labeling
- Variable labeling costs:
  - using current trained model to assist in the labeling of query instances (pre-labeling)
  - explicitly accounting for varying label costs e.g. "robot scientist" example[King et al., 2004] considers cost of materials
- Alternative query types:
  - instances grouped into bags, and the bags are labeled
  - query on features rather than instances

More Practical Considerations

- Multi-task active learning:
  - alternating selection/rank combination
  - taking mutual information among labels for dependent tasks
- Changing model classes
  - random sampling may be better
- Stopping Criteria
  - can be set theoretically, but usually it stops early due to economic or other external factors

Section 7

Related Areas

Semi-Supervised Learning

- common: making the most out of unlabeled data
- self-traing[Yarowsky, 1995]: adds most confident unlabeled instances to training set
- co-training and multi-view training: uses ensemble mehthods as in query-by-committee
- same problem from opposite directions
  - SSL: exploit what the learner thinks it knows about the unlabeled data
  - AL: attempt to explore the unknown aspects

Reinforcement Leraning

- learner must be proactive in order to perform well.
- exploration-exploitation tradeoff
- active learning of relocation of state to reduce number of actions required to find optimal policy in Q-learning [Mihalkova and Mooney, 2006]
  - When:
    - agent is in trouble: decreasig Q-values
    - agent is bored: change in Q-values are small
  - Where:
    - should be likely to be encountered while following an optimal policy
    - agent is uncertain about the best action

Reinforcement Learning (2)

[Hsu and Lin, 2015] Active learning by learning:

- Interpret k-learner system as a multi-armed bandit
- multi-armed bandit
- a gambler is given K bandit machines, a budget of T iterations
- the gambler sequentially decides which machine to pull in each iteraton
- the bandit machine randomly provides a reward from a machine-specific distribution unknown to the gampler
- goal: to maximize the total rewards earned through the sequence of decisions
- trade-off between exploitation and exploration
- analogy: bandit machine - selection algorithm
- careful selection of bandit method and reward scheme is needed
    - Exp4.P: performance guarantee on adversarial settings [Beygelzimer et al., 2011]
    - Importance-weight accuracy for reward