CSCE990
FEBRUARY 22 & 24, 2011

Active Learning Summary and Applications

L. Dee Miller

Overview

2

- Machine Learning (ML) Review and Connections (Feb. 22)
- Active Learning (AL) summary (Feb. 22 & 24)
- AL application papers (Feb. 24)
- Discussion relevant to MAS (Anytime)

2

ML Review

3

- Datasets
- Taxonomy
- Supervised Learning (SL)

ML Review—Dataset 1

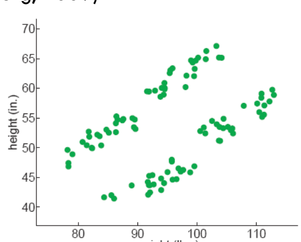
4

- Dataset consists of set of instances
- An instance (i.e., data point) consists of *D*-dimensional feature vector (x)
- Features (i.e., attributes) can be numeric or discrete values
- An instance may have a desired prediction or label (y)
- Assumption: instances used for training are sampled independently from underlying distribution

ML Review—Dataset 2

5

- Example dataset "Little Green Men" (Zhu & Goldberg, 2009)



ML Review—Taxonomy

6

- Supervised learning (focus)
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

ML Review—SL 1

7

- Uses training sample of instances **with** labels
- Common Tasks:
  - Regression
  - Classification
    - Train a function (i.e., classifier) to predict the correct label for unknown data points from the same joint probability distribution as the training sample
    - Function divides feature space into decision regions where instances share the same label

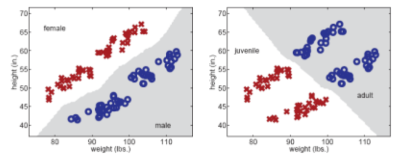
ML Review—SL 2

8

- K-Nearest-Neighbor Classifier

Input: Training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$*; distance function* $d()$*; number of neighbors* k*; test instance* $\mathbf{x}^*$

1. *Find the* k *training instances* $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$ *closest to* $\mathbf{x}^*$ *under distance* $d()$.
2. *Output* y^* *as the majority class of* $y_{i_1}, \ldots, y_{i_k}$. *Break ties randomly.*

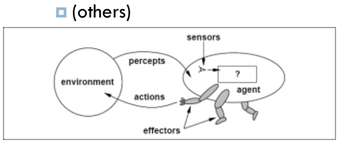


(a) classification by gender (b) classification by age

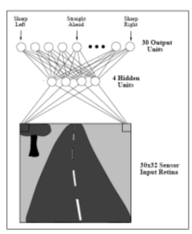
Connections 1

9

- Artificial Intelligence (AI) Areas
  - Machine Learning (ML)
  - Multi-Agent Systems (MAS)
  - (others)



Intelligent Agent (Russell & Norvig, 1995)

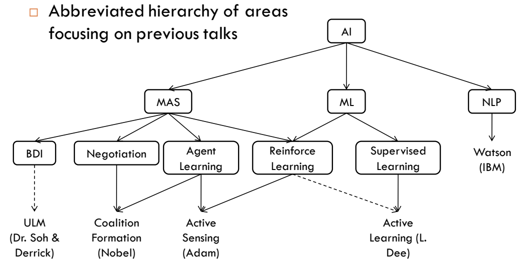


ALVINN (Mitchell, 1997)

Connections 2

10

- Abbreviated hierarchy of areas focusing on previous talks



Connections 3

11

- Addressing "AI factionalism" not primary focus!
- Similarity between intelligent agents and SL
- Review of environment characteristics (Russell & Norvig, 1995)
  - Accessible ➡ sensors have complete access
  - Deterministic ➡ next state from current state & action
  - Episodic ➡ experience divided into episodes
  - Static ➡ environment does not change during deliberation
  - Discrete ➡ limited number of actions

Connections 4

12

- Environment characteristics for SL
  - Accessible ➡ Yes, dataset has all relevant features
  - Deterministic ➡ No, label for next point does not depend on current
  - Episodic ➡ Yes, predicts labels individually
  - Static ➡ Yes, "concept" for dataset does not change
  - Discrete ➡ Yes, labels are fixed

Environment	Chess	Poker	SL
Accessible	Yes	No	Yes
Deterministic	Yes	No	No
Episodic	No	No	Yes
Static	Yes	Yes	Yes
Discrete	Yes	Yes	Yes

AL Summary

13

- Overview
- Strategies
- Interestingness Measures
- Analysis (Empirical/Theoretical)
- Problem Variants
- Practical Considerations

AL Summary—Overview 1

14

- Also called query learning or optimal experimental design (in statistics)
- Problem: Instances are cheap but labels may be expensive
  - Speech recognition ➡ annotation is time consuming and requires trained linguists
  - Information extraction ➡ trained using documents with detailed annotations
  - Classification/filtering requires ➡ users must provide many annotations

AL Summary—Overview 2

15

- Solution:
  - Choose instances for oracle to label
  - Re-learn the model (i.e., function)

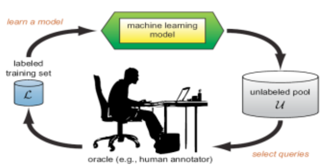


Figure 1: The pool-based active learning cycle.

AL Summary—Strategies 1

16

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

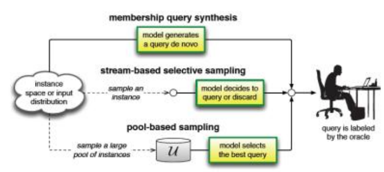


Figure 4: Diagram illustrating the three main active learning scenarios.

AL Summary—Strategies 2

17

- Membership query synthesis
  - Requests labels for synthesized queries (i.e., points) created de novo (i.e., anew)
  - Can automatically discover interesting experiments (e.g., mutant yeast)
  - May find queries which are not meaningful to the human annotator

AL Summary—Strategies 3

18

- Stream-based selective sampling
  - Assumes obtaining an unlabeled instance is free
  - Decides whether or not to query for the label
    - Use informativeness measure or query strategy (examples given later)
    - Compute region of uncertainty still ambiguous to the learner (Boundary of Use)
  - Useful when memory or processing power may be limited

AL Summary—Strategies 4

19

- Pool-based sampling
  - Assume large amount of unlabeled instances together with small amount of labeled instances
  - Query in greedy fashion based on informativeness measure
  - Ranks all instances together (more common) rather than sequentially as in stream-based

AL Summary—Measures 1

20

- Uncertainty sampling
  - Query the instance the learner is least certain how to label
    - Least confident $x^*_{LC} = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x),$
    - Margin sampling $x^*_M = \underset{x}{\operatorname{argmin}}\ P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x),$
    - Entropy $x^*_H = \underset{x}{\operatorname{argmax}} -\sum_i P_\theta(y_i|x) \log P_\theta(y_i|x),$
  - None of the three are "best", but entropy minimizes log-loss while the other two reduce classification error

AL Summary—Measures 2

21

- Query-by-committee
  - Maintain a committee of models (i.e., ensemble of classifiers)
    - Construct a small committee of models (e.g., HMM, boosting, bagging)
    - Use entropy or Kullback-Leibler divergence to measure consensus of committee
  - Informativeness measured using disagreement

AL Summary—Measures 3

22

- Expected model change
  - Select instance that would cause the greatest change to current model if label was known
    - Can measure for any function using gradient decent (e.g., artificial neural networks) by measuring change in the weights
  - Choosing point can be computational expensive if set of features and labels is large
    - Genetic Algorithm Classifier System

AL Summary—Measures 4

23

- Expected error reduction
  - Select instance that would minimize the generalization error in current model
  - Choosing this point can be computationally expensive because the function must be re-trained after labeling each point
    - Approximate over all possible labels with the current model!

AL Summary—Measures 5

24

- Variance reduction
  - Reduce the generalization error indirectly by minimizing variance
  - For gradient descent methods, we can reduce the variance by using the Fisher information matrix. Optimizing on this matrix can be tricky and there are several strategies:
    - A-optimality minimize the trace of the inverse matrix (most common)
    - D-optimality minimize the determinant of the inverse matrix
    - E-optimality minimize the max eigenvalue of the inverse matrix
  - Classifier does <u>not</u> need to be retrained

AL Summary—Measures 6

25

- Density-weighted methods
  - Previous methods are vulnerable to outliers
  - Want to find query points which are representative of underlying distribution
    - Weight uncertainty metric by similarity of instance to other instances in training data
    - Use a density method which clusters the unlabeled instances and query cluster centroid

AL Summary—Analysis 1

26

- Empirical
  - Majority of papers say AL reduces number of labeled instances need to achieve desired accuracy
  - However, training data created is biased towards function rather than underlying distribution
  - AL sometimes requires more labels to do well than passive and/or do worse than random sampling

AL Summary—Analysis 2

27

- Theoretical (limited advances)
  - Some work on how many random labeled instances are needed to achieve the maximum desired error rate for pool-based AL
  - Pool-based AL with linear classifiers shown to have worst-case performance equivalent to supervised learning
  - Theoretical frameworks are <u>not extendible</u> to all SL algorithms

AL Summary—Variants 1

28

- AL for structured outputs
  - Extension to probabilistic finite state machines (HMMs, context-free grammars, etc.)
- Active feature acquisition
  - Extension to request missing feature data
  - Goal: select most informative features (e.g., budgeted learning)

AL Summary—Variants 2

29

- Active class selection
  - Assumes labels are freely available but there is cost associated with instances
  - Fairly new problem variant
- Active clustering
  - Extension to unsupervised clustering used to organize data into meaningful patterns
  - Goal: choose instances which self-organize into groups with less overlap (improve cluster assumption) ➡ semi-supervised clustering

AL Summary—Considerations 1

30

- Batch-mode active learning
  - Query instances in groups
  - Cannot simply select Q-best because of overlap—must consider "diversity" in Q-best
- Noisy oracles
  - Quality of the label could vary (e.g., crowd-sourcing)
  - Learner must decide whether to query label for new instance or re-query label for existing instance

AL Summary—Considerations 2

31

- Variable learning cost
  - Cost for labels could vary
  - Previous work generally assumes annotation costs are known and modify measure to balance annotation/misclassification cost
- Alternative query types
  - Instances are grouped into bags (e.g., bag: document, instances: passages)
  - Queries are made about bags rather than instances (higher level)

AL Summary—Considerations 3

32

- Multi-task AL
  - Instances could have multiple, correlated labels
  - Take into account mutual information among different labels
- Changing model classes
  - AL chooses instances biased towards classifier used which may reduce accuracy for others
  - Only a problem when classifier could change

AL Application

33

- Attenburg (2010) Why Label when you can search?
  - Active class selection
  - Noisy oracles (crowd-sourcing)
- Donmez (2008) Paired Sampling in Density Active Learning
  - Active clustering
  - Density-weighted method

AL Summary—Attenburg 1

34

- The authors are interested in safe advertising—deciding whether web pages contain questionable content (e.g., porn)
- Humans examining text for every page would be expensive. However, humans can examine some of the pages using crowd-sourcing
- SL can learn functions to decide, but accuracy depends on pages provided for training

AL Summary—Attenburg 2

35

- AL could be used to find training instances and use human to provide label
- Problem: Extreme class imbalance
  - Only a tiny fraction of pages contain questionable content (1/100)
  - Active learning rarely chooses any instances with positive labels resulting in class imbalance
  - SL systems do not learn well from training data with class imbalance (even distribution is best)

AL Summary—Attenburg 3

36

- Solution: use both AL and guided learning (i.e., active class selection)
  - Guided learning uses oracle to search for instances satisfying some criteria (e.g., instances with positive labels from questionable content)
  - Note: Guided learning subsumes AL and should have higher cost
    - Guided Learning: search + label
    - Active Learning: label

AL Summary—Attenburg 4

37

- AL Measures Used
  - Uncertainty sampling (query instance closest to decision boundary) ➡ best
  - Boosted disagreement query-by-committee (query instance with most disagreement)
  - Density sensitive pre-clustering (query instance nearest cluster centroid)

AL Summary—Attenburg 5

38

- Guided learning (simulated)
  - Previous work on SL shows even split in training data generally gives highest test accuracy
  - Therefore, guided learning should request new instances with even split of labels
  - Authors simulate guided learning using equal sampling technique on dataset
    - Points are sampled equally and u.a.r. from bins with different labels until minority bin is empty

AL Summary—Attenburg 6

39

- Experimental Setup
  - Dataset from Open Directory project containing 4,000,000 urls
  - Uses logistical regression model (i.e., linear classifier)
    - SL is efficient during training which is important for large datasets
    - Smaller-scale experiments (i.e., sanity-checks) show benefits of approach are independent of SL used
    - All experiments use receiver operating characteristic curve important for class imbalance

AL Summary—Attenburg 7

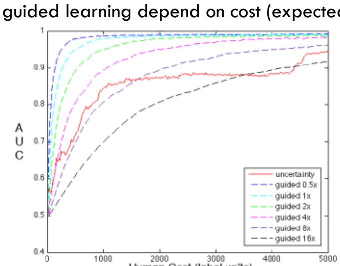
40

- Results
  - Searching for instances with balanced label proportion gives better results without AL (trivial)
  - Natural clusters of instances which are strongly misclassified but not high priority for exploration
    - Density-sensitive AL does not work well when concepts are disjunctive (i.e., label dispersion)

AL Summary—Attenburg 8

41

- Results (cont.)
  - Benefits of guided learning depend on cost (expected)



AL Summary—Donmez 1

42

- The authors are interested in AL for:
  - Balanced sampling on both sides of decision boundary (overcome cold-start problem)
  - Exploiting natural clustering of instances
- Analogy: easier to obtain geological data on regions with/without oil than to drill multiple test holes

AL Summary—Donmez 2

43

- Density-sensitive distance estimation
  - Assumes decision boundary lies in low density region (i.e., cluster assumption)
  - Clusters using fully-connected graph with edge weights from Euclidean distance
  - Density-sensitive distance based on longest distance edge
    - Can give poor results when two points are connected by a long path of short edges
    - Need to balance inter-cluster and intra-cluster distance ➡ use multi-dimensional scaling

AL Summary—Donmez 3

44

- Density-sensitive paired sampling
  - Uses logistical regression model (i.e., linear classifier)
  - Pairs of points sampled with opposite labels and high uncertainty
  - Also consider points in high density regions to increase confidence in labels for neighbors

AL Summary—Donmez 4

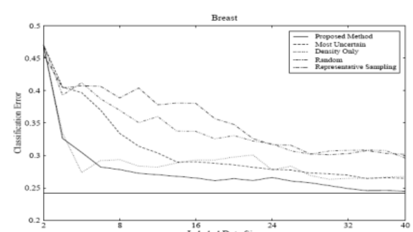
45

- Overall: Balance density estimate with uncertainty from SL
  - Avoid querying labels for points in "successful" regions where SL has high confidence
- Function used is quite convoluted, but favors pairs of points from large neighborhoods which have different (i.e., uncertain) labels

AL Summary—Donmez 5

46

- Results
  - Balance gives better results than individual AL measures



AL Conclusions

47

- Questions
  - What areas of MAS can benefit from AL?
  - Which strategies, measures, etc. should we use for MAS?
  - Who is going to win the epic badminton match?

References

48

- Attenberg, J., & Provost, F. Why Label when you can Search?, *ACM SIGKDD*, pp. 423-432, 2010.
- Donmez, P., & Carbonell, J. Paired-Sampling in Density-Sensitive Active Learning. *Proceedings of the 10th International Symposium on Artificial Intelligence and Mathematics*, 2008.
- Mitchell, T., *Machine Learning*, McGraw-Hill, 1997.
- Russell and S., Norvig, P., *Artificial Intelligence A Modern Approach*, Prentice Hall, 1995.
- Settles, B., Active Learning Literature Survey, *University of Wisconsin—Madison*, No. 1648, 2010.
- Zhu, X. and Goldberg, A.B., *Introduction to Semi-Supervised Learning*, Morgan & Claypool Publishers, 2009