Comp 135 Introduction to Machine Learning and Data Mining

Fall 2016

Professor: Roni Khardon
Computer Science
Tufts University

Contributes from prior slides by Carla Brodley, Kevin Small and Byron Wallace and material from the survey by Burr Settles gratefully acknowledged.

Supervised Learning

- All algorithms so far (kNN, DT, Nbayes, Percptron, …, SVM) are passive
- Input: is set of examples and labels
- Output: is classifier.
- What if the learner can be active?
- What does that mean?

Focusing on Labels

- Often labels are expensive because they require work/time of a domain expert
- or scarce/unavailable because we do not have access to "ground truth"
- Recent methods try to use crowd sourcing to obtain labels – this is related to our topic but out of scope for the course

Focusing on Labels

- Unsupervised - No labels
- Semi-supervised - Some Labels
- Active - Ask for Labels

  When? How? What is reasonable/permissible?
- Supervised - All labels

Active Learning Variants

- Where/how can the leaner ask for labels?
- We will discuss 3 settings.
- Query Learning
- Stream based action learning
- Pool based active learning

Query Learning

- Various types of questions possible:
- In Classification
  - Equivalence queries
  - Subset queries
  - Membership queries
- In preference elicitation:
  - Ranking queries
- …

Membership query Learning

- Consider learning a discrete threshold function (x>= n) where n is an integer in the range [-1000,1000]
- Learning algorithm using membership queries?
- Naïve query strategy: …
- Smart query strategy: …

Membership query Learning

- Consider learning a discrete 2D rectangle where x,y boundaries are integers in the range [-1000,1000]
- Learning algorithm using membership queries?
- Naïve query strategy: …
- Smart query strategy: …
  This is much harder

Membership query Learning

- General principle? Ask about an example which will shrink the range of options to the largest extent.
- Since the answers are not known in advance: ask about an example which will shrink the range of options for any possible answer.
- In noise free case: the version space is the set of hypotheses that are consistent with current data
- Seek to shrink the version space

Notable Query Learning Systems

- Program synthesis/debugging [S82]: developed methods for interactive construction of Prolog programs via equivalence queries (and counter examples) and membership (I/O) queries.
- Robot scientist [K04,K09]: automate process of investigation re which genes (in yeast) encode certain enzymes.

Program debugging

$$isort([X|Xs],Ys) \leftarrow isort(Xs,Zs),\ insert(X,Zs,Ys).$$
$$isort([\,],[\,]).$$

$$insert(X,[Y|Ys],[Y|Zs]) \leftarrow Y>X,\ insert(X,Ys,Zs).$$
$$insert(X,[Y|Ys],[X,Y|Ys]) \leftarrow X \leq Y.$$
$$insert(X,[\,],[X]).$$

We first test *isort* on [2,1,3],

| ?– isort([2,1,3],X).

$X = [2,3,1]$

Example/Image from [Shapiro 1982]

Program debugging

| ?– fp(isort([2,1,3],[2,3,1]),C).

query: *isort*([],[])? y.

query: *insert*(3,[],[3])? y.

query: *isort*([3],[3])? y.

query: *insert*(1,[],[1])? y.

query: *insert*(1,[3],[3,1])? n.

$C = insert(1,[3],[3,1]) \leftarrow 3>1,\ insert(1,[\,],[1])))$

yes

fp returned a false instance of the first clause of *insert*. Examining it shows that the arguments for the > test are exchanged. We fix that bug, and try *isort* again,

Example/Image from [Shapiro 1982]

The Robot Scientist

- Automatically design yeast growth experiments
- Implement in hardware
- Observe results, and continue
- 1000 combinations a day
- Goal: identify which genes encode specific enzymes/functions
- Novel discoveries

Example/Image from [King et al 2009]

comp135 Roni Khardon, Tufts University

Membership query Learning

- Hard to apply in practice
- What does a membership query mean when learning to classify molecules?
- What does a membership query mean when learning to classify images (e.g. character recognition)?

comp135 Roni Khardon, Tufts University

Query Learning: summary

- Main idea: try to reduce uncertainty regarding possible true concepts
- Impressive results in some cases
- Hard to apply in general
- Mainly because of the need to construct meaningful examples or questions

Stream based and Pool based active learning avoid this difficulty by allowing questions only on examples that already exist

comp135 Roni Khardon, Tufts University

Pool Based Active Learning

- Assume that a set of unlabeled examples is given
- Learner can ask for labels of examples in this set
- A simple baseline will draw a random subset of examples and ask for their labels.
- Can we do better?

comp135 Roni Khardon, Tufts University

Pool Based Active Learning

1) Obtain initial classifier
2) While expert is willing to label
   a) Make predictions with current classifier
   b) Identify "useful instance(s)"
   c) Request labels for "useful instances"
   d) Retrain and Goto 2

- How to select instances in step 2?
  - Classifier specific schemes?
  - Classifier independent schemes?

comp135 Roni Khardon, Tufts University

Pool Based Active Learning

- Uncertainty sampling:
- Instance selected for which current hypothesis is least confident in its prediction
  - Uses a single hypothesis to determine uncertainty
  - How do we define uncertainty for k-NN, decision trees, N Bayes, SVM?

comp135 Roni Khardon, Tufts University

Pool Based Active Learning

- If we have or can approximate a probabilistic prediction and label probabilities are (from max to min)

  $p_1, p_2, \dots, p_k$

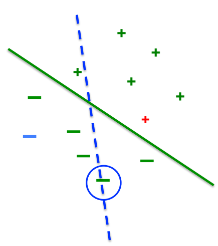
- Instance with smallest p_1
- Instance with smallest $(p_1 - p_2)$
- Instance with largest entropy$(p_1, p_2, \dots, p_k)$

(same for 2-class but different in multi-class)

Pool Based Active Learning

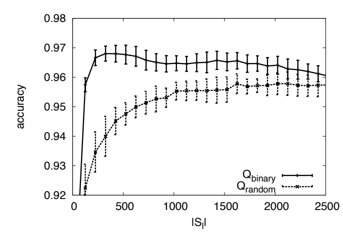
- For SVM: all separators of current labeled examples are in version space
- Least confident close to Max Margin separator → ask about examples with smallest margin

Example



Experiments on **ad** dataset

- Distinguish Advertisement Images from Non-Ads (UCI)
- 3279 instances, 458 ads, $\{0,1\}^{1555}$ input, $|S_l(0)|=25$, $|S_{select}|=5$
- Used Averaged Perceptron with Margin

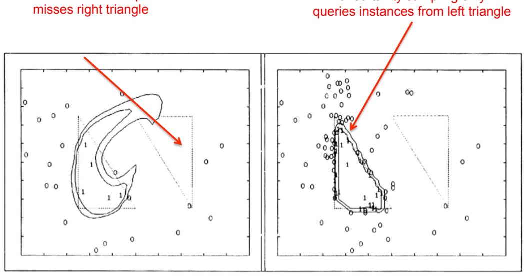


[Small 2010]

Pathological Example for Uncertainty Sampling

initial random sample misses right triangle

uncertainty sampling only queries instances from left triangle



[Cohn, Atlas & Ladner 94]

Pool Based Active Learning

- Uncertainty sampling
  - Easy to implement in many algorithms
  - Successful in many cases
  - Can be trapped into missing important distinctions
  - Decisions based on one/current hypothesis and may not reflect diversity in version space

Pool/Stream based Active Learning

- Ideally take a vote among all the hypotheses in the version space
- Picking a point with high disagreement will reduce the version space regardless of what the true label is!
- and will reduce uncertainty
- But this is almost never feasible
- Alternative: form a "committee"

Query by Committee

- Ideally take a vote among all the hypotheses in the version space
- Instead take a vote among a few selected hypotheses – the committee
- Pick an example that maximizes disagreement
- How to pick committee? And its size?
- Works for Pool based or stream based

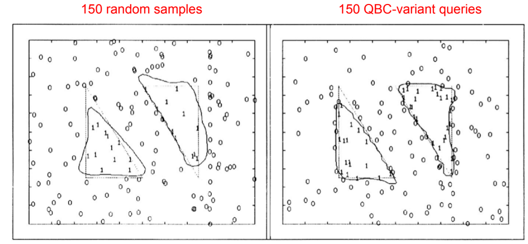
Stream Based Active Learning

- Query by Committee

1) Calculate initial hypothesis space H_1 and prior π
2) For t = 1,…,T
   a) Retrieve unlabeled instance x_t
   b) Draw $h_1, h_2 \sim (\pi, H_t)$
   c) If $h_1(x_t) \neq h_2(x_t)$, request label from expert
   d) Generate H_{t+1} and recalculate π

- Here committee size is 2
- And we are assuming ability to sample from "posterior" over hypotheses

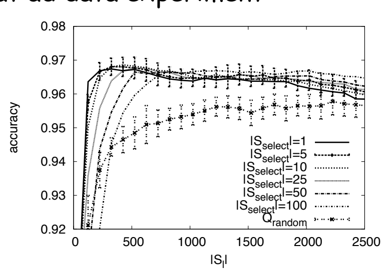
Query by Committee



[Cohn, Atlas & Ladner 94]

Size of Query Batch

- Repeat ad data experiment



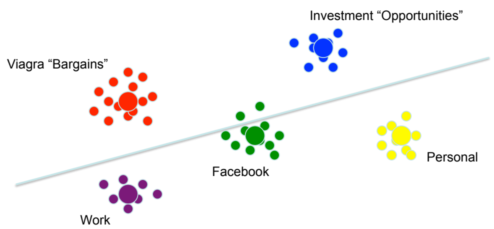
- Small size works well

[Small 2010]

Size of Query Batch

- Small size often works well
- But not practical in many cases where we want to send a batch of questions
- How should the batch be chosen?

Active Learning using Pre-clustering



- If data clusters, we only require a few representative instances from each cluster to label data

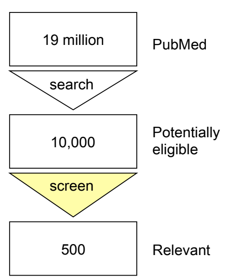
[Ngyuen & Smeulders 2004]

Systematic Reviews [WSBT2011]

- **Systematic review:** an exhaustive assessment of all the published medical evidence regarding a precise clinical question
- e.g., "Is aspirin better than leeches in inducing more than 50% relief in patients with tension headaches?"
- ***Must find all relevant studies***

Systematic Reviews [WSBT2011]

- The problem



Systematic Reviews [WSBT2011]

Doctors read these. They'd rather be doing something else.

Citation Screening [WSBT2011]

- An additional challenge of performing active learning when the class distribution is very skewed (very few positive instance)
- Must get 100% recall
- …
- Reliably reduce labeling effort by X%
- System deployed and used in practice

Active Learning

- Query learning allows to synthesize query examples
- Active Learning focus on existing examples
- Pool based/stream based selection
- Uncertainty sampling is simple and effective – based on the current hypothesis
- Query by Committee uses disagreement amongst an ensemble of classifiers
- Several refinements and improvements
- Margin based selection with linear separators a popular approach