主动学习：奇特的AI算法

ACTIVE LEARNING:CURIOIS AI ALGORITHMS

胡伟龙
huweilong@whu.edu.cn

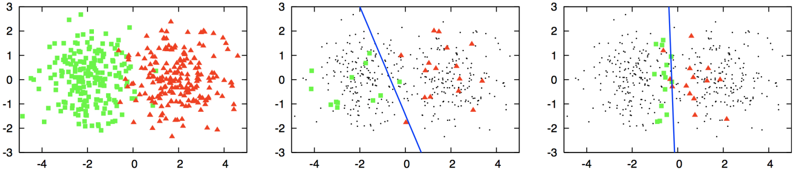
2018年6月1日

武汉大学 WUHAN UNIVERSITY | 计算机学院 Computer School of Wuhan University

Contents

INTRODUCTION

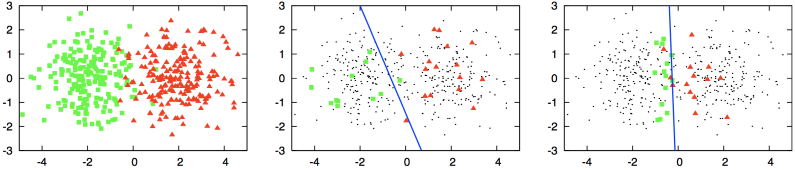
Motivation



Explanation:

- Time consuming, e.g., document classification.
- Expensive, e.g.,medical decision (need doctors).
- Sometimes dangerous, e.g., landmine detection.

Motivation



Explanation:

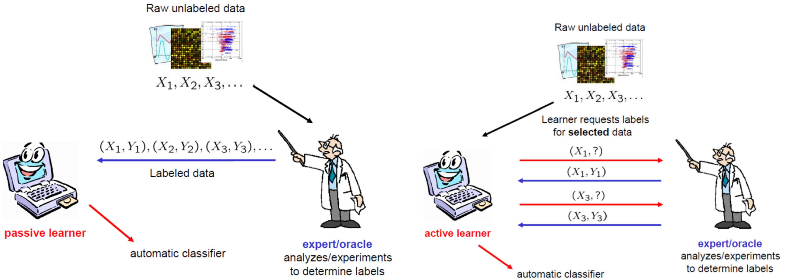
- Do not know the labels(red or green) and it's expensive to find the labels.
- Poor selection of data points for logistic regression.
- Select superior data points to create a good decision boundary.

Modern Research

- Use deep learning algorithms like CNNs and LSTMS as the learner and improve their efficiency when using active learning frameworks.(Kronrod and Anandkumar, 2017; Sener and Savarese, 2017)
- Implemente Generative Adversarial Networks (GANs) into the active learning framework.(Zhu and Bento, 2017)
- Reframe active learning as a reinforcement learning problem.(Fang et. al, 2017)
- Learn active learning strategies via a meta-learning setting.(Fang et. al, 2017)

Active Learning vs. Semi-supervised Learning

- The same goal:
  - Attain good learning performance without demanding too many labeld examples.
- Different approaches
  - Semi-supervised learning: use unlabeled data
  - Active learning: choose labeled examples



DEFINATION AND CONCEPTS

Passive Learning & Active Learning

HYPOTHESIS: Choosing superior data can surpass traditional methods with substantially less data for training.

PASSIVE LEARNING: Gather a large amount of data randomly sampled from the underlying distribution and use this dataset to train a model.

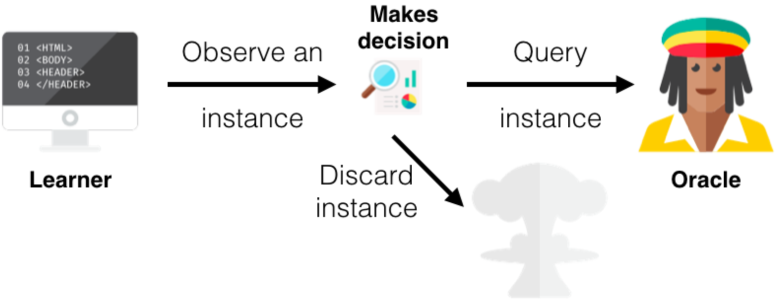
Certain Criteria on Studying Pancreatic Cancer

- If the patient drinks alcohol and is over 40 years.
- If the patient is over 50 years old.

Scenarios Membership Query Synthesis

- The learner generates an instance from some underlying natural distributions.
- The instance is sent to the oracle to label.

Scenarios Stream-Based Selective Sampling



- Get an unlabelled instance is free(assumption).
- Determine whether needs to be labelled or discarded.
- To determine informativeness of the the instance, you use a QUERY STRATEGY.

Scenarios Pool-Based Sampling

- A large pool of unlabelled data.
- The most informative instance(s) are selected based on some informativeness measure.

Scenarios Pool-Based Sampling

- A large pool of unlabelled data.
- The most informative instance(s) are selected based on some informativeness measure.

How to get informativeness of an instance?

QUERY STRATEGIES

Query Strategies

DIFFERENCE:The ability to query instances based upon past queries and the responses (labels) from those queries.

Common Query Strategies

- UNCERTAINTY SAMPLING
- QUERY-BY-COMMITTEE(QBC)
- Expected Model Change
- Expected Error Reduction
- Variance Reduction
- Density-Weighted Methods

Uncertainty Sampling

Instances	Label A	Label B	Label C
d_1	0.9	0.09	0.01
d_2	0.2	0.5	0.3

LEAST CONFIDENCE: Selects the instance for which it has the least confidence in its most likely label.

- The leaner is pretty confident about the label for d_1.
- d_2's probabilities are more spread, so less confident.

SHORTCOMING: Only consider the most probale label and disregard the others.

Uncertainty Sampling

Instances	Label A	Label B	Label C
d_1	0.9	0.09	0.01
d_2	0.2	0.5	0.3

MARGIN SAMPLING: Selecte the instance that has the smallest difference between the first and second most probable labels.

- d_1's difference is 0.81(0.9-0.09).
- d_2's difference is 0.2(0.5-0.3). Hence, select d_2.

SHORTCOMING: Still did not consider all possible label probabilities.

Uncertainty Sampling

Instances	Label A	Label B	Label C
d_1	0.9	0.09	0.01
d_2	0.2	0.5	0.3

Entropy Sampling: Selecte the instance with the largest entropy.

- d_1 has a value of 0.155.
- d_2 has a value of 0.447. Hence, select d_2 again.

Uncertainty Sampling

Instances	Label A	Label B	Label C
d_1	0.9	0.09	0.01
d_2	0.2	0.5	0.3

ENTROPY SAMPLING: Selecte the instance with the largest entropy.

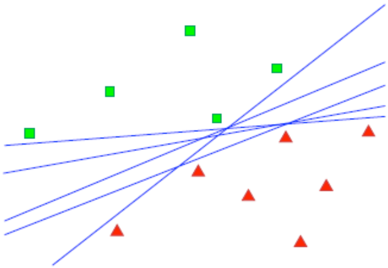
- d_1 has a value of 0.155.
- d_2 has a value of 0.447. Hence, select d_2 again.

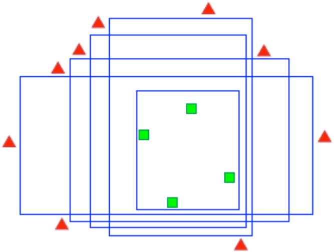
ANOTHER PROBLEM

How to get probabilities of all label?

Query-By-Committee

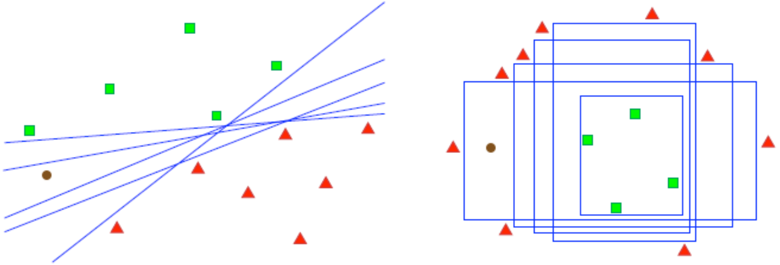
Basic idea: A committee $\mathscr{C} = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$ of models are trained on the labeled set $\mathscr{L}$, but represent competing hypotheses. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree.





Query-By-Committee

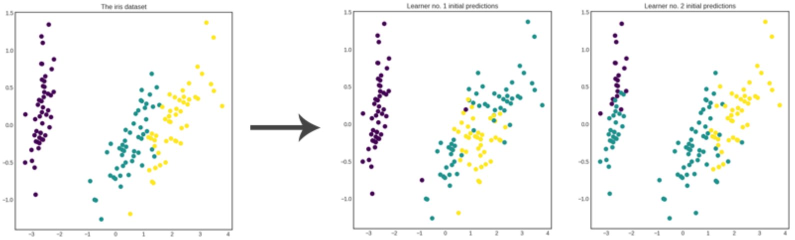
Basic idea: A committee $\mathscr{C} = \{\theta^{(1)}, \ldots, \theta^{(C)}\}$ of models are trained on the labeled set $\mathscr{L}$, but represent competing hypotheses. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree.



Query-By-Committee

In order to implement a QBC selection algorithm:

- Be able to construct a committee of models that represent different regions of the version space.
- Have some measure of disagreement among committee members.



Query-By-Committee

For measuring the level of disagreement:

- Vote entropy.
- Kullback-Leibler (KL) divergence.

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

$V(y_i)$ is the number of "votes" of a label, C is the committee size.

Query-By-Committee

For measuring the level of disagreement:

- Vote entropy.
- Kullback-Leibler (KL) divergence.

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

where:

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

Here $\theta^{(c)}$represents a particular model in the committee and $P_C(y_i|x) = \sum_{c=1}^{C} P_{\theta(c)}(y_i|x)$ is the "consensus" probability that y_i is the correct label.

Query-By-Committee

- Query by committee

Query-By-Committee

- Query by committee
  - Keep a committee of classifiers

Query-By-Committee

- Query by committee
  - Keep a committee of classifiers
  - Query the instance that the committee members disagree

Query-By-Committee

- Query by committee
  - Keep a committee of classifiers
  - Query the instance that the committee members disagree
- QBC as version space reduction

Query-By-Committee

- Query by committee
  - Keep a committee of classifiers
  - Query the instance that the committee members disagree
- QBC as version space reduction
  - Committee is an approximation to the version space

Query-By-Committee

- Query by committee
  - Keep a committee of classifiers
  - Query the instance that the committee members disagree
- QBC as version space reduction
  - Committee is an approximation to the version space
- QBC as uncertainty sampling

Query-By-Committee

- Query by committee
  - Keep a committee of classifiers
  - Query the instance that the committee members disagree
- QBC as version space reduction
  - Committee is an approximation to the version space
- QBC as uncertainty sampling
  - Use committee members to measure the uncertainty

AN EXAMPLE

An Example

Gather Data

- Ensure that the dataset you gather is representative of the true distribution.
- Impossible to have a totally representative sample.

Instances	Feature A	Feature B
d_1	10	0
d_2	4	9
d_3	8	5
d_4	3	3
d_5	5	5

An Example

STEP1: Split into Seed and Unlabelled Dataset

- Label a small part of the dataset as seed.
- Typically, a fully labelled dataset is used

Instances	**Feature A**	**Feature B**	**Label**
d_1	10	0	Y
d_2	4	9	-
d_3	8	5	N
d_4	3	3	-
d_5	5	5	-

An Example

STEP2: Train a learner

- Use the seed to train a learner.
- Use leaners that give a probabilistic response to whether an instance has a particular label.

Instances	Feature A	Feature B	Label
d_1	10	0	Y
d_2	4	9	-
d_3	8	5	N
d_4	3	3	-
d_5	5	5	-

An Example

STEP3: Choose unlabelled instances

- Determine the type of scenario.
- Determine the query strategy.

Instances	Feature A	Feature B	Label
d_1	10	0	Y
d_2	4	9	Y
d_3	8	5	N
d_4	3	3	N
d_5	5	5	-

Use pool-based sampling with a batch size of 2. Query strategy is LC

An Example

STEP4: Stopping criteria

- The number of instances queried.
- The number of iterations of steps 2 and 3.
- After the performance does not improve significantly

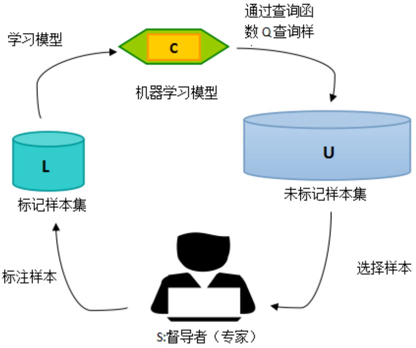
Instances	**Feature A**	**Feature B**	**Label**
d_1	10	0	Y
d_2	4	9	Y
d_3	8	5	N
d_4	3	3	N
d_5	5	5	-

Abstraction

Active Learning can be summarized as:

$$A = (C, Q, S, L, U)$$

where C denotes classifier(s), L denotes labeled data, Q denotes query function, S denotes experts, U represents unlabeled data.



Resources

modAL

A modular active learning framework for Python3

https://github.com/google/active-learning

http://burrsettles.com/pub/settles.activelearning.pdf