Active Learning and Semi-Supervised Learning

COMP90049

Semester 2, 2021

Qiuhong Ke, CIS

Acknowledgement: Tim Baldwin, Karin Verspoor & Kris Ehinger

Roadmap

So far:

- Supervised learning
- Unsupervised learning

Today:

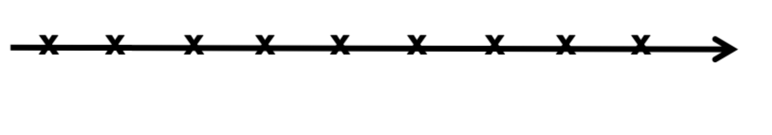
- Active learning
  - Query Scenarios
  - Query Strategies
- Semi-supervised learning
  - Combine unsupervised and supervised algorithm
  - Self-training

Active Learning

Active Learning

- Motivation: labelling is a finite resource, which should be expended in a way which optimises machine learning effectiveness.
- Key idea: the learner
  - has access to raw unlabeled data,
  - make a query about the label to an **oracle** (e.g. a human annotator)
- Goal: train a good classifier with reduced annotation cost.

Toy Example: 1D classifier



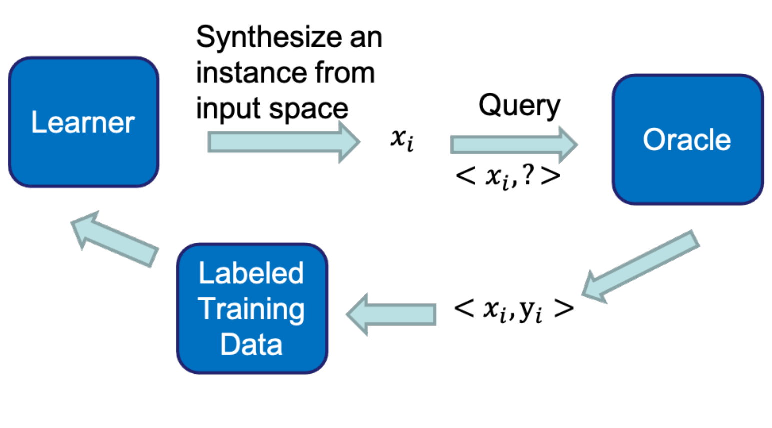
- Input: Unlabeled data, labels are all 0 then all 1 (left to right)
- Goal: find classifier (threshold function between 0 and 1)
- Naive method: annotate all data points
- Better method: use binary search to reduce annotations

Query Scenarios

- Query Synthesis
- Stream-based Sampling
- Pool-based Sampling

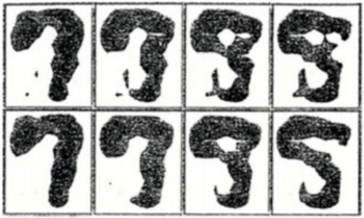
Query Synthesis I

Learner constructs an instance from input space or distribution from scratch



Query Synthesis II

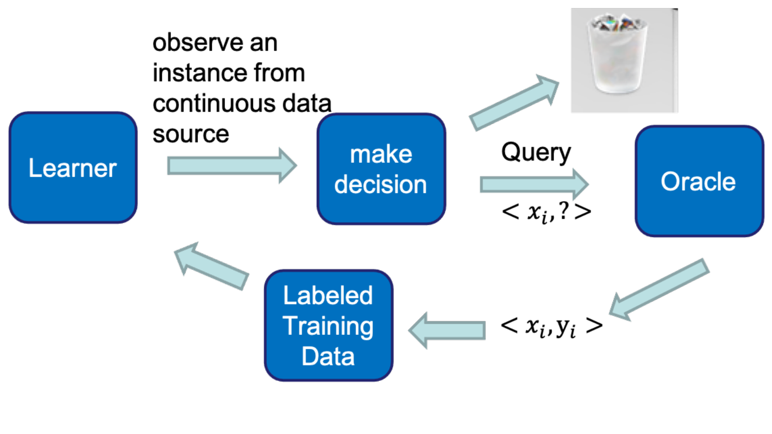
- Problem: Human annotator might not recognize the pseudo instance



Source: Kevin J. Lang and Eric B Baum. Query Learning Can Work Poorly when a Human Oracle is Used, 1992

Stream-based Sampling I

Learner decides query or ignore the observed instance from the continuous data source

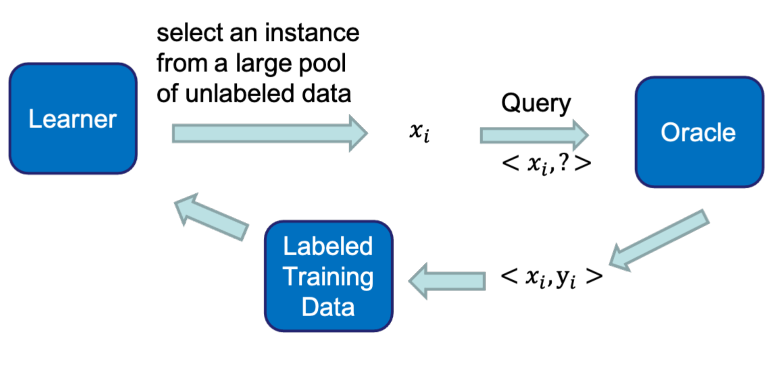


Stream-based Sampling II

- Query data from true distribution
- Useful if the dataset is too large to load
- Assumption: drawing an instance is less expensive than labeling, e.g., downloading video vs annotating actions

Pool-based Sampling I

Learner chooses the best instance from a large pool of unlabeled examples to query.



Stream-based vs Pool-based Sampling

- Stream-based: the learner observes one instance at a time and makes the decision individually.
- Pool-based: the learner observes whole dataset and choose the best one.

Query Strategies

- Uncertainty Sampling
- Query by Committee

Query Strategies I: Uncertainty Sampling

- Least Confident
- Margin Sampling
- Entropy Sampling

Uncertainty Sampling I: Least Confident

- Query instances where the classifier is least confident of the classification

$$x^* = \underset{x}{argmin}(P_\theta(\hat{y}|x))$$

$$where \qquad \hat{y} = \underset{y}{argmax}(P_\theta(y|x))$$

- Example: select instance 2 as the query

	y_1	y_2	y_3
Instance 1	0.01	0.9	0.09
Instance 2	0.5	0.3	0.2

Uncertainty Sampling II: Margin Sampling

- Selects queries where the classifier is least able to distinguish between the first and second most probable categories, e.g.:

$$x = \underset{x}{argmin}(P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x))$$

- where $\hat{y}_1$ and $\hat{y}_2$ are the first- and second-most-likely labels for x
- Example: select instance 2 as the query

	y_1	y_2	y_3
Instance 1	0.25	0.5	0.25
Instance 2	0.5	0.4	0.1

Example

- Which instance should be the query based on the strategy of least confidence?
- Which instance should be the query based on the strategy of margin sampling?

	y_1	y_2	y_3	y_4
Instance 1	0.2	0.4	0.2	0.2
Instance 2	0.5	0.35	0.1	0.05

Uncertainty Sampling III: Entropy Sampling

- Use entropy as an uncertainty measure to utilize all the possible class probabilities:

$$x = \underset{x}{argmax} - \sum_i P_\theta(\hat{y}_i|x) log_2 P_\theta(\hat{y}_i|x)$$

Applications & Other Uncertainty Sampling Methods

1. Speech Recognition
2. Machine Translation
3. Text Classification
4. Word Segmentation: classifier margin

1 Hakkani-Tür, Dilek, Giuseppe Riccardi, and Allen Gorin. ”Active learning for automatic speech recognition.” ICASP 2002.

2 Haffari, Gholamreza, Maxim Roy, and Anoop Sarkar. ”Active learning for statistical phrase-based machine translation.” ACL 2009.

3 Lewis, David D., and William A. Gale. ”A sequential algorithm for training text classifiers.” SIGIR 1994.

4 Sassano, Manabu. ”An empirical study of active learning with support vector machines forjapanese word segmentation.” ACL 2002.

THE UNIVERSITY OF MELBOURNE

Query Strategies II: Query by Committee

- Use multiple classifiers to predict on unlabelled data, and select instances with the highest disagreement between classifiers
- Assumes that all the classifiers learn something different, so can provide different information
- Disagreement can be measured by:
  - Vote entropy
  - KL divergence

Disagreement Measure I: Vote Entropy

Select instance with highest vote entropy for query:

$$x = \underset{x}{argmax} - \sum_{y_i} (\frac{V(y_i)}{N}) log_2 (\frac{V(y_i)}{N})$$

- $V(y_i)$: number of “votes” that label y_i receives.
- N: total number of “votes” (classifiers).

Example

$V(y_1) = 0, V(y_2) = 4, V(y_3) = 0$

$H = 0$

classifier	y_1	y_2	y_3
C_1	0.2	0.7	0.1
C_2	0.2	0.6	0.2
C_3	0.05	0.9	0.05
C_4	0.1	0.8	0.1

Example

$H = ?$

classifier	y_1	y_2	y_3
C_1	0.2	0.7	0.1
C_2	0.1	0.3	0.6
C_3	0.8	0.1	0.1
C_4	0.3	0.5	0.2

THE UNIVERSITY OF MELBOURNE

Disagreement Measure II: KL Divergence

$$x = \underset{x}{argmax} \frac{1}{N} \sum_{i=1}^{N} D(P_i || P_m)$$

- P_m: mean probability distribution of all the N models.
- Kullback Leibler (KL) divergence (relative entropy)

$$D(P_i||P_m) = -\sum_{j=1}^{n_c} P_i(j) \left[\log_2 P_m(j) - \log_2 P_i(j))\right] = \sum_{j=1}^{n_c} P_i(j) \log_2 \frac{P_i(j)}{P_m(j)}$$

  - $P_i = [P_i(1), P_i(2), \cdots, P_i(n_c)]$
  - $P_m = [P_m(1), P_m(2), \cdots, P_m(n_c)]$
  - $P_i(j)$: probability of the j^{th} class in the probability distribution P_i
  - $P_m(j)$: probability of the j^{th} class in the probability distribution P_m

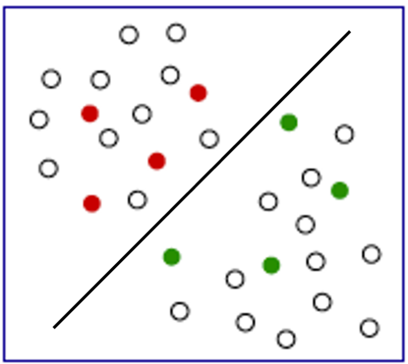
Semi-supervised learning

Semi-supervised learning

- Semi-supervised learning is learning from both labelled and unlabelled data
- **Semi-supervised classification**:
  - L is the set of labelled training instances $\{x_i, y_i\}_{i=1}^{l}$
  - U is the set of unlabelled training instances $\{x_i\}_{i=l+1}^{l+u}$
  - Often $u \gg l$
  - Goal: learn a better classifier from $L \cup U$ than is possible from L alone

Semi-Supervised Learning Approach I

- A simple approach: combine a supervised and unsupervised model
- e.g., Find clusters, choose a label for each (most common label?) and apply it to the unlabelled cluster members



Semi-Supervised Learning Approach II

Self-Training (Also known as "Bootstrapping")

- Assume you have $L = \{x_i, y_i\}_{i=1}^{l}$ labelled and $U = \{x_i\}_{i=l+1}^{l+u}$ unlabelled training instances
- Repeat
  - Supervised learning: Train a model f on L
  - Prediction: $y = f(U)$ to predict the labels on each instance in U
  - Identify a subset U' of U with "high confidence" labels
  - $L \leftarrow L \cup < U', f(U') >$
  - $U \leftarrow U \setminus U'$
  - Until L does not change

Active Learning vs Semi-supervised learning

- Same goal: reduce human annotation effort
- semi-supervised learning:
  - Learner produce labels automatically (e.g., on the data with high confidence)
- active learning:
  - Learner select unlabeled data (e.g., with low confidence/high uncertainty) to make a query
  - Oracle annotates the query

Summary

Summary

- What is active learning?
- What are the main sampling strategies in active learning?
- Outline a selection of query strategies in active learning.
- What is semi-supervised learning?
- What is self-training, and how does it operate?

References

- Burr Settles. Active learning literature survey. Technical report, Department of Computer Sciences, University of Wisconsin, Madison, 2010.
- Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison, 2005.
- Xiaojin Zhu. Tutorial on semi-supervised learning.
- Edith Law. Introduction to machine learning-active learning