Active Learning

Part of the Summerschool Data Science: Machine Learning with Python

Georg Krempl

g.m.krempl@uu.nl
Algorithmic Data Analysis Group
Information and Computing Sciences
Utrecht University, The Netherlands

Utrecht, July 26th 2024

Utrecht, July 26th 2024

Active Learning

Part of the Summerschool Data Science: Machine Learning with Python

Georg Krempl

g.m.krempl@uu.nl

Algorithmic Data Analysis Group
Information and Computing Sciences
Utrecht University, The Netherlands

Active Learning

Outline & Tentative Schedule

- 09:00 – 10:30 Lecture
  - Opening
  - Evaluating
  - Broadening the View
- 10:45 – 11:45 Practical
- 11:45 – 12:15 Discussion

Getting Ready

- Have scikit-learn and scitkit activeml installed:

```
pip install scikit-learn
pip install scikit-activeml
```

Outline & Tentative Schedule

- 09:00 – 10:30 Lecture
  - Opening
  - Evaluating
  - Broadening the View
- 10:45 – 11:45 Practical
- 11:45 – 12:15 Discussion

Getting Ready

- Have scikit-learn and scitkit activeml installed:

```
pip install scikit-learn
pip install scikit-activeml
```

Motivation: Exemplary Applications

Diagnosis Support System

- Objective: Clinical Image Classification
- Input: Images, ...
- Output: Class (e.g., benign vs malignant)
- Labelling requires medical expert, lab tests, ...

Brain Computer Interfaces / Intelligent Prosthesis

- Objective: Predict the action the user desires
- Input: Sensors / EEG patterns
- Output: Desired action
- (Re-)Calibration is tedious

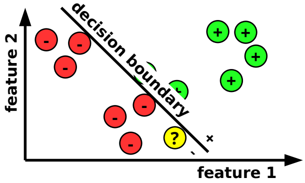
Motivation: (Supervised) Machine Learning

order value	product	...	default?
2.2	A	...	no
1.7	A	...	yes
1.3	B	...	yes
3.1	A	...	no
...	...	...	...

instances

features / explanatory variables

class label / dependent var.



Supervised Learning

$$f : \underbrace{x}_{\textit{feature}} \rightarrow \underbrace{y}_{\textit{label}}$$

- **Collect training data** from previous customers
- **Train and test on that data**
- **Deploy in production** predictions on new customers

Motivation: Data / Supervision Challenge

- Key to successful supervised models:

  Sufficient high-quality labelled training data

- Labelling often requires querying **oracles**, e.g.,
  - human domain experts
  - tedious-to-perform experiments
  - expensive-to-acquire third-party data

- How to build an equally good model with less data?

Active Learning: When & Why?

Motivation

- lot's of (automatically) generated data, but
- (human) annotation capacities remain limited

Context of Active Learning

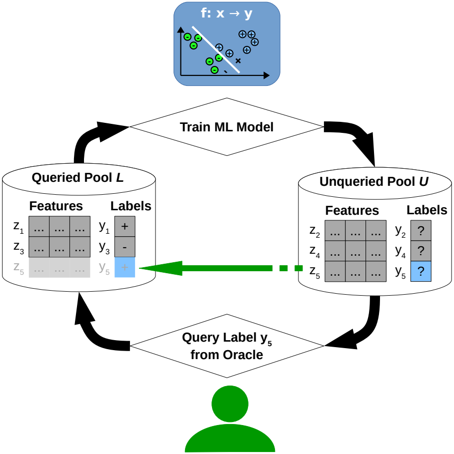
- unlabelled data $\mathcal{U}$ is abundant
- annotation is costly (paucity of labelled data $\mathcal{L}$)
- control over label selection process

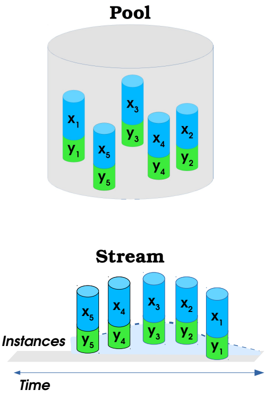
Aim of Active Learning

- select the most valuable (informative) instances for labelling

Active Learning: How?

Data Collection Process for Active Labelling

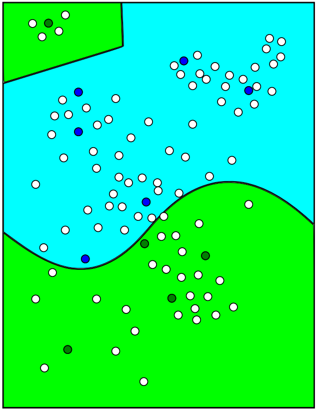




Active Learning: Illustrative Example

Which instance would you select?

Active Learning: Selection Criteria



What factors influence the decision?

- Density (improve the classifier, where decisions are important)
- Decision boundary (be specific, where change is expected)
- Label density (explore unexplored regions)

Influence Factors:

- **Decision boundary**: main criterion for decision making (prediction)
  - Proxy: posterior probability, margin, etc.
- **Reliability of decision**: identifies how sure one can be that the decision is already correct
  - Proxy: classifier ensemble diversity, labels distribution
- **Influence**: the influence of one instance for the complete dataset
  - Proxy: density, simulation
- **Class distribution**: are classes equally often represented
  - Proxy: class prior

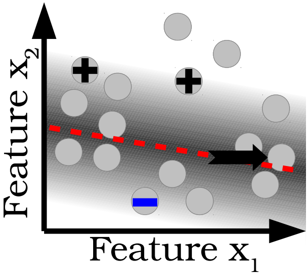
Active Learning Strategies: Overview

- Uncertainty Sampling:
  selects instances near the decision boundary
- Query by Committee:
  minimizes classifier variance
- Expected Error Reduction:
  simulates acquisition of each candidate and each possible outcome
- Probabilistic Active Learning:
  calculates expected performance locally
- ... (there exist many more methods)

Random Sampling ‣ skip

- Also called passive sampling
- Selects instances randomly for labeling
- Competitive approach
- Standard baseline
- Free of heuristics

Uncertainty Sampling [Cohn et al., 1990] ▸ skip



Idea

Select those instances where we are least certain about the label

Approach:

- 3 labels preselected
- Linear classifier
- Use *distance to the decision boundary* as *uncertainty measure*

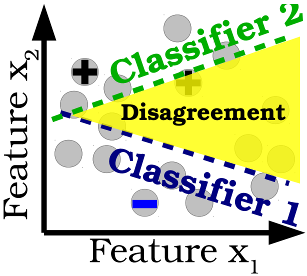
Discussion of Uncertainty Sampling

- $\oplus$ easy to implement
- $\oplus$ fast

- $\ominus$ no exploration (often combined with random sampling)
- $\ominus$ impact not considered (density weighted extensions exist)
- $\ominus$ problem with complex structures (performance can be even worse than random)

Influence factors: Decision boundary

Ensemble-Based Strategy [Seung et al., 1992] skip



Idea

Use disagreement between base classifiers

Approach

1. Get an initial set of labels
2. Split that set into (overlapping) subsets
3. On each subset, train a different base-classifier
4. Repeat until stop
5. On each unlabeled instance do
6. Apply all base-classifiers
7. Request label, if base-classifiers disagree
8. Update all base-classifiers
9. Go to step 4

Discussion of QbC

$\oplus$ applicable to every classifier (even discriminative ones)

$\ominus$ need more labels as some are hidden for some classifiers

$\ominus$ training of multiple classifiers

Influence factors: Decision boundary, Reliability of decision

Expected Error Reduction [Roy and McCallum, 2001] ▸ skip

- Simulates the acquisition of each label candidate and each possible outcome (class)
- Calculates the generalization error of the simulated new model
- Chooses the label with lowest generalization error

$$x^* = \mathrm{argmin}_x \sum_{i \in \{1,\ldots,C\}} P_\theta(y_i \mid x) \left(\sum_{x' \in \mathcal{U}} 1 - P_{\theta + (x, y_i)}(\hat{y} \mid x') \right)$$

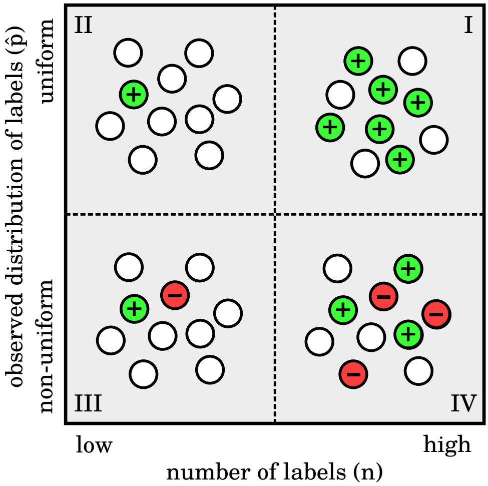
Discussion of Expected Error Reduction

$\oplus$ decision theoretic model

$\ominus$ long execution time (closed form solutions for specific classifiers, approximations for speed up)

Influence factors: Decision boundary, Reliability of decision, Impact

Exemplary AL Situations



OPAL

- a label's value depends on the label information in its neighbourhood
- label information
  - number of labels
  - share of classes
- uncertainty sampling ignores **the number of similar labels**

Probabilistic Active Learning [Krempl et al., 2015b]

- Models the *true* posterior as being Beta-distributed
  - variance of posterior is correlated with the number of local observations
  - thereby omit the complex simulation of expected error reduction
- Calculates the performance improvement of the model

$$G_{\mathrm{OPAL}}(\mathit{ls}, m) = \frac{1}{m} \cdot \mathtt{E}_p\Big[\mathtt{E}_k\left[\mathrm{gain}_p(\mathit{ls}, k, m)\right]\Big]$$

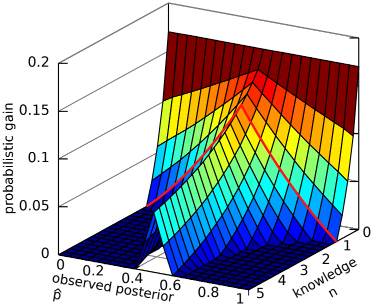
with:

- $\mathit{ls} = (n, \hat{p})$: Label statistics
- p: True posterior at candidate's position
- m: Number of candidates to be acquired (budget)
- k: Number of candidates with positive label realisations

- with performance gain as difference between future and current performance:

$$\mathrm{gain}_p(\mathit{ls}, k, m) = \mathrm{perf}_p\left(\frac{n\hat{p} + k}{n + m}\right) - \mathrm{perf}_p(\hat{p})$$

Probabilistic Active Learning [Krempl et al., 2015b] ▸ skip



- Models the *true* posterior as being Beta-distributed
  - variance of posterior is correlated with the number of local observations
  - thereby omit the complex simulation of expected error reduction
- Calculates the performance improvement of the model

Discussion of Probabilistic Active Learning

- $\oplus$ decision theoretic model
- $\oplus$ fast w.r.t. expected error reduction

- $\ominus$ local number of labels required

Influence factors: Decision boundary, Reliability of decision, Impact

Active Learning Strategies: Overview

- Uncertainty Sampling:
  selects instances near the decision boundary
- Query by Committee:
  minimizes classifier variance
- Expected Error Reduction:
  simulates acquisition of each candidate and each possible outcome
- Probabilistic Active Learning:
  calculates expected performance locally
- ... (there exist many more methods)

Active Learning

Outline & Tentative Schedule

- 09:00 – 10:30 Lecture
  - Opening
  - Evaluating
  - Broadening the View
- 10:45 – 11:45 Practical
- 11:45 – 12:15 Discussion

Getting Ready

- Have scikit-learn and scitkit activeml installed:

```
pip install scikit-learn
pip install scikit-activeml
```

Evaluation: Objectives & Criteria

Main Objectives

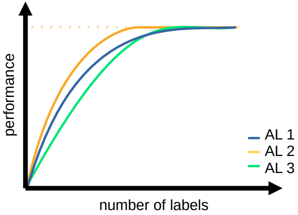
- Maximise classification performance
- Minimise labelling costs / label requests

Criteria

- Performance of classifier, depending on
- Number of acquired labels / spent budget

- Exploration of data space?
- Explainability?
- Query runtime?

AL Strategies - Evaluation: Learning Curve



Plots

- (classification) performance versus
- used budget / number of label requests

Expected behaviour:

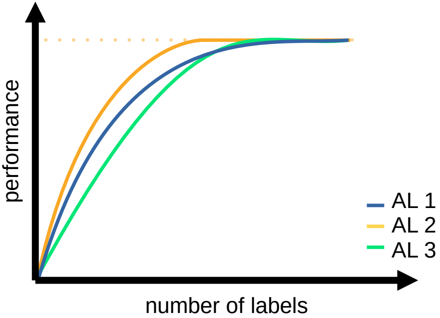
- Identical performance for budget = 0
- Performance increases with number of labels
- Convergence: after ∞ label requests, all strategies should have the same performance

Caveat

Always compare using same classifier and data

How to interpret the results of a learning curve?

- Converging as fast as possible
- Converging to the highest overall value

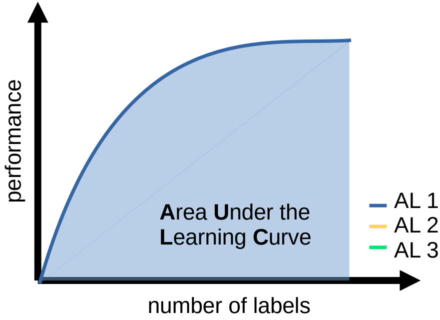


Aggregated Measures

How to summarize results from a learning curve?

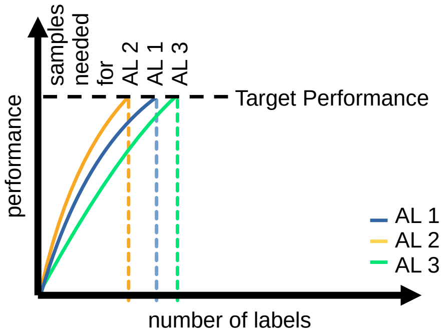
- Table at specific time points (early, mid, late)
- Area under the learning curve, mean (depends on stopping point) [Culver et al., 2006]
- Data Utilisation Rate [Reitmaier and Sick, 2013]

Area Under the Learning Curve (AULC)[Culver et al., 2006]



- AULC above that of a random-sampling learner
- Calculated for maximum budget, thus **sensitive to budget**
- Negative value indicates worse-than-random performance

Data Utilization Rate (DUR) [Reitmaier and Sick, 2013]



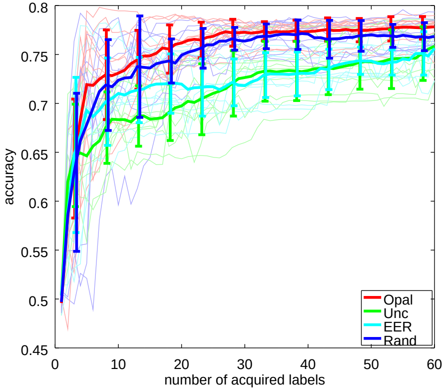
- The **minimum number of samples needed** to reach a **target accuracy**, divided by the number of samples needed by a random sampling learner
- Indication of efficiency for selecting of data
- Sensitive to choice of target accuracy, ignores performance changes at other points

How to evaluate statistical significance?

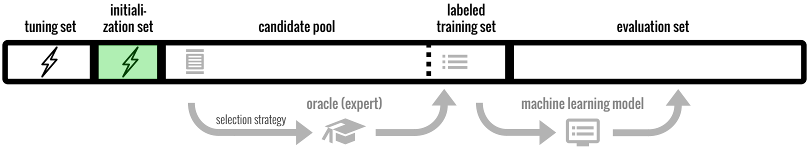
- Which values to compare?
  - **not** across label acquisitions (highly correlated) but across multiple repetitions
  - at which point in time?
- Statistical tests
  - t-Test cmp. mean (assumes that mean is normal distributed)
  - Wilcoxon Signed Rank Test cmp. tendency (parameter-free test)

How many repetitions are required?

Comparison of algorithms using 5-fold cross validation

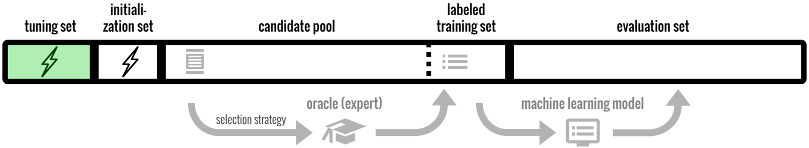


Initialization of Instance Selection



- Cannot be class-specific, as labels are unknown
- Often random (How to tune the number of random samples?)

Parameter Tuning



1. Determine hyperparameter and fix them across selection methods
2. How to tune without labels?

Parameter Tuning

- tuning instances should be considered in the number of acquisitions
- how many instances should be used for tuning? (many classifiers are sensitive to the number of instances)
- normally, no instances for supervised parameter tuning available
- tuning parallel to sampling may be complicated

Evaluation Challenges

Real applications oft are more challenging

- Often highly specialized (hard to transfer approaches to related domains)
- Imperfect labelers (experts might be wrong)
- In real-world only one shot (mean results are not representative)
- Labels are not always available (in time and space)
- Performance guarantees (cmp. random sampling)
- Assess online performance of an actively trained classifier
- Different costs for different annotations or classes
- Ground truth might not be available

Evaluation: Recommendations [1]

- Use exactly the **same robust classifier** for every AL method when comparing and try to sync the parameters of these classifiers.
- Capture the effect of different AL methods on multiple datasets using **at least 50 repetitions**.
- Start with an **initially unlabeled set**. If you need initial training instances, sample randomly and explain when to stop.
- Use either an **apriori defined stopping criterion** or enough label acquisitions (sample until convergence).
- Show **learning curves** (incl. quartiles) with reasonable performance measures.
- Present **pairwise differences** in terms of significance and effect size (Wilcoxon signed rank test).

[1] See [Kottke et al., 2017].

Outline & Tentative Schedule

- 09:00 – 10:30 Lecture
  - Opening
  - Evaluating
  - Broadening the View
- 10:45 – 11:45 Practical
- 11:45 – 12:15 Discussion

Getting Ready

- Have scikit-learn and scitkit activeml installed:

```
pip install scikit-learn
pip install scikit-activeml
```

Beyond pool-based scenarios

Beyond pool-based scenarios

Aims

- **Broadening view** on active learning
- **Overview** on different variants of the active learning task
- **Pointers** to surveys / key papers for each variant
- **Challenges/caveats** and exemplary approaches

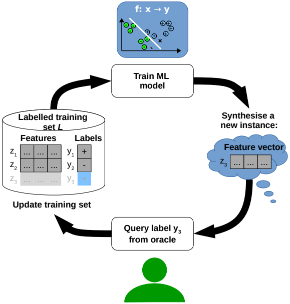
Active Learning: Scope

processing scenarios	learning objective	initiation of interaction	selected information
pool-based	inductive	active learning	of labels (labelling)
stream-based	transductive	machine teaching	of features (feature acquistion)
query synthesis			of instances (class selection)
			of classification features / instances

Processing Scenarios

processing scenarios		learning objective		initiation of interaction		selected information
pool-based	✖	inductive	✖	active learning	✖	of labels (labelling)
stream-based		transductive		machine teaching		of features (feature acquistion)
query synthesis						of instances (class selection)
						of classification features / instances

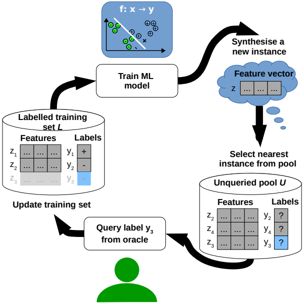
Processing Scenarios: Query Synthesis



Query Synthesis Scenario

- No **pool**
- **Ad hoc generation** of queried instances
- **Membership query**: Query class membership of generated instance
- See [Angluin, 2004] (introduction)

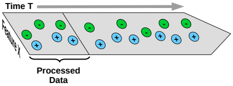
- **Challenge: creating meaningfull instances**



Hybrid Query Synthesis/Pool Scenario

- **Aim: creating meaningfull instances**
- **Combination with pool-based AL**: [Wang et al., 2015]
  - given a (too) large pool of unlabelled data
  - synthesize instance close to decision boundary
  - select the nearest neighbouring real instance
  - faster than pool-based AL, meaningful queries

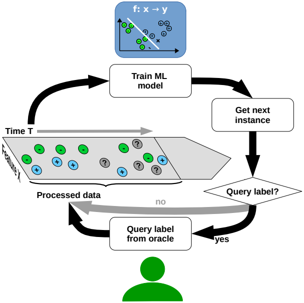
Processing Scenarios: Stream



Stream-Based Selective Sampling Scenario

- **Sequential arrival**, **no repeated access**
- **Online** active learning as synonym
- **No/few initial labels**
- **Possibly infinite** number of instances
- **Efficient processing** and limited storage
- **Non-stationary** distributions (concept drift)
- **Adaptation** (forgetting) needed
- "Big Data" is often streaming data

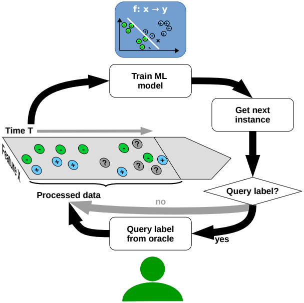
Processing Scenarios: Stream



Stream-Based Selective Sampling Scenario

- **Decide upon arrival** of new instance whether to query that instance's label or not
- **Update classifier** if label was queried, otherwise skip
- **Continue** for as long as new instances arrive

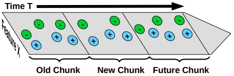
Processing Scenarios: Stream



Recommended literature

- [Cacciarelli and Kulahci, 2023] (survey)
- [Zliobaitė et al., 2013] (concept drift)
- [Kottke et al., 2015] (budget management)
- [Pham et al., 2022] (verification latency)

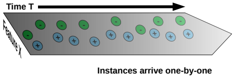
Processing Scenarios: Stream



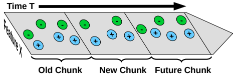
Chunk-based processing

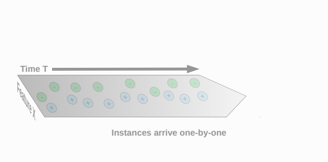
versus

Instance-wise processing



Processing Scenarios: Stream



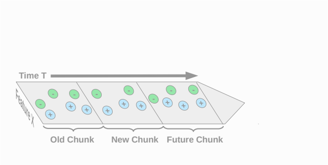


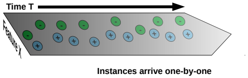
Chunk-based processing

- **Split data chronologically into chunks**
- AL on each chunk is similar to pool-based AL
- Often, ensemble with one new classfier per chunk is trained [a]
- Alernative: Clustering-based approaches [b]

[a] E.g., [Ryu et al., 2012, Zhu et al., 2010, Zhu et al., 2007]
[b] E.g., [Krempl et al., 2015a, Ienco et al., 2013]

Processing Scenarios: Stream





Instance-wise processing

- **Instances arrive one-by-one**
- Decision to query or not must be taken at once
- **Budget:** Trade-off between spatial and temporal usefulness [a]

[a] See [Kottke et al., 2015]

Active Learning: Learning Objective

processing scenarios		**learning objective**		initiation of interaction		selected information
pool-based	✖	inductive	✖	active learning	✖	of labels (labelling)
stream-based		**transductive**		machine teaching		of features (feature acquistion)
query synthesis						of instances (class selection)
						of classification features / instances

Georg Krempl g.m.krempl@uu.nl Utrecht University

Inductive

- Training and test data are different
- Objective: Generalising to unseen data

Transductive

- Same data used for training needs to be classified
- Objective: Mastering given (training) data set

Learning Objective: Inductive vs. Transductive ▸ skip

Particularities of Transductive AL

- **Evaluation data is known beforehand**, as test and train set are identical, no need to build a generalised model
- **Excluding** instances from being predicted by the classifier is possible by querying them from the oracle

Implications

- Ignore high aleatoric uncertainty for inductive setting
- Remove such instances by labelling for transductive setting
- See [Kottke et al., 2022]

Learning Objective: Inductive vs. Transductive

Transductive Gain

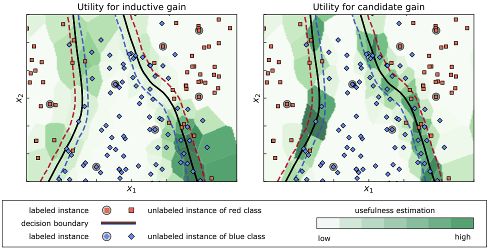


Figure: Transductive gain as sum of the utilities of inductive gain (left), and of candidate gain (right) [Kottke et al., 2022, Fig.1]

Active Learning: Initiatior of Interaction

processing scenarios
- pool-based
- stream-based
- query synthesis

✖

learning objective
- inductive
- transductive

✖

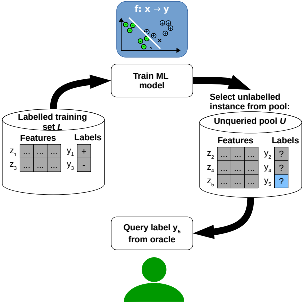
initiation of interaction
- active learning
- machine teaching

✖

selected information
- of labels (labelling)
- of features (feature acquistion)
- of instances (class selection)
- of classification features / instances

Georg Krempl g.m.krempl@uu.nl Utrecht University

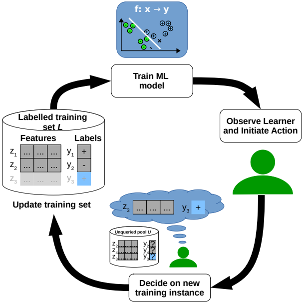
Initiatior of Interaction: Machine (Active Learning)



Active Learning

- **Machine** is proactive in the interaction

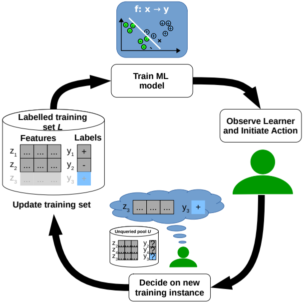
Initiatior of Interaction: Human (Machine Teaching)



Machine Teaching

- **Human** is proactive in the interaction
- **No direct knowledge transfer** between teacher (human) and learner (machine)
- **Aim is designing an optimal training set**
- See [Tegen, 2022] (PhD thesis) and [Tegen et al., 2021] (review)

Initiatior of Interaction: Human (Machine Teaching)



Triggers for human to add instances to training set might be

- Trigger by **error**
- Trigger by **state change**
- Trigger by **time**
- Trigger by **user factors**

Active Learning: Selected Information

processing scenarios		learning objective		initiation of interaction		**selected information**
pool-based	✖	**inductive**	✖	**active learning**	✖	**of labels (labelling)**
stream-based		**transductive**		**machine teaching**		**of features (feature acquistion)**
query synthesis						**of instances (class selection)**
						of classification features / instances

Question?

Bibliography I

Angluin, D. (2004).
Queries revisited.
Theoretical Computer Science, 313(2):175–194.

Cacciarelli, D. and Kulahci, M. (2023).
A survey on online active learning.
arXiv preprint, (arXiv:2302.08893).

Cohn, D., Atlas, L., Ladner, R., El-Sharkawi, M., Marks, R., Aggoune, M., and Park, D. (1990).
Training connectionist networks with queries and selective sampling.
In *Advances in Neural Information Processing Systems (NIPS)*. Morgan Kaufmann.

Culver, M., Kun, D., and Scott, S. (2006).
Active learning to maximize area under the roc curve.
In *Sixth International Conference on Data Mining (ICDM'06)*, pages 149–158. IEEE.

Holzinger, A. (2016).
Interactive machine learning for health informatics: When do we need the human-in-the-loop?
Brain Informatics, 3:119–131.

Bibliography II

Ienco, D., Bifet, A., Zliobaite, I., and Pfahringer, B. (2013).
Clustering based active learning for evolving data streams.
In fürnkranz, J., Hüllermeier, E., and Higuchi, T., editors, *Proceedings of the 16^{th} Int. Conf. on Discovery Science (DS), Singapore*, volume 8140 of *Lecture Notes in Artificial Intelligence*, page 79–93. Springer.

Kok, T., Brouwer, R. M., Mandl, R. M., Schnack, H. G., and Krempl, G. (2021).
Active selection of classification features.
In *Advances in Intelligent Data Analysis XIX. IDA 2021*, volume 12695 of *LNCS*, pages 184–195. Springer.

Kottke, D., Herde, M., Minh, T. P., Benz, A., Mergard, P., Roghman, A., Sandrock, C., and Sick, B. (2021).
scikitactiveml: A Library and Toolbox for Active Learning Algorithms.
Preprints.

Kottke, D., Huseljic, D., Calma, A., Krempl, G., and Sick, B. (2017).
Challenges of reliable, realistic and comparable active learning evaluation.
In *Proc. of the Workshop and Tutorial on Interactive Adaptive Learning*, volume 1924 of *Workshop Proceedings*. CEUR.

Kottke, D., Krempl, G., and Spiliopoulou, M. (2015).
Probabilistic active learning in data streams.
In Fromont, E., Bie, T. D., and Leeuwen, M. v., editors, *Advances in Intelligent Data Analysis XIV*, volume 9385 of *LNCS*, page 145–157. Springer.

Bibliography III

Kottke, D., Krempl, G., Stecklina, M., Styp von Rekowski, C., Sabsch, T., Pham Minh, T., Deliano, M., Spiliopoulou, M., and Sick, B. (2016).
Probabilistic active learning for active class selection.
In Mathewson, K., Subramanian, K., and Loftin, R., editors, *Proc. of the NIPS Workshop on the Future of Interactive Learning Machines*.

Kottke, D., Sandrock, C., Krempl, G. K., and Sick, B. (2022).
A stopping criterion for transductive active learning.
In *Proc. of the Europ. Conf. on Machine Learning and Principles of Knowledge Discovery in Databases (ECMLPKDD 2022)*.

Krempl, G., Ha, T. C., and Spiliopoulou, M. (2015a).
Clustering-based optimised probabilistic active learning (COPAL).
In Japkowicz, N. and Matwin, S., editors, *Proc. of the 18th Int. Conf. on Discovery Science*, volume 9356 of *LNCS*, page 101–115. Springer.

Krempl, G., Kottke, D., and Lemaire, V. (2015b).
Optimised probabilistic active learning (OPAL) for fast, non-myopic, cost-sensitive active classification.
Machine Learning, 100(2).

Mosqueira-Rey, E., Hernandez-Pereira, E., Alonso-Rios, D., Bobes-Bascaran, J., and Fernandez-Leal, A. (2022).
Human-in-the-loop machine learning: a state of the art.
Artificial Intelligence Review.

Bibliography IV

Pham, T., Kottke, D., Krempl, G., and Sick, B. (2022).
Stream-based active learning for sliding windows under verification latency.
Machine Learning.

Reitmaier, T. and Sick, B. (2013).
Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4ds.
Information Sciences, 230:106–131.

Roy, N. and McCallum, A. (2001).
Toward optimal active learning through sampling estimation of error reduction.
In *Proc. of the 18th Int. Conf. on Machine Learning, ICML 2001, Williamstown, MA, USA*, page 441–448, San Francisco, CA, USA. Morgan Kaufmann.

Ryu, J. W., Kantardzic, M. M., Kim, M.-W., and Khil, A. R. (2012).
An efficient method of building an ensemble of classifiers in streaming data.
In *Big Data Analytics*, page 122–133. Springer.

Settles, B. (2012).
Active Learning.
Number 18 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.

Bibliography V

Seung, H. S., Opper, M., and Sompolinsky, H. (1992).
Query by committee.
In M.K., W. and L.G., V., editors, *Proc. of the fifth workshop on computational learning theory*. Morgan Kaufmann.

Tegen, A. (2022).
Interactive Online Machine Learning.
PhD thesis, Malmö University.

Tegen, A., Davidsson, P., and Persson, J. A. (2021).
A taxonomy of interactive online machine learning strategies.
In Hutter, F., Kersting, K., Lijffijt, J., and Valera, I., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 137–153, Cham. Springer International Publishing.

Wang, L., Hu, X., Yuan, B., and Lu, J. (2015).
Active learning via query synthesis and nearest neighbour search.
Neurocomputing, 147:426–434.

Zhu, X., Zhang, P., Lin, X., and Shi, Y. (2007).
Active learning from data streams.
In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, page 757–762, Washington, DC, USA. IEEE Computer Society.

Zhu, X., Zhang, P., Lin, X., and Shi, Y. (2010).
Active learning from stream data using optimal weight classifier ensemble.
IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 40(6):1607 – 1621.

Zliobaitė, I., Bifet, A., Pfahringer, B., and Holmes, G. (2013).
Active learning with drifting streaming data.
IEEE Transactions on Neural Networks and Learning Systems, 25(1):27–39.