Active learning (machine learning)

Active learning is a special case of machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points.[1] [2] [3] In statistics literature it is sometimes also called optimal experimental design. [4]

There are situations in which unlabeled data is abundant but manually labeling is expensive. In such a scenario, learning algorithms can actively query the user/teacher for labels. This type of iterative supervised learning is called active learning. Since the learner chooses the examples, the number of examples to learn a concept can often be much lower than the number required in normal supervised learning. With this approach, there is a risk that the algorithm is overwhelmed by uninformative examples. Recent developments are dedicated to multi-label active learning[5], hybrid active learning[6] and active learning in a single-pass (on-line) context,[7] combining concepts from the field of machine learning (e.g. conflict and ignorance) with adaptive, incremental learning policies in the field of online machine learning.

Contents

Definitions

Let T be the total set of all data under consideration. For example, in a protein engineering problem, T would include all proteins that are known to have a certain interesting activity and all additional proteins that one might want to test for that activity.

During each iteration, i, T is broken up into three subsets

1. $\mathbf{T}_{K,i}$: Data points where the label is **known**.
2. $\mathbf{T}_{U,i}$: Data points where the label is **unknown**.
3. $\mathbf{T}_{C,i}$: A subset of $T_{U,i}$ that is **chosen** to be labeled.

Most of the current research in active learning involves the best method to choose the data points for $T_{C,i}$.

Query strategies

Algorithms for determining which data points should be labeled can be organized into a number of different categories, based upon their purpose:[1]

- **Balance exploration and exploitation**: the choice of examples to label is seen as a dilemma between the exploration and the exploitation over the data space representation. This strategy manages this compromise by modelling the active learning problem as a contextual bandit problem. For example, Bouneffouf et al.[8] propose a sequential algorithm named Active Thompson Sampling (ATS), which, in each round, assigns a sampling distribution on the pool, samples one point from this distribution, and queries the oracle for this sample point label.
- **Expected model change**: label those points that would most change the current model
- **Expected error reduction**: label those points that would most reduce the model's generalization error
- **Exponentiated Gradient Exploration for Active Learning**:[9] In this paper, the author proposes a sequential algorithm named exponentiated gradient (EG)-active that can improve any active learning algorithm by an optimal random exploration.
- **Membership Query Synthesis**: This is where the learner generates its own instance from an underlying natural distribution. For example, if the dataset are pictures of humans and animals, the learner could send a clipped image of a leg to the teacher and query if this appendage belongs to an animal or human. This is particularly useful if your dataset is small.[10]
- **Pool-Based Sampling**: In this scenario, instances are drawn from the entire data pool and assigned an informative score, a measurement of how well the learner "understands" the data. The system then selects the most informative instances and queries the teacher for the labels.[11]
- **Stream-Based Selective Sampling**: Here, each unlabeled data point is examined one at a time with the machine evaluating the informativeness of each item against its query parameters. The learner decides for itself whether to assign a label or query the teacher for each datapoint.
- **Uncertainty sampling**: label those points for which the current model is least certain as to what the correct output should be
- **Query by committee**: a variety of models are trained on the current labeled data, and vote on the output for unlabeled data; label those points for which the "committee" disagrees the most
- **Querying from diverse subspaces or partitions** [12]: When the underlying model is a forest of trees, the leaf nodes might represent (overlapping) partitions of the original feature space. This offers the possibility of selecting instances from non-overlapping or minimally overlapping partitions for labeling.
- **Variance reduction**: label those points that would minimize output variance, which is one of the components of error

A wide variety of algorithms have been studied that fall into these categories.[1][4]

Minimum Marginal Hyperplane

Some active learning algorithms are built upon Support vector machines (SVMs) and exploit the structure of the SVM to determine which data points to label. Such methods usually calculate the margin, W, of each unlabeled datum in $T_{U,i}$ and treat W as an n-dimensional distance from that datum to the separating hyperplane.

Minimum Marginal Hyperplane methods assume that the data with the smallest W are those that the SVM is most uncertain about and therefore should be placed in $T_{C,i}$ to be labeled. Other similar methods, such as Maximum Marginal Hyperplane, choose data with the largest W. Tradeoff methods choose a mix of the smallest and largest Ws.

Meetings

- 2016 "Workshop Active Learning: Applications, Foundations and Emerging Trends" at iKNOW, Graz, Austria[13]
- 2018 "Interactive Adaptive Learning" Workshop at ECML PKDD, Dublin, Ireland[14]

See also

- Proactive learning
- List of datasets for machine learning research

Notes

1. Settles, Burr (2010), "Active Learning Literature Survey" (http://pages.cs.wisc.edu/~bsettles/pub/settles.activelearning.pdf) (PDF), *Computer Sciences Technical Report 1648. University of Wisconsin–Madison*, retrieved 2014-11-18
2. Rubens, Neil; Elahi (https://www.linkedin.com/in/mehdielahi), Mehdi; Sugiyama, Masashi; Kaplan, Dain (2016). "Active Learning in Recommender Systems" (https://rd.springer.com/chapter/10.1007/978-1-4899-7637-6_24). In Ricci, Francesco; Rokach, Lior; Shapira, Bracha. *Recommender Systems Handbook* (https://rd.springer.com/book/10.1007/978-1-4899-7637-6) (2 ed.). Springer US. doi:10.1007/978-1-4899-7637-6 (https://doi.org/10.1007%2F978-1-4899-7637-6). ISBN 978-1-4899-7637-6.
3. Das, Shubhomoy; Wong, Weng-Keen; Dietterich, Thomas; Fern, Alan; Emmott, Andrew (2016). "Incorporating Expert Feedback into Active Anomaly Discovery". In Bonchi, Francesco; Domingo-Ferrer, Josep; Baeza-Yates, Ricardo; Zhou, Zhi-Hua; Wu, Xindong. *IEEE 16th International Conference on Data Mining* (https://doi.org/10.1109/ICDM.2016.0102). IEEE. doi:10.1109/ICDM.2016.0102 (https://doi.org/10.1109%2FICDM.2016.0102).
4. Olsson, Fredrik. "A literature survey of active machine learning in the context of natural language processing" (http://eprints.sics.se/3600/).
5. Yang B, Sun J T, Wang T J, et al.(2009), "Effective Multi-Label Active Learning for Text Classification". (https://www.microsoft.com/en-us/research/wp-content/uploads/2009/01/sigkdd09-yang.pdf) Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 917-926.
6. E. Lughofer (2012), Hybrid Active Learning (HAL) for Reducing the Annotation Efforts of Operators in Classification Systems. (http://www.sciencedirect.com/science/article/pii/S0031320311003463) Pattern Recognition, vol. 45 (2), pp. 884-896, 2012.
7. E. Lughofer (2012), Single-Pass Active Learning with Conflict and Ignorance. (https://link.springer.com/article/10.1007/s12530-012-9060-7) Evolving Systems, vol. 3 (4), pp. 251-271, 2012.
8. Bouneffouf et al. (2014), Contextual Bandit for Active Learning: Active Thompson Sampling. (https://hal.archives-ouvertes.fr/hal-01069802) Neural Information Processing - 21st International Conference, ICONIP 2014
9. Bouneffouf et al. (2016), Exponentiated Gradient Exploration for Active Learning. (http://www.mdpi.com/2073-431X/5/1/1) Computers, vol. 5 (1), 2016, pp. 1-12
10. "Active learning via query synthesis and nearest neighbour search" (https://www.sciencedirect.com/science/article/pii/S0925231214008145). *Neurocomputing*. **147**: 426–434. 2015-01-05. doi:10.1016/j.neucom.2014.06.042 (https://doi.org/10.1016%2Fj.neucom.2014.06.042). ISSN 0925-2312 (https://www.worldcat.org/issn/0925-2312).
11. "How does Active Learning Work in Practice?" (https://deepai.org/machine-learning-glossary-and-terms/active-learning). *deepai.org*.
12. S. Das, [1] (https://github.com/shubhomoydas/ad_examples#query-diversity-with-compact-descriptions) https://github.com/shubhomoydas/ad_examples#query-diversity-with-compact-descriptions
13. http://vincentlemaire-labs.fr/iknow2016/
14. http://www.uni-kassel.de/go/ial2018

Other references

- N. Rubens (http://ActiveIntelligence.org), M. Elahi (https://www.linkedin.com/in/mehdielahi), M. Sugiyama, D. Kaplan. Recommender Systems Handbook: Active Learning in Recommender Systems (eds. F. Ricci, P.B. Kantor, L. Rokach, B. Shapira). Springer, 2015 [2] (https://www.researchgate.net/publication/296481444_Active_Learning_in_Recommender_Systems), [3] (http://activeintelligence.org/research/al-rs/).
- Active Learning Tutorial (http://hunch.net/~active_learning/), S. Dasgupta and J. Langford.