1 Good Tutorial [1]

The tutorial starts with a motivation: Why we need active learning? There are a lot of data available out there (documents off the web, speech samples, images, videos etc.), but they are unlabeled and labeling them all can be expensive. Therefore, a successful labelling heuristic could really help. The general acative learning heuristic suggested by Dasgupta is as follows:

- At the beginning we have a pool of unlabeled data points
- Pick a few random points and get their label
- Repeat
  - Fit a classifier to points seen so far
  - Query points closest to boundary (or most uncertain or most likely to decrease overall uncertainty)

However, the above heuristic bears the problem of sampling bias: As learner queries the points closer to its current boundary at each iteration, the training set quickly diverges from the underlying distribution of the data.

Active learning has similarities to supervised and unsupervised learning. Like supervised learning, the ultimate goal is learning a classifier. Like unsupervised learning labels are unknown and labels come with a certain cost. Idea is to query as few points as possible, yet attaining an accurate classifier with a lower cost than regular supervised learning algorithm.

Recent literature uses 2 ways to explain when active learning is helpful: 1) Efficient search through hypothesis space and 2) expliting cluster structure in data. In the former explanation, each label helps to shrink the set of plausible classifiers and by using active learning one can choose the points that would help to shrink this set as fast as possible. The latter explanation assumes that the data forms clusters and by discovering these clusters we can label the unlabaled points. Most primitive form is to find clusters in data, query random points and assign each cluster to its majority class, then finally to build the classifier on these clusters.

The theory tells us that active learning can be applied anywhere supervised applies and current methods apply best when a discrete loss is approach is adopted, a low noise data is used, the number of possible predictors are limited and when we have low data (where we do not have lots of labeled data.)

2 Hierarchical sampling for active learning [2]

In the paper authors present an active learning scheme, where they use cluster structures in data. Their method first applies hierarchical clustering on the data and the root of the forming tree is assumed to be the cluster containing all the intances. Then random instances in the tree are queried for their labels and depending on these labels, statistics for each pruning of the tree to see whether a particular pruning is mixed with different labels or is relatively pure with one class. Querying can be stopped at any point and the clusters are assigned their labels depending on the majorith vote. However, the stopping occurs whenever prunings of the tree are statistically judged to be pure. Their method is statistically consistent and never worse than supervised techniques. Furthermore when compared to baseline active learning method of random sampling, it reduces the number of points queried significantly. Yet the success of this

method depends mainly on how well actual labels are aligned with the cluster structure in the data.

3 On proper unit selection in active learning: co-selection effects for named entity recognition [3]

Due to biased sampling active learning may fail to explore large regions and ultimately may miss clusters in data, which lowers the recall and reduces learning for infrequent classes. Authors study the missed class effect, which is a form of missed cluster effect (complete class is overlooked by learner) and they focus on missed class effect in named entity recognition (NER) context, which is a common natural language processing (NLP) task. They show that missed classes can be avoided : They change the sampling granularity from single-unit instances (tokens) to multi-unit instances (sentences) where they make use of the co-selection effect. Co-selection effect: When an AL model uses sequences instead of tokens, instances of different classes co-occur and an active learner selecting uncertain examples of one class implicity selects examples from other classes as well. In experiments authors compare AL based on tokens to AL based on sentences and they use 3 different datasets. As a result, authors show that missed class problem in sequence classification tasks can be avoided using sentences as units of labeling. They have also found that the missed class effect is more critical, if single tokens rather than sentences are used for labeling and the co-selection of tokens provides an implicit exploratory aspect.

4 Active Learning for Biomedical Citation Screening [4]

Interesting take-away techniques in this paper. Typically papers with medical data have a good supply of oracles so Wallace capitalized on this.

Co-Feature: Use several different oracles (1-3) to assign labels to data points. Wallace's approach trusts the oracles' a priori domain knowledge (as opposed to using a prediction based system) where the oracles agree and address the points where they differ.

Concept Drift: Ideas (or labels) associated with input points change over time, perhaps as more information becomes available. Addressed by weighting the most recently labeled features.

5 Multi-Label Learning by Exploiting Label Dependency [5]

LEAD: Instead of using the original labels present in the sample as a source for query based learning methods, LEAD uses a Bayesian Network to construct classifiers for all labels indepently of one another. Recursively, this procedure produces accurate classifiers while reducing the error rate associated with classifying instances using their original labels.

6 Employing EM in Pool-Based Active Learning for Text Classification [6]

Pool Based Learning: Usage of keeping a pool of unlabeled examples from which to query. This practice is common to almost all active learning procedures.

Query by Committee (QBC): Samples several times from a classifier parameter distribution that results from training data in order to create a committee of classifier variants. The committee is meant to approximate the entire classifier distribution. QBC measures variance by examining disagreements between committee members. When members disagree strongly (are different enough), those members are selected for labeling requests.

7 Analysis of a greedy active learning strategy [7]

**Discredits use of synthetic query points.

Dasgupta's binary search of the space requires, in most cases, d log m labels. This is a substantially more appealing than selecting all m labels or randomly querying the space.

The greedy choice made by his approach, to always ask for the label which most evenly divides the space, satisfies the greedy choice property. This does not necessarily minimize the number of queries that must be made but typically requires at most O(ln —H—), where H is the target hypothesis. This is optimal based on the literature.

8 The Use of Simulated Experts in Evaluating Knowledge Acquisition [8]

Covers the use of common learning algorithms such as Naive Bayes to stand in for domain expert, since domain experts aren't necessarily easy to come by. What the stand in learner will do is create a classifier that can be used to make decisions about labeling or clustering techniques.

Given that the expertise provided by a learning algorithm cannot be an ideal substitute for a living, breathing expert; the simulated expert improved the knowledge acquisition process.

** Relevant? An interesting point Compton usually makes is that when a domain expert is asked why they made a certain decision they can't give you a straight answer.

9 Multi-View Multi-Label Active Learning for Image Classification [9]

Two Dimensional Active Learning (2DAL): Selects sample-label pairs instead of samples in each iteration. Used in multi-label problems.

Multi-view learning: Reduces the amount of labeled samples required for learning. Works by creating/training multiple hypotheses (classifiers) for the same label set and then utilizing their agreement (or disagreement) among different learners to improve overall classification performance.

10 Improving Generalization with Active Learning [10]

Old paper but explains why active learning is superior to randomized learning techniques since with active learning we have control over the input space. In random examples (locating a boundary on the unit line inverval) a random query engine would need $O(\frac{1}{\epsilon}ln(\frac{1}{\epsilon}))$ labels in order to arrive at an expected position error less than ϵ. Another concept discussed is the **region of uncertainty**, an area in the sample space where we believe that misclassification is still possible based on any information already given/gathered. This is the region we want to exclusively sample because we're comfortable accepting the high cost of classifying/labeling these points.

11 Agnostic Active Learning [11]

The work is first to be proposed as an active larning algorithm A^2 (Agnostic Active) that works in the presence of any form of (limited) noise. The assumption of the algorithm is that samples come drawn i.i.d. from a fixed distribution. A^2 is proposed to be a robust version of the selective sampling proposed by Cohen et. al. [10]. The difference of A^2 to selective sampling is that unlike selective sampling, it is an agnostic approach and does not eliminate a hypothesis depending on a single query. Authors show that in particular that A^2 achieves exponential speedups in several settings previously analyzed without a noise. A^2 performs particularly well for the simple case of learning threshold functions (this holds for arbitrary distributions as well, provided that the noise rate is low).

12 Coarse Sample Complexity Bounds for Active Learning [12]

Previous work has shown that in an active learning setting we need at most $O(dlogd/\epsilon)$ labels to learn a classifier with error rate less than ϵ. This is exponentially smaller than the sample complexity of learning a linear classifier in a supervised setting. However, the above situation is valid only if certain conditions are met: If the hypothesis class in an active learning setting is homogenous linear separators, if the data is distributed uniformly on the surface of a unit sphere and if data corresponds perfectly to either one of the hypotheses (i.e. seperable case). Generalization of this result to non-homogenous case is quite difficult and the complexity changes dramatically depending on the target hypothesis. In this paper authors characterize the sample complexity of active learning according to a parameter which takes into account: Distribution over the input space, specific target hypothesis and the wanted error rate ϵ. In particular authors define a splitting index ρ that induces a topology on the set of hypotheses and they show that ρ fairly captures the sample complexity of active learning. Depending on the splitting index authors define lower and upper bounds for sample complexity of active learning.

13 Importance Weighted Active Learning [13]

Importance Weighted Active Learning (IWAL): Beygelzimer et al. propose an agnostic routine based on importance weighting (loss functions) to correct sampling

bias, control variance, and to ultimately reduce the amount of labels needed to build an accurate classifier.

Loss Functions represent the loss (cost) associated with an estimate being wrong. Loss functions serve as the metric for a rejection threshold.

Basic Outline of IWAL: Run through the sample space and upon seeing instance x_t call rejection threshold subroutine. The rejection threshold subroutine considers x_t and all past instances and returns the probability p_t of requesting y_t. Thus, the algorithm shrinks the hypothesis class H_t to the subset of H_t whose empirical loss is close to the smallest empirical loss in H_t.

14 Active learning in the non-realizable case [14]

The **realizability assumption**: The learner's hypothesis class is assumed to contain a target function that perfectly classifies all training and test examples. This assumption is hardly attainable and in order to make the theory of active learning relevant to practice, the realizability assumption must be relaxed. Basically, we're shooting too high.

This paper analyzes bounding the noisy case in which much noise exists in the data. One would assume that active learning can't do much in here, since there's not much for the learner to intuit and it holds true since the learner has only the sample space to query. Kääriäinen places a lower bound of $\Omega(1/\epsilon^2)$ on active learning in the noisy case which means that active learning has no advantage over passive learning.

Kääriäinen goes on to argue that the difficulty that arises from high-noise rates isn't due to non-realizability and that noise-free problems are quite common in many domains.

15 Active Learning in the Drug Discovery Process [15]

Warmuth applies an iterative approach to selectively sampling previously unlabeled data for finding molecules that will bind with a particular molecule. It's certainly not time efficient to test each example when your pool could be quite large and classification is a time instensive procedure. 5% batches of the data are first chosen at random until a positive batch and a negative batch are found. All further batches are chosen using several selection strategies. Positive is the goal state for a batch since those compounds will bind. Labels were hidden from the learner in order to simulate the experiment and performance is benchmarked by reporting false positives.

- **Strategy 1, Nearest Neighbor:** Having normalized the data, batches are considered for being closest to the already labeled batches' hyperplanes.
- **Strategy 2, SVM:** Uses the maximum margin hyperplane produced by a support vector machine (SVM).
- **Strategy 3, VoPerc:** Store the weight vector of each example, remembering the vectors recorded in each pass for 100 random permutations of the labeled examples. Each weight vector is assigned a $\pm$ vote and the prediction on an example is positive if the resultant vote is greater than zero. Unlabeled examples are chosen that are closest to zero.

Results show that all strategies perform better than corresponding strategies and that VoPerc performs slightly better, having less variance.

16 Active Learning for Visual Object Detection [16]

This paper discusses an interactive system used to label images collected from street cameras as containing or not containing pedestrians. The idea here is to build an accurate classifier for detecting pedestrians automatically. Initially, the system requires a user to mark an image positive if it contains a pedestrian and also to place a rectangle on the pedestrians position in the image. It takes a human roughly 20 seconds to mark a pedestrian and 3 seconds to draw a rectangle around them. Images marked negative are sent to a cache. The system then interactively asks the user to label the hard to classify cases. Abramson's research uses a boosting algorithm based on features in the images to determine if a pedestrian may or may not be present.

References

1. S. Dasgupta and J. Langford, "Tutorial summary: Active learning," in *ICML*, p. 178, 2009.
2. S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *ICML '08: Proceedings of the 25th international conference on Machine learning*, (New York, NY, USA), pp. 208–215, ACM, 2008.
3. K. Tomanek, F. Laws, U. Hahn, and H. Schütze, "On proper unit selection in active learning: co-selection effects for named entity recognition," in *HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, (Morristown, NJ, USA), pp. 9–17, Association for Computational Linguistics, 2009.
4. B. C. Wallace, K. Small, C. E. Brodley, and T. A. Trikalinos, "Active learning for biomedical citation screening," in *Knowledge Discovery and Data Mining (KDD)*, 2010.
5. M.-L. Zhang and K. Zhang, "Multi-label learning by exploiting label dependency," in *Knowledge Discovery and Data Mining (KDD)*, 2010.
6. A. Mccallum and K. Nigam, "Employing em in pool-based active learning for text classification," 1998.
7. S. Dasgupta, "Analysis of a greedy active learning strategy," in *In Advances in Neural Information Processing Systems*, pp. 337–344, MIT Press, 2004.
8. P. Compton, P. Preston, and B. Kang, "The use of simulated experts in evaluating knowledge acquisition," in *University of Calgary*, pp. 12–1, 1995.
9. X. Zhang, J. Cheng, C. Xu, H. Lu, and S. Ma, "Multi-view multi-label active learning for image classification," in *ICME'09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, (Piscataway, NJ, USA), pp. 258–261, IEEE Press, 2009.
10. D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.
11. M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," in *ICML '06: Proceedings of the 23rd international conference on Machine learning*, (New York, NY, USA), pp. 65–72, ACM, 2006.
12. L. S. Dasgupta, "Coarse sample complexity bounds for active learning," in *Advances in Neural Information Processing Systems 18*, 2005.
13. A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," in *ICML 09: Proceedings of the 26th Annual International Conference on Machine Learning*, (New York, NY, USA), pp. 49–56, ACM, 2009.
14. M. Kääriäinen, "Active learning in the non-realizable case," in *NIPS Workshop on Foundations of Active Learning*, 2006.
15. M. K. Warmuth, G. Rätsch, M. Mathieson, J. Liao, and C. Lemmen, "Active learning in the drug discovery process," 2002.
16. Y. Abramson and Y. Freund, "Active learning for visual object recognition," 2006.