Subset Selection in Machine Learning: Theory, Applications, and Hands On

Rishabh Iyer[1], **Abir De**[2], **Ganesh Ramakrishnan**[2], **Jeff Bilmes**[3]

[1] University of Texas at Dallas, [2] Indian Institute of Technology, Bombay, [3] University of Washington, Seattle
rishabh.iyer@utdallas.edu, abir@cse.iitb.ac.in, ganesh@cse.iitb.ac.in, bilmes@uw.edu

Sub-area within AI/Keywords

Subset Selection, Combinatorial Optimization, Coresets, Combinatorial Information Measures, Efficient Training, Active Learning, Robustness, Fairness, Personalization, Human Assisted AI, Feature Selection and Model Compression

Suggested Duration

The suggested duration of the tutorial is 3 hours, 30 mins, plus 30 mins break, i.e., a half-day tutorial.

Tutorial Outline

Machine learning, and specifically deep learning has transformed numerous application domains like computer vision and video analytics, speech recognition, natural language processing, and so on. As a result, significant focus of researchers in the last decade has been on obtaining the most accurate models, often matching and sometimes surpassing human level performance in these areas. However, deep learning is also unlike human learning in many ways. To achieve the human level performance, deep models require large amounts of labeled training data, several GPU instances to train, and massive size models (ranging from hundreds of millions to billions of parameters). In addition, they are often not robust to noise, imbalance, and out of distribution data, and can also easily inherit the biases in the training data. Motivated by these desiderata and many more, we will present a rich framework of subset selection and coreset based approaches. We will begin by studying the theoretical aspects of subset selection, such as the modeling paradigms of coresets and submodularity, and the resulting optimization algorithms and theoretical properties. We will then study the application of subset selection to a number of areas like: a) compute-efficient training of deep models, b) label efficient methods like active learning, c) feature selection and model compression, d) targeted subset selection for robust, fair, and personalized learning, and e) human assisted learning. An important component of this tutorial will be a hands-on session where we will present a number of toolkits developed as a part of DECILE (www.decile.org), which include SUBMODLIB (submodular optimization), CORDS (coresets and subset selection for compute efficient training), DISTIL (Active learning), and TRUST (targeted subset selection).

Goal of the tutorial

The goal of this tutorial is to provide a gentle introduction to ideas in combinatorial optimization, coresets and submodularity to the broader machine learning and deep learning researchers, and ground this in applications. Specifically, we believe that the applications presented in this tutorial in areas such as label efficient, compute efficient, robust, fair and personalized learning will enable researchers to think beyond just improving the model accuracy and in broader yet important aspects like Green AI, fairness, robustness, personalization, data efficiency and so on. Furthermore, the hands-on demonstrations will also be useful to students and researchers from industry to get oriented in and practically stated with these topics. Another goal of this tutorial is to connect researchers working on theoretical and algorithmic areas to the numerous applications where their work can have impact, and vice-versa. The target audience of this tutorial are practitioners in deep learning and machine learning as well as researchers working on more theoretical areas in optimization in machine learning.

Content of the Tutorial

A growing number of machine learning problems involve finding subsets of data points. Examples range from selecting subset of labeled or unlabeled data points, to selecting subsets of features or parameters of a deep model, to selecting subsets of data for outsourcing predictions to humans (human assisted machine learning). The tutorial would encompass a wide variety of topics ranging from theoretical aspects of subset selection *e.g.*, coresets, submodularity, determinantal point processes, to several practical applications, *e.g.*, time and energy efficient learning, learning under resource constraints, active learning, human assisted learning, feature selection, model compression, feature induction, fair, robust and personalized machine learning *etc.*

We believe that this tutorial will prove very useful since, a) subset selection is naturally emerging and has often been considered in isolation in several of the above applications, and b) by connecting researchers working on both the theoretical and application domains above, we can foster a much needed discussion on reusing several technical innovations across these subareas and applications. Furthermore, we would also like to connect researchers working on the theoretical foundations of subset selection (in areas such as coresets and submodularity) with researchers working in applications (such as feature selection, active learning, data efficient learning, model compression, and human assisted machine learning).

This tutorial will focus on the following broad areas:

Theoretical and Algorithmic directions: We will explore several closely related concepts such as coresets (Feldman 2020; Bachem, Lucic, and Krause 2017), determinantal point processes (Kulesza and Taskar 2012) and submodularity (Fujishige 2005; Tohidi et al. 2020; Krause and Guestrin 2008). We will also cover some recent advances in combinatorial information measures (Iyer et al. 2021).

Applications: Ideas from subset selection have been demonstrated to yield promising results in a wide range of applications. These include, a) selecting a subset of a labaled training dataset to reduce running time, energy and costs (Wei, Iyer, and Bilmes 2015; Kaushal et al. 2019; Mirzasoleiman, Bilmes, and Leskovec 2020; Killamsetty et al. 2021a,b) (some of the state-of-the-art approaches have demonstrated 5x - 10x speedups and energy reductions with negligible accuracy loss); b) selecting a subset of unlabaled data (either one shot or in an active learning setting) to reduce labeling costs (Kaushal et al. 2019; Ash et al. 2019); c) selecting a subset of data for human assisted learning (to offload certain critical and hard decisions to the human) (De et al. 2020, 2021); d) targeted subset selection with goals like robustness, fairness, and personalization (Kaushal et al. 2021; Kothawade et al. 2021), e) feature selection to reduce dimensionality or feature acquisition costs (Das, Iyer, and Natarajan 2021; Peng, Long, and Ding 2005; Derezinski, Khanna, and Mahoney 2020; Elenberg et al. 2018) and Model compression to deploy deep learning models in resource constrained scenarios (Mariet and Sra 2015; Mussay et al. 2019; Dubey, Chatterjee, and Ahuja 2018); and finally, f) the combination of subset selection ideas in related emerging topics such as weak supervision, semi-supervised learning, self-supervision, and meta learning (Maheshwari et al. 2020).

Hands-On Session: In addition to going over the theory of subset selection and the numerous applications discussed above, we will also have an hands-on session, which we describe in the next section in detail.

Topics and Plan: The following will be a summary of the topics covered and time spent on each topic.

- Part I: Theory of Subset Selection (45 mins)
  - Submodular Optimization (15 mins)
  - Coresets (15 mins)
  - Combinatorial Information Measures and Optimization (15 mins)
- Part II: Applications (1 Hour, 30 Mins)
  - Compute Efficient Training of Deep Models (15 Mins)
  - Active Learning (15 Mins)
  - Targeted Subset Selection (15 mins)
  - Human Assisted Learning (15 mins)
  - Feature Selection and Model Compression (20 Mins)
  - Other Applications of Subset Selection to Semi-Supervised Learning, Meta Learning, and Weak Supervision (10 Mins)
- Part III: Hands-On Session (1 Hour 15 mins)
  - SubmodLib (15 Mins)
  - CORDS (15 Mins)
  - DISTIL (15 Mins)
  - APRICOT (15 mins)
  - TRUST (15 Mins)

Hands-On Session

A critical part of this tutorial will be a hands-on session where we will go over a few demos on subset selection algorithms and applications.

- SUBMODLIB: We will begin by going over submodlib (https://github.com/decile-team/submodlib) which implements a number of submodular functions and optimization algorithms (*e.g.*, variants of the greedy algorithm). We will also go over the submodular mutual information and conditional gain functions. We will review the utility of these in two concrete usecases, namely data subset selection for efficient training and data summarization. SUBMODLIB also serves as the backbone for the libraries below.
- CORDS: Next, we will also cover CORDS (https://github.com/decile-team/cords) to provide an overview of coresets and subset selection approaches for speeding up deep learning training.
- DISTIL: We will also cover DISTIL (https://github.com/decile-team/diltil) to provide an overview of active learning approaches for deep learning.
- APRICOT: We will also cover APRICOT (https://github.com/jmschrei/apricot) to provide an overview of submodular optimization for machine learning.
- TRUST: Finally, we will also cover TRUST (https://github.com/decile-team/trust) which uses subset selection for robust learning, fairness, and personalization.

In each of the cases above, we will prepare google colab notebooks and go over them in the tutorial. The hope is that after the tutorial, the participants can get hands-on experience and get started on using the stated libraries and tools.

Other Details

Estimated number of participants: Based on similar tutorials in the past, we estimate the size of the audience to be around 100. This is based on similar workshops which were organized at leading conferences (*e.g.*, ICML, NIPS, IJCAI, CVPR, etc.) and the recent SubSetML 2021 workshop at ICML 2021 which had close to 50 papers submitted and around 100 attendees.

Prerequisite knowledge: To ensure broad reach, we will make the tutorial as self-contained as possible. The only major pre-requisite knowledge is a reasonable understanding of machine learning and a basic understanding of deep learning and optimization. We expect most attendees of AAAI to have the prerequisite knowledge.

Supplementary Content: For completeness, we add the video links to some previous tutorials which cover similar topics.

IJCAI 2020 Tutorial by Rishabh and Ganesh: https://www.youtube.com/playlist?list=PLGod0_zT9w93RRj00LSVUJlOROmakvZTw

ISIT 2018 Tutorial by Jeff: https://www.youtube.com/watch?v=6TEneSHM6M0

ICML 2021 SubSetML Workshop: https://slideslive.com/icml-2021/subset-selection-in-machine-learning-from-theory-to-applications.

Similar Past Events

This is the first time this exact workshop is being organized. However, there are a few prior prior tutorials and workshops that are related. This is a testimony to the importance of this area. Furthermore, some of the tutorials and workshops below have run by a subset of the organizers. Most of the tutorials/workshops below have been on specific topic areas (*e.g.*, DPPs, submodularity, learning with less data, *etc*). In this tutorial, we will bring several of these sub-communities together under the umbrella of subset selection. *Furthermore, different from previous tutorials, this tutorial will have a heavy focus on applications and hands on components.* In fact, this is the first tutorial, which will have a significant hands-on component for practitioners to get their hands dirty with. Furthermore, many of the tutorial organizers have presented tutorials on related topics in the past (*e.g.*, Rishabh and Ganesh at IJCAI 2020 and ECAI 2020, Jeff at NIPS 2013, CVPR 2016, and ISIT 2018). Moreover, all four of the tutorial speakers co-organized the SubSetML 2021 workshop at ICML 2021 which was a grand success (around 50 submissions and 10 invited talks, and over 100 attendees of the workshop).

1. IJCAI 2020 and ECAI 2020 Tutorials: ”Combinatorial Approaches for Data, Feature and Topic Selection and Summarization” (co-organized by Ganesh Ramakrishnan and Rishabh Iyer)
2. ISIT 2018 tutorial on Submodularity in Information and Data Science by Jeff Bilmes
3. ICML 2013 Tutorial on Submodularity in Machine Learning co-organized by Stefanie Jegelka and Andreas Krause
4. NIPS 2013 Tutorial on “Deep Mathematical Properties of Submodularity with Applications to Machine Learning“ co-organized by Jeff Bilmes
5. CVPR 2019 Tutorial: ”Recent Advances in Visual Data Summarization”:https://rpand002.github.io/cvpr19_sumt.html
6. ICML 2020 Tutorial: ”Submodular Optimization: From Discrete to Continuous and Back”: https://icml.cc/virtual/2020/tutorial/5755
7. ICML 2021 Workshop: Subset selection in machine learning: https://sites.google.com/view/icml-2021-subsetml/home
8. ICML 2020 Workshop: ”Negative Dependence and Submodularity: Theory and Applications in Machine Learning”, https://icml.cc/Conferences/2020/Schedule?showEvent=5727
9. CVPR 2020 Workshop: Visual Learning with Limited Labels - https://www.learning-with-limited-labels.com/

This tutorial is NOT part of a continuing series and this is the first instance of this tutorial.

Organizers

Rishabh Iyer
Email: rishabh.iyer@utdallas.edu
Website: https://cs.utdallas.edu/people/faculty/iyer-rishabh/
Google Scholar: https://scholar.google.com/citations?user=l_XxJ1kAAAAJ&hl=en
Bio: Rishabh Iyer is currently an Assistant Professor at University of Texas at Dallas where he heads the Machine Learning and Optimization Lab. Prior to this, he was a Research Scientist at Microsoft where he spent three years. He finished his PostDoc and Ph.D from the University of Washington, Seattle. He has worked on several problems including discrete and submodular optimization, large scale data selection, robust and efficient machine learning, visual data summarization, active and semi-supervised learning. His work has received best paper awards at ICML 2013 and NIPS (now NeurIPS), 2013. He also won the Microsoft Ph.D fellowship, a Facebook Ph.D Fellowship, and the Yang Outstanding Doctoral Student Award from University of Washington. He has organized tutorials on summarization and data subset selection in WACV 2019, ECAI 2020 and IJCAI 2020.

Abir De:
Email: abir@cse.iitb.ac.in
Webpage: https://abir-de.github.io
Google Scholar: https://scholar.google.com/citations?user=_9ZKKbIAAAAJ
Bio: Abir is an Assistant Professor at the Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, India. Prior to this, he was a postdoctoral researcher at the Max Planck Institute for Software Systems. His main research interests broadly lie in the area of machine learning and its applications on human assisted learning and on networks. His current focus is modeling, learning and control of networked dynamical processes.

Ganesh Ramakrishnan
Email: ganesh@cse.iitb.ac.in
Webpage: https://www.cse.iitb.ac.in/~ganesh/
Google Scholar: https://scholar.google.co.in/citations?user=W1ZpREMAAAAJ&hl=en
Bio: Ganesh is currently serving as a Professor at the Department of Computer Science and Engineering, IIT Bombay and is Professor-in-charge of the Koita Centre for Digital Health, IIT Bombay. His areas of research include Human Assisted Machine Learning, symbolic representation in machine learning and computational linguistics in Indian languages. In the past, he has received awards such as IBM Faculty Award and research awards from Adobe

Research, Microsoft Research Award, Yahoo!, *etc.*, as well as IIT Bombay Impactful Research Award. He also held the J.R. Isaac Chair at IIT Bombay. Ganesh is very passionate about boosting the AI research eco-system for India and toward that, the research by him and his students as well as collaborators has resulted in a couple of startups which he continues to mentor. Ganesh has organized workshops at KDD 2014, COLING 2012 and has delivered tutorials at ECAI 2020, IJCAI 2020 (on summarization and data subset selection) and IJCAI 2007 (on graphical models). Ganesh has served a senior PC member for AAAI, IJCAI over the past few years and is also tutorial co-chair for PAKDD 2021.

Jeff Bilmes
Email: bilmes@uw.edu
Webpage: https://melodi.ece.uw.edu/people/bilmes
Google Scholar: https://scholar.google.com/citations?user=L9QufAsAAAAJ&hl=en
Bio: Jeffrey A. Bilmes is a professor at the Department of Electrical and Computer Engineering at the University of Washington, Seattle, where he leads the MELODI (MachinE Learning for Optimization and Data Interpretation) lab at UW. He received his Ph.D. from the department of Electrical Engineering and Computer Science, University of California in Berkeley and a masters degree from MIT. Prof. Bilmes is a 2001 NSF Career award winner, a 2002 CRA Digital Government Fellow, a 2008 NAE Gilbreth Lectureship award recipient, and a 2012/2013 ISCA Distinguished Lecturer. He was the program chair and general chair at UAI in 2009 and 2010 respectively, the tutorial chair (2011) and tutorial chair (2014) at NIPS/NeurIPS. Prof. Bilmes has also pioneered (starting in 2003) the development of submodularity within machine learning, and he received a best paper award at ICML 2013 and a best paper award at NIPS 2013 in this area. In 2014, Prof. Bilmes also received a most influential paper in 25 years award from the International Conference on Supercomputing, Prof. Bilmes is also founder and CEO of a company (Summary Analytics Inc. or smr.ai) that makes submodular optimization scalable to massive (e.g., peta-bytes) real-world data sets commonly found in the commercial world and makes using it quick, easy, and accessible for anyone. He has ran several workshops including the "Discrete Optimization in Machine Learning" workshop series at NIPS/NeurIPS from 2010 - 2014 and has given tutorials at NIPS/NeurIPS 2013, CVPR 2016 and ISIT 2018.

References

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Bachem, O.; Lucic, M.; and Krause, A. 2017. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*.

Das, S.; Iyer, R.; and Natarajan, S. 2021. A Clustering based Selection Framework for Cost Aware and Test-time Feature Elicitation. In *8th ACM IKDD CODS and 26th COMAD*, 20–28.

De, A.; Koley, P.; Ganguly, N.; and Gomez-Rodriguez, M. 2020. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2611–2620.

De, A.; Okati, N.; Zarezade, A.; and Gomez-Rodriguez, M. 2021. Classification Under Human Assistance. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Derezinski, M.; Khanna, R.; and Mahoney, M. W. 2020. Improved guarantees and a multiple-descent curve for Column Subset Selection and the Nystrom method. *Advances in Neural Information Processing Systems*, 33.

Dubey, A.; Chatterjee, M.; and Ahuja, N. 2018. Coreset-based neural network compression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 454–470.

Elenberg, E. R.; Khanna, R.; Dimakis, A. G.; Negahban, S.; et al. 2018. Restricted strong convexity implies weak submodularity. *Annals of Statistics*, 46(6B): 3539–3568.

Feldman, D. 2020. Core-sets: Updated survey. *Sampling Techniques for Supervised or Unsupervised Tasks*, 23–44.

Fujishige, S. 2005. *Submodular functions and optimization*. Elsevier.

Iyer, R.; Khargoankar, N.; Bilmes, J.; and Asanani, H. 2021. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, 722–754. PMLR.

Kaushal, V.; Iyer, R.; Kothawade, S.; Mahadev, R.; Doctor, K.; and Ramakrishnan, G. 2019. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1289–1299. IEEE.

Kaushal, V.; Kothawade, S.; Ramakrishnan, G.; Bilmes, J.; and Iyer, R. 2021. PRISM: A Unified Framework of Parameterized Submodular Information Measures for Targeted Data Subset Selection and Summarization. *arXiv preprint arXiv:2103.00128*.

Killamsetty, K.; Sivasubramanian, D.; Mirzasoleiman, B.; Ramakrishnan, G.; De, A.; and Iyer, R. 2021a. GRAD-MATCH: A Gradient Matching Based Data Subset Selection for Efficient Learning. *arXiv preprint arXiv:2103.00123*.

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; and Iyer, R. 2021b. GLISTER: Generalization based Data Subset Selection for Efficient and Robust Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kothawade, S.; Beck, N.; Killamsetty, K.; and Iyer, R. 2021. SIMILAR: Submodular Information Measures Based Active Learning In Realistic Scenarios. *arXiv preprint arXiv:2107.00717*.

Krause, A.; and Guestrin, C. 2008. Beyond convexity: Submodularity in machine learning. *ICML Tutorials*.

Kulesza, A.; and Taskar, B. 2012. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*.

Maheshwari, A.; Chatterjee, O.; Killamsetty, K.; Iyer, R.; and Ramakrishnan, G. 2020. Data Programming using

Semi-Supervision and Subset Selection. *arXiv preprint arXiv:2008.09887*.

Mariet, Z.; and Sra, S. 2015. Diversity networks: Neural network compression using determinantal point processes. *arXiv preprint arXiv:1511.05077*.

Mirzasoleiman, B.; Bilmes, J.; and Leskovec, J. 2020. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, 6950–6960. PMLR.

Mussay, B.; Osadchy, M.; Braverman, V.; Zhou, S.; and Feldman, D. 2019. Data-independent neural pruning via coresets. *arXiv preprint arXiv:1907.04018*.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8): 1226–1238.

Tohidi, E.; Amiri, R.; Coutino, M.; Gesbert, D.; Leus, G.; and Karbasi, A. 2020. Submodularity in action: From machine learning to signal processing applications. *IEEE Signal Processing Magazine*, 37(5): 120–133.

Wei, K.; Iyer, R.; and Bilmes, J. 2015. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, 1954–1963. PMLR.