Explainable Artificial Intelligence via Bayesian Teaching

Scott Cheng-Hsin Yang
Department of Mathematics & Computer Science
Rutgers University–Newark
Newark, NJ 07102
scott.cheng.hsin.yang@gmail.com

Patrick Shafto
Department of Mathematics & Computer Science
Rutgers University–Newark
Newark, NJ 07102
patrick.shafto@gmail.com

Abstract

Modern machine learning methods are increasingly powerful and opaque. This opaqueness is a concern across a variety of domains in which algorithms are making important decisions that should be scrutable. The explainabilty of machine learning systems is therefore of increasing interest. We propose an explanation-by-examples approach that builds on our recent research in Bayesian teaching in which we aim to select a small subset of the data that would lead the learner to similar conclusions as the entire dataset. We discuss this approach, explicating several key advantages. First, the ability to cover any model with a probabilistic interpretation including supervised, unsupervised, and reinforcement learning (including deep learning). Second, we discuss the empirical foundations of this approach in the cognitive science of learning from other agents. Third, we outline challenges to full realization of the promise of this approach. We conclude by discussing implications for machine learning and applications to real-world problems.

1 Introduction

Recent advances in machine learning have yielded remarkable gains in performance on important, hard learning problems. For instance, algorithms have recently shown remarkable success at learning to play ATARI games [1], automatically categorizing images [2], and even beating champions at the game of Go [3]. These machine learning methods are often built upon clever inference methods applied to models that posit enormous numbers of latent variables (e.g., on the order of 20,000 nodes for AlphaGo [4]). While the flexibility provided by these vast pools of latent variables allows these models to fit an enormous array of possible data, they also yield complex interdependence amongst the variables that renders the model inference difficult to interpret.

The opaqueness of the models is addressed in practice through a collection of ad hoc methods that attempt to coax understanding. These methods fall into a variety of categories. The first group are methods that are applied to otherwise opaque models. These include a host of techniques for visualizing layers of neural networks [5, 6] and automatic generation of text or captions to explain specific predictions in circumscribed domains [7, 8, 9]. A second group comprises machine learning models that are explicitly designed to be explainable by adopting structured, symbolic representations [e.g. 10], sometimes in combination with probabilistic inference [e.g. 11]. A third group treats the

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

behavior of the machine learning model as the target of training through optimizing the prediction of the model's output behavior in response to different inputs [e.g. 12]. A fourth group is to rely on technical experts to explain the model.

These existing methods for understanding the results of machine learning models have critical limitations. Ad hoc application of machine learning tools or visualizations may yield insights about the original model; however, it does not inform us about when to apply different methods or when or why we would expect these methods to mislead or fail. Approaches that are tailored to specific use cases, such as text generation, fail to provide the general principles that guide programmatic development of explainable AI. For many symbolic models, while the individual symbolic components may be interpretable, navigating through their interactions can be difficult. Moreover, inference in these models often yield a distribution over structures and distributions that do not avail themselves to explainability. Also, any approach which relies on human technical experts to explain results to human domain experts faces bottlenecks due to the limited number and high cost of such technical expertise. Moreover, the degree to which technical experts understand their own models is open to debate in light of recent demonstrations of the failure modes of popular models [13, 14].

A common theme of the first three groups of approaches is to use interpretable machine learning to explain opaque machine learning. These approaches avoid the core problem of understanding how learning and explanation are related. We argue that this relationship can be revealed by thinking of explanation as the inverse of modeling, presenting domain (but not necessarily technical) experts with the most relevant data for training the model. This view relies on the two fundamental observations that that all machine learning models is trained on data and that data is the natural common language between a user and any model. While explanation by examples alone do not capture the breadth of what philosophers and cognitive scientists may mean by explanation [15], it does provide a powerful subset that harnesses people's proclivities toward inductive inference. In light of this, we propose that Bayesian teaching, a method that samples example data to teach a model's inferences, is a general, model-agnostic way to explain a broad class of machine learning models. In the following sections, we will introduce Bayesian teaching along with the scope of its application (Section 2), present empirical support of the method (Section 3), raise challenges that will broaden, increase, and validate its utility (Section 4), discuss its relation to other works (Section 5), and make some concluding remarks (Section 6).

2 Bayesian teaching

Using examples as explanations is ubiquitous. Humans have the ability to induce principles and understandings from a few examples [16]. Thus, Bayesian teaching, an explanation-by-example method at heart, can help users understand the model in a general way and implicitly support user decisions on all levels, from those based on the model's individual decisions, to those based on conditions under which the model succeeds and fails. Furthermore, Bayesian teaching has its root in cognitive science and was developed to understand human inference in interactive settings [17, 18]. The chosen examples have been shown to match what humans find representative of the underlying generative process [19]. We have also shown how the chosen examples facilitate and explain human learning [20, 21]. As such, this framework fits well with the human-machine interaction inherent in explainable artificial intelligence, as explanation typically requires back-and-forth communication between the explainer and explainee.

In Bayesian teaching, the teaching problem is formalized as selecting a small subset of the data that will, with high probability, lead a learner model to the correct inference. The framework can be applied to any probabilistic model, and it uses whatever training data is input to the learning model. The equation for Bayesian teaching is [17, 22]

$$P_T(x|\Theta) = \frac{P_L(\Theta|x)P(x)}{\int_{x'} P_L(\Theta|x')P(x')dx'}, \tag{1}$$

where x can be any subset of the training data; Θ denotes the target model, which can be an entire model or particular substructures, such as latent features, relations, grammars, programs, or combinations of these; $P_T(x|\Theta)$ is the probability of choosing x as the teaching examples for explaining target model Θ; $P_L(\Theta|x)$ is the learner's posterior inference after receiving x; $P(x)$ describes the bias for certain kind of examples (e.g., favoring smaller subsets); and the integral is

over all partitions of the training data (i.e., if the size of x is m and the size of the entire training corpus is N, there are N-choose-m partitions).

2.1 Characterization of the scope of explainable models and problems

Bayesian teaching can be applied to any model that can be cast as Bayesian inference. One can characterize the scope of these models based on the way that they are typically introduced. This characterization carves the scope into three types of model: (1) all generative models, for which the probabilities on data and model variables are explicitly defined; (2) models that are first introduced in a decision-theoretic framework with loss functions, which were later interpreted as negative log likelihood and negative log prior; and (3) non-Bayesian models that are extended to become Bayesian. Thus, the scope of Bayesian models is comprehensive [cf. 23] and covers all areas of machine learning.

Below, we survey the problem areas of supervised, unsupervised, reinforcement, and deep learning, first describing what x and Θ corresponds to, then providing notable examples of each of the three types of models mentioned in the previous paragraph. Note that deep learning is really a class of models rather than a problem area; however, we will treat it as a separate branch of machine learning because it is currently the dominant approach, as well as arguably the most opaque class of models.

For supervised learning, the x in Eq. 1 is examples and labels, and the Θ is typically the weights that determined the shape of the regression function or the classification boundary. A powerful and expanding class of supervised generative models is those based on Gaussian process [24]. Discriminative models are usually defined by a loss function. Interestingly, almost all of them turn out to have a Bayesian interpretation, including the popular support vector machine [25, 26]. Bayesian adaptive regression trees [27] and Mondrian forests [28] can be considered as Bayesian version of random forests, a popular black-box model that does not have clear Bayesian interpretation because of its algorithmically based development.

For unsupervised learning, x is the examples, and Θ is typically the latent structures, such as the parameters of the hidden mixtures in clustering, topics in models of text semantics, and states and transition matrix in time-series analysis. Models based on the Dirichlet process, such as infinite Gaussian mixture models [29] and latent Dirichlet allocation [30], form a powerful and expanding class of generative models. Principle component analysis is an example of the second type where the probabilistic interpretation is discovered many years after its introduction [31]. Probabilistic latent semantic analysis [32] and probabilistic matrix factorization [33] are examples of probabilistic versions of popular, non-Bayesian models useful for recommender systems.

For reinforcement learning, x is taken from the history of the actions, observations (or states visited), and rewards experienced by the agent, and Θ is the learned policy and the learned model of the world. Reinforcement learning algorithms come in two flavors: model-based and model-free. On the side of model-based learning, the problem of finding an optimal policy has a Bayesian interpretation and can be translated exactly into a Bayesian inference problem in both deterministic and partially-observable Markov-decision process settings [34, 35]. The translation is done by constructing a generative model for the world model under consideration. On the model-free side, two popular, non-Bayesian algorithms for policy improvement—temporal difference and State-Action-Reward-State-Action algorithm—have been made Bayesian by putting a Gaussian-process prior on state-action values [36]. These are examples of the third type.

For deep learning, x is again the training examples, and Θ is the weights of deep network. Three well-known generative deep models are deep belief network [37], deep Boltzmann machines [38], and deep Gaussian process [39]. Two well-known discriminative deep models are deep convolutional and recurrent neural networks [40]. They are currently the best machine learning algorithms for many supervised learning tasks. Recent work has shown how the modern way of training these deep networks relates to performing variational approximation for deep Gaussian process, shedding light on their Bayesian interpretation [41]. The implications is that a fully Bayesian way of training deep Gaussian process will make it the Bayesian extension of these deep networks. Such training has recently seen a big breakthrough and has allowed deep Gaussian process to be applied to large-scale regression problems for the first time [42].

In summary, the scope of Bayesian models is extensive and covers all areas of machine learning. Furthermore, Bayesian models are expanding quickly in not only machine learning, but also in

neuroscience and cognitive science. Since Bayesian teaching is applicable to any model that can be cast Bayesian, it will have far-reaching impact in advancing explainable artificial intelligence.

3 Empirical support

The Bayesian teaching approach is rooted in the cognitive science of teaching and learning. The cognitive science literature has investigated this approach under the banner of pedagogical reasoning. [17] investigated teaching and learning in a simple rule-based concept learning setting in which concepts were axis-aligned rectangles on a game board and examples were points labeled as inside or outside the true concept [cf. 43]. They found that participant's choices of examples when teaching, and inferences about concepts when learning, were well characterized by a model that assumed mutual cooperation [17]. [18] extended this research to explore human teaching and learning of prototype categories (Gaussian distributions; [cf. 19]) and causal networks. [44] found that children, after learning from a knowledgeable and helpful informant, showed reduced exploratory play, as was predicted if they assumed the examples were selected to teach [see also 45], a pattern that was not observed after learning from a naive informant. This work focused on a model predicated on cooperation between the teacher and the learner, formalized as a recursive reasoning between the teacher and learner.

The Bayesian teaching approach described in Section 2 is a one-step approximation of the pedagogical reasoning model described in [17, 18]. [19] proposed a similar formalism to characterize people's judgments of representativeness of examples with respect to a category. This model considered categories individually, modeled as Gaussian distributions in a multidimensional similarity space. Similar approaches has been adopted in the language literature to explain modifications in the grammatical structure of language input to children [46], and capture the effects of pragmatics on speakers and listeners [47].

From a modeling perspective, Bayesian teaching requires computing the marginal likelihood of every candidate subset of data. For this reason, the empirical work described above has been largely contained to highly circumscribed contexts with simple stimuli and concepts. Scaling models beyond these simple contexts to more realistic problems of the kind one would encounter when explaining machine learning models is a challenging computational problem. We have begun to address this problem by leveraging Markov Chain Monte Carlo (MCMC) techniques to approximate the marginal likelihood via sequential importance sampling [21, 22]. The approach is general and scalable and has been applied to infinite Gaussian mixture model (IGMM) for unsupervised clustering problems [21, 20] and to latent Dirichlet allocation (LDA) for topic modelling [22].

In [21], we investigated infant-directed speech—the curious manner in which we talk to infants—as optimal input examples for teaching phonetic categories of the language. Previous results had shown a confusing pattern of results: whereas the space occupied by vowel categories increased in infant-directed speech, some vowel categories actually get closer together and some increase in (co)variance. While the first feature would seemingly make learning vowel categories easier, the others do not seem consistent with that possibility. We modelled infant's clustering of phonetic sounds as an IGMM and applied Bayesian teaching to select sounds that best help the infant learn to infer the phonetic categories found in adult speech. We showed that the sounds selected by this approach are consistent with infant-directed speech. This consistency suggests that infant-directed speech may be a means of teaching infants about the adult language, and that Bayesian teaching aligns with human behavior in a complex, real-world domain.

In [20], we developed a pipeline that automated the quantification, analysis, and presentation of data to teach naive human users about novel statistical patterns in data. This pipeline took in a corpus of images, automatically inferred visual categories from the corpus using IGMM, and applied Bayesian teaching to output a small set of images as the best summary to explain the extracted categories. Naive human learners (on MTurk) who saw these teaching examples performed better in a categorization task (i.e. more accurately inferred model's categories) than those who, for example, saw the highest likelihood examples. This result suggests that Bayesian teaching can help naive human users to make better predictions about a model's behavior.

In summary, Bayesian teaching has been shown to match the way human teach human in a variety of perceptual and cognitive domains. When used as an automatic summarizing tool, it has been shown

to push human's judgment to align with those of machine learning models. These results are evidence that Bayesian teaching can be used to transmit the knowledge of a machine learning model to human.

4 Challenges

Although there have been some advances in making Bayesian teaching more general and scalable as discussed in the previous section, there remains many challenges for making Bayesian teaching a general and useful tool for explaining artificial intelligence. In this section, we list what we feel are the essential challenges.

4.1 Model expressivity

The first challenge is to extend Bayesian teaching to a broader collection of models as well as more expressive models such as deep Gaussian processes models. One general approach toward this goal is to integrate Bayesian teaching with probabilistic programming. The two main, practical benefits of probabilistic programming are that inference is largely automated once the model to infer is specified and that complex models are typically much easier to specify. Such integration makes Bayesian teaching a general procedure, just like inference, to automate the example selection process. Of course, general-purpose approximations that are both efficient and accurate is a challenging problem, but recent headway in the machine learning literature may suggest fruitful approaches [e.g. 48].

4.2 Explanation expressivity

The second challenge is to make Bayesian teaching capable of selecting multiple examples that condition on the model's substructures. Providing sets of examples will allow the users to leverage comparison and diversity of the examples to understand inferences about complexity and uncertainty. Conditioning on model substructures will allow the users to tease apart the explanatory role of different aspects of the model by decomposing the model's behavior based on the model's substructure. This is useful because model substructures appear across modeling frameworks. In the graphical model framework, a model substructure is any inferred variable. Similarly, in the probabilistic programming framework, a model substructure is any function that contains inferred variables. In deep learning, a model substructure is any hidden layer.

The main technical difficulty is that the number of partitions, or the number of ways x can be chosen, grows combinatorially with the size of x. This implies two problems: first, one will need to compute the numerator in Eq. 1 many more times; second, the space of x to explore expands drastically. A first line of attack may involve using a combination of variational inference [49], advanced MCMC methods [50, 51, 52], and active learning [53]. As another route, there are promising paths for tractable inference which considers subclasses of models that have geometric [54] and topological [55] properties that can be exploited to provide efficient, approximate inference.

4.3 Interface and evaluation

Explanation is rarely a one-shot event but typically an interactive process between the explainer and explainee. Thus, an effective interface is inherently a part of the problem of explanation. A good interface may want to provide the following two aspects of functionality: first, a visualization of the summary of the training data that accommodates a broad array of data types, including numerical (continuous and categorical), text, images, and video; and second, an interactive interface to query additional examples conditioned on user-chosen aspects of the model's substructure and to submit user-curated examples to elicit model predictions. This functionality will allow users to interactively explore the model and its component parts, as a means of understanding the relatively roles of different aspects of the model with respect to different aspects of the data.

Because the goal of explainable artificial intelligence is to enhance human understanding, empirical evaluation of the framework with human experiment is essential. Following [56], we suggest that evaluation should be done to test whether explanations can help user: 1) predict the model's predictions, 2) understand why the model makes predictions the way they do, 3) understand when the model would fail, 4) develop trust towards the model, and 5) know how to correct the model. Toward these goals, it would be useful to leverage research in the cognitive science literature, which has

developed a rich set of methods for assessing the structure of people's concepts and their implications for future learning. A potential challenge is that the cognitive science literature has mainly focused on highly controlled settings, while the goal of explainable machine learning is inherently about real-world data. This raises questions about how to assess the efficacy of an explanation or a system that provides explanations when there may not be a ground truth. Two possibilities are to consider whether specific explanations do or do not support identification of similar cases, and, for the purposes of testing, to modify the data to ask whether system-level explanations facilitate the discovery of these induced errors.

5 Related work

The connection that we have drawn between teaching models and explainable artificial intelligence has far-reaching implications in that any teaching model that transmits knowledge through data can in principle be used for generating explanation. In this section, we mention several kinds of these teaching models, including Machine Teaching, coresets, algorithmic teaching, and inverse reinforcement learning.

Machine teaching is typically framed as an optimization problem where the loss function comprises of a term for how well the data selected will induce the target model and another term on the effort required to select those data [57]. It has been successfully applied to Bayesian learning models in the exponential family [58] and has guarantees for linear learning models [59]. It differs from Bayesian teaching in that it aims to deterministically select the best examples.

The method of coresets has its roots in computational geometry [60, 61]. Like machine teaching, it also involves a loss function that takes the data selected and target model as inputs, but instead of optimizing the data selection, any dataset within a tolerance loss level is considered a coreset. Coresets has been applied to many learning models, including large-scale dimension reduction [62] and mixture models [63] (see [61] for a recent review). It differs from Bayesian teaching mainly in its framing, but its construction is often done by sampling as well.

Algorithmic teaching is focused on a theoretical perspective that has roots in the algorithmic learning literature. These approaches focus on identifying teaching sets, which are sets of examples that rule out all concepts other than the target [64, 65]. Algorithmic teaching has mainly been focused on concept learnability and quantifying the example-based complexity of learning via teaching dimension [66]. It differs from Bayesian teaching in focusing on theoretical results and in the deterministic nature of the models.

Recent research has also leveraged inverse reinforcement learning (IRL) to address educational issues. For example, recent research has investigated performance of IRL-based approaches to guiding inquiry [67] and to personalization [68]. More recent research has developed approaches that model learner's beliefs by forming policies over actions that include providing data and quizzing the learner [69]. These approaches differ in their focus on education and in adopting a planning perspective.

6 Conclusion

The development of new, more-opaque machine learning methods has led to revolutionary advances in learning, but at the cost of explainability. This paper argues that Bayesian teaching is a way to remove the current tension between powerful learning and explainability. The Bayesian teaching approach explicitly builds from foundational results in cognitive science. It leverages the natural common language for understanding model behavior—the data—to explain opaque models through the examples from the original data that are most representative of the inference. In doing so, it *integrates* learning and explanation by taking the learning model as input into the explanation process, and outputs an explanation in terms of the examples from the original data.

Because the approach is phrased in the language of probabilistic inference, the explanation applies to any model that can be cast as Bayesian inference, which covers many of the most influential classes of machine learning models including deep learning, generative and discriminative approaches, and probabilistic programs. Furthermore, it is model-agnostic in nature and hence a very general approach. It takes the model together with the data as the object to be understood, and by formalizing explanation as teaching examples, the explanations become model-free, and thus no special modeling

expertise is required to understand the explanation that was not already required to understand the data. Such model-free explanation by examples is a key strength which may significantly increase the usefulness of machine learning methods.

Acknowledgments

This material is based on research sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-17-2-0146 to P.S. and S.Y. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[3] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[4] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[5] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341, 2009.

[6] Alexander Mordvintsev, Michael Tyka, and Christopher Olah. Deepdream. `https://github.com/google/deepdream`, 2015.

[7] Hui Cheng, Jingen Liu, Saad Ali, Omar Javed, Qian Yu, Amir Tamrakar, Ajay Divakaran, Harpreet S Sawhney, R Manmatha, James Allan, et al. Multimedia event detection and recounting. *SRI-Sarnoff AURORA at TRECVID 2014*, 2014.

[8] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.

[9] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. *arXiv preprint arXiv:1603.08507*, 2016.

[10] Marta Garnelo, Kai Arulkumaran, and Murray Shanahan. Towards deep symbolic reinforcement learning. *arXiv preprint arXiv:1609.05518*, 2016.

[11] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *arXiv preprint arXiv:1602.04938*, 2016.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[14] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436. IEEE, 2015.

[15] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

[16] John Stuart Mill. *A system of logic ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Harper, 1884.

[17] Patrick Shafto and Noah D. Goodman. Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th annual conference of the Cognitive Science Society*, Austin, TX, 2008. Cognitive Science Society.

[18] Patrick Shafto, Noah D. Goodman, and Thomas L. Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55–89, 2014.

[19] Joshua B Tenenbaum, Thomas L Griffiths, et al. The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the Cognitive Science Society*, page 103641. Citeseer, 2001.

[20] April Schweinhart, Baxter S. Eaves, and Patrick Shafto. Automating the recoding, analysis, and interpretation pipeline using naturalistic visual scenes. In *IJCAI 2016 workshop on Closing the Cognitive Loop*, 2016.

[21] Baxter S. Eaves, Naomi H. Feldman, Thomas L. Griffiths, and Patrick Shafto. Infant-Directed Speech Is Consistent With Teaching. *Psychological Review*, 2016.

[22] Baxter S. Eaves and Patrick Shafto. Toward a general, scalable framework for Bayesian teaching with applications to topic models. In *IJCAI 2016 workshop on Interactive Machine Learning*, 2016.

[23] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[24] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. the MIT Press, 2006.

[25] Nicholas G Polson, Steven L Scott, et al. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

[26] Vojtech Franc, Alexander Zien, and Bernhard Schölkopf. Support vector machines as probabilistic models. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 665–672, 2011.

[27] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, pages 266–298, 2010.

[28] Balaji Lakshminarayanan, Daniel M Roy, and Yee Whye Teh. Mondrian forests: Efficient online random forests. In *Advances in neural information processing systems*, pages 3140–3148, 2014.

[29] Carl Edward Rasmussen. The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

[30] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[31] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[32] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[33] Ruslan R Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1257–1264. Curran Associates, Inc., 2008.

[34] Marc Toussaint and Amos Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the 23rd international conference on Machine learning*, pages 945–952. ACM, 2006.

[35] Finale Doshi-Velez, David Pfau, Frank Wood, and Nicholas Roy. Bayesian nonparametric methods for partially-observable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):394–407, 2015.

[36] Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 201–208. ACM, 2005.

[37] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[38] Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In *AISTATS*, volume 1, page 3, 2009.

[39] Andreas C. Damianou and Neil D. Lawrence. Deep gaussian processes. *AISTATS*, 2013.

[40] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[41] Yarin Gal and Zoubin Ghahramani. On modern deep learning and variational inference. *Advances in Approximate Bayesian Inference workshop, NIPS*, 2015.

[42] Thang D Bui, Daniel Hernández-Lobato, Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. Deep gaussian processes for regression using approximate expectation propagation. *arXiv preprint arXiv:1602.04133*, 2016.

[43] Solla S. Kearns, M. and D. Cohn. Bayesian modeling of human concept learning. In *Advances in neural information processing systems*, pages 59–65, 1999.

[44] Elizabeth Bonawitz, Patrick Shafto, Hyowon Gweon, Isabel Chang, Sydney Katz, and Laura Schulz. The double-edged sword of pedagogy: Modeling the effect of pedagogical contexts on preschoolers' exploratory play. *Proceedings of the 31st annual conference of the Cognitive Science Society*, 2009.

[45] Elizabeth Bonawitz, Patrick Shafto, Hyowon Gweon, Noah D Goodman, Elizabeth Spelke, and Laura Schulz. The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330, 2011.

[46] Anna Rafferty and Thomas Griffiths. Optimal language learning: The importance of starting representative. In *Proceedings of the Cognitive Science Society*, volume 32, 2010.

[47] N.D. Goodman and A. Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5, 2013.

[48] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[49] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

[50] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. Dram: efficient adaptive mcmc. *Statistics and Computing*, 16(4):339–354, 2006.

[51] Dougal Maclaurin and Ryan P Adams. Firefly monte carlo: Exact mcmc with subsets of data. *arXiv preprint arXiv:1403.5693*, 2014.

[52] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[53] Michael A. Osborne, David Duvenaud, Roman Garnett, Carl Edward Rasmussen, Stephen J. Roberts, and Zoubin Ghahramani. Active learning of model evidence using Bayesian quadrature. In *Advances in Neural Information Processing Systems 25*, pages 46–54, 2012.

[54] Dan Feldman, Matthew Faulkner, and Andreas Krause. Scalable training of mixture models via coresets. In *Advances in neural information processing systems*, pages 2142–2150, 2011.

[55] Matthew T Pratola et al. Efficient metropolis–hastings proposal mechanisms for bayesian regression tree models. *Bayesian Analysis*, 2016.

[56] David Gunning (Program Manager). Broad agency announcement (darpa-baa-16-53): Explainable artificial intelligence (xai). Technical report, DARPA, 2016.

[57] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, pages 4083–4087, 2015.

[58] Xiaojin Zhu. Machine teaching for bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems*, pages 1905–1913, 2013.

[59] Ji Liu and Xiaojin Zhu. The teaching dimension of linear learners. *Journal of Machine Learning Research*, 17(162):1–25, 2016.

[60] Dan Feldman. *Coresets and Their Applications*. Tel Aviv University, 2010.

[61] Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

[62] Dan Feldman, Mikhail Volkov, and Daniela Rus. Dimensionality reduction of massive sparse datasets using coresets. In *Advances in Neural Information Processing Systems*, pages 2766–2774, 2016.

[63] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training mixture models at scale via coresets. *arXiv preprint arXiv:1703.08110*, 2017.

[64] Frank J Balbach. Measuring teachability using variants of the teaching dimension. *Theoretical Computer Science*, 397(1-3):94–113, 2008.

[65] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Teaching dimensions based on cooperative learning. In *COLT*, pages 135–146, 2008.

[66] Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, vc-dimension and sample compression. *Journal of Machine Learning Research*, 15(1):3107–3131, 2014.

[67] Libby F Gerard, Kihyun Ryoo, Kevin W McElhaney, Ou Lydia Liu, Anna N Rafferty, and Marcia C Linn. Automated guidance for student inquiry. *Journal of Educational Psychology*, 108(1):60, 2016.

[68] Anna N Rafferty, Rachel Jansen, and Thomas L Griffiths. Using inverse planning for personalized feedback. In *EDM*, pages 472–477, 2016.

[69] Anna N Rafferty, Emma Brunskill, Thomas L Griffiths, and Patrick Shafto. Faster teaching via pomdp planning. *Cognitive science*, 40(6):1290–1332, 2016.