Automated cognitive modeling with Bayesian active model selection

Abstract

Behavioral experiments are often feed-forward: they begin with designing the experiment, and proceed by collecting the data, analyzing it, and drawing inferences from the results. Active learning is an alternative approach where partial experimental data is used to iteratively design subsequent data collection. Here, we study experimental application of Bayesian Active Model Selection (BAMS), which designs trials to discriminate between a set of candidate models. We consider a model set defined by a generative grammar of Gaussian Process kernels that can model both simple functions and complex compositions of them. To validate the method experimentally, we use BAMS to discover how factors such as contrast and number affect numerosity judgements. We compare the rate of convergence of the active-learning method to a baseline passive-learning strategy that selects trials at random. Active learning over a structured model space may increase the efficiency and robustness of behavioral data acquisition and modeling.

Keywords: Bayesian active model selection; Bayesian statistics; active model selection; active learning; numerosity;

Background

The history of active learning is rich in its implementation across various scientific domains, from cancer and genomics, to pedestrian detection in video surveillance systems (?, ?). There has also been major efforts in studying the computational optimization of active learning querying strategies (?, ?). However, there is room for further studies in applying active learning and active model selection to behavioral experiments in psychology and cognitive science. Behavioral experiments often require gathering data that may be difficult, expensive, and time consuming to obtain. Using strategies to optimize experimental efficiency can help mitigate these issues in the behavioural sciences. BAMS constructs a learner that maintains multiple hypotheses with differing degrees of confidence. These models construct a matrix of representations of various degrees of belief across a broad range of hypotheses. The learner then probes a participant by changing experiment variables, in order to fit, select, and filter for the model with the highest confidence, while reducing the degree of uncertainty in models that do not fit to the data. (?, ?).

The plan of the paper is as follows: We first describe a few common use cases of active learning in behavioral academic literature, and we speculate how this model might further benefit behavioral experiments. Next, we examine the abundance of numerosity behavioral research, and introduce why this experiment is suitable for Bayesian active model selection. We will describe our specific experiment design, the experimental manipulations, as well as which models each experiment and strategy discovered. Here we will compare our participant results, and the results described in the literature to the model generated results in order to determine the model's efficiency in describing behavior. We conclude with a discussion on the potential of BAMS, as well as improvements that can be made to this technique.

Active learning in psychology

Active learning has successfully been used in a number of classic behavioral experiments. One experiment which examined a method similar to Bayesian active model selection in psychology used a method called QUEST to model the human psychometric function, or the strength of psychophysical responses to stimuli. The experiment used active, or "adaptive" learning, because of its time and resource efficient advantages, using prior psychological literature on the shape of the psychometric function to quickly find the threshold of that function. The experimenters assumed, based on prior research, that the psychometric function remains the same in human behavioral experiments and can be modeled with the function of log intensity, differing only in position defined by a threshold parameter. During the experiment, a posterior probability function is used to determine confidence. It contains information about the threshold given a participant's feedback from prior trials. This is continuously used to determine how to construct the proceeding trial to best approximate the psychometric function's threshold (?, ?).

Another popular paradigm for decision-based behavioral studies is the multi-armed bandit problem. This behavioral experiment presents a scenario where a participant is tasked with maximizing profit by making a number of decisions, each which have an unknown payoff distribution. The experiment's name is taken from the nickname "single-arm bandits", given to slot machines due to their tendency to unsuspectingly drain money from its players over time. The multi-arm bandit presents a problem where a participant can choose from a number of slot machines, or slot machine-like options, each with different distributions of reward. The participants must then make decisions based off limited information on how to optimally determine the arm, or combination of arms which return the greatest reward. Optimal solutions are difficult to compute due to the large amount of uncertainty in

these problems (?, ?). This sort of uncertainty benefits greatly from using an active learning strategy, which can iteratively use feedback from the slot machines to quickly rule out, or gain confidence in strategic models. One popular solution that does this is Bayesian Bandits, or Thompson sampling, which begins with a prior distribution for each belief and takes the average expected reward for each arm, then plays the arm which has the highest expected reward for a denoted duration of time. It attempts to balance the trade-off of maximizing immediate reward, while still allowing search exploration of potential rewards.

Bayesian Active Structure Learning

Suppose we face a regression problem defined on an input space $\mathcal{X}$ and output space $\mathcal{Y}$. Our goal is to model the relationship between these two spaces using a set of training observations $\mathcal{D} = (\mathrm{X}, \mathrm{y})$, where X represents the design matrix of explanatory variables $\mathrm{x}_i \in \mathcal{X}$, and $y_i \in \mathcal{Y}$ is the respective output response to be predicted.

Usually, the experimenters use their own expertise to choose a (probabilistic) model $\mathcal{M}$ suitable to explain the phenomenon under investigation. During the data analysis, they might realize that their first guess was not ideal, and, as a consequence, they will need to invest more time in model selection. In practice, this model-investigation-cycle might occur many times, which is not only inefficient but also hard to replicate.

In the following, we describe a methodology that makes model selection both: 1) *explicity* with respect to the set of models being considered and 2) emphactive, meaning, it leverages the potential to design experimental configurations to reveal the most information about which model best explains the relationship between $\mathcal{X}$ and $\mathcal{Y}$.

Model class We assume that our observations were generated according to a latent function $f\colon \mathcal{X} \to \mathbb{R}$ via a fixed probabilistic observation mechanism $p(\mathrm{y} \mid \mathrm{f})$, where $f_i = f(\mathrm{x}_i)$. A standard nonparametric Bayesian approach is to consider regression problems with additive Gaussian observation model, where we have placed a Gaussian process prior distribution on f. Formally, a GP is a (potentially infinite) collection of random variables such that the joint distribution of any finite number of them is multivariate Gaussian distribution.

Similarly to the multivariate Gaussian distributions, a GP on f can be fully specified by its first two moments: $p(f) = \mathcal{GP}(f; \mu_f, K_f)$, where $\mu_f\colon \mathcal{X} \to \mathbb{R}$ is a mean function (first moment) and $K_f\colon \mathcal{X}^2 \to \mathbb{R}$ is a positive-definite covariance function or kernel, representing the second moment. We can also condition on the observed data to form a posterior distribution $p(f \mid \mathcal{D})$, which is typically an updated Gaussian process. Finally, we make predictions at a new input x^* via the predictive distribution $p(y^* \mid \mathrm{x}^*, \mathcal{D}) = \int p(y^* \mid f^*, \mathcal{D}) p(f^* \mid \mathrm{x}^*, \mathcal{D}) \, \mathrm{d}f^*$, where $f^* = f(\mathrm{x}^*)$. For a complete review on Gaussian processes see (?, ?).

Each choice of mean function μ_f and covariance function (or kernel) K_f defines a specific probabilistic model $\mathcal{M}$. Both μ_f and K_f might have hyperparameters which are concatenated in a single vector $\theta \in \Theta_{\mathcal{M}}$, where $\Theta_{\mathcal{M}}$ is the corresponding parameter space of the model. For simplicity, we will assume that all models have the same zero-mean function, making the kernel choice the crux of model selection. Notice that different kernels might be used to model different trends of data such as such linearity, periodicity or function values smoothness. Essentially, several *structural* assumption about the data can be modeled by kernels.

Active model selection In Bayesian active structure learning, we wish to automatically identify the most likely model to be the source of data. We will frame this problem as a special case of active machine learning.

We begin describing how to compare models using the Bayesian framework. First, we quantify the uncertainty over models using the *model evidence*, the probability of generating the observed data given a model $\mathcal{M}$:

$$p(\mathrm{y} \mid \mathrm{X}, \mathcal{M}) = \int_{\Theta_{\mathcal{M}}} p(\mathrm{y} \mid \mathrm{X}, \theta, \mathcal{M}) \, p(\theta \mid \mathcal{M}) \, \mathrm{d}\theta. \qquad (1)$$

The evidence (also called *marginal likelihood*) integrates over θ to account for all possible explanations of the data offered by the model, under a prior $p(\theta \mid \mathcal{M})$ associated with that model.

Given a set of candidate models $\{\mathcal{M}_i\}_{i=1}^{M}$, and the computed evidence for each, we apply Bayes' rule to compute the *posterior probability* of each model given the data:

$$p(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathrm{y} \mid \mathrm{X}, \mathcal{M}) p(\mathcal{M})}{p(\mathrm{y} \mid \mathrm{X})} = \frac{p(\mathrm{y} \mid \mathrm{X}, \mathcal{M}) p(\mathcal{M})}{\sum_i p(\mathrm{y} \mid \mathrm{X}, \mathcal{M}_i) p(\mathcal{M}_i)}, \qquad (2)$$

where $p(\mathcal{M})$ represents the prior probability distribution over the models.

Now, we want to use an active learning strategy to select the new data pair—$\mathrm{x}^* \in \mathcal{X}$ and respective label $y^* = y(\mathrm{x}^*)$—to add to our dataset $\mathcal{D} = (\mathrm{X}, \mathrm{y})$, in order to better distinguish the candidate models $\{\mathcal{M}_i\}_{i=1}^{M}$. After making this observation, we will form an augmented dataset $\mathcal{D}' = \mathcal{D} \cup \{(\mathrm{x}^*, y^*)\}$, from which we can recompute a new model posterior $p(\mathcal{M} \mid \mathcal{D}')$. Our strategy for actively selecting observations is the Bayesian active model selection (BAMS) (?, ?), which maximizes the *mutual information* between the new observation value y^* and the unknown model:

$$I(y^*; \mathcal{M} \mid \mathrm{x}^*, \mathcal{D}) = H[y^* \mid \mathrm{x}^*, \mathcal{D}] - \mathbb{E}_{\mathcal{M}}\big[H[y^* \mid \mathrm{x}^*, \mathcal{D}, \mathcal{M}]\big], \qquad (3)$$

where H indicates (differential) entropy. Other active model selection approaches were proposed but BAMS was shown to be sample-efficient and less computation demanding than previous approaches (?, ?).

Numerosity

Numerosity experiments have been extensively studied in behavioral psychology experiments that investigate both human

and animal perception. It is considered by some to be an elementary quality of perception, and both animal and human research suggests dedicated brain structures for numerosity perception. It is generally accepted that the ability to perceive a number of objects is an important survival skill for many animal species. Humans can estimate numerosity at a young age during development, and accuracy is also thought to further improve with age up until age 30 (?, ?).

Popular numerosity experiments have examined the role of density and area influencing numerosity perception (?, ?) and examining the effects of size adaption on human numerosity judgements (?, ?). This research makes the numerosity experiment a good candidate for BAMS due to well-studied behavioral effects, as well as the general simplicity of the experiment implementation and manipulation.

Experiment Method

We applied BAMS as a method for model discovery in a simple numerosity experiment. We selected so that the active learner could manipulate numerosity stimuli presented to participants, in order to improve the search for the kernel grammar with the highest confidence of modeling the behavioral data.

Experiment stimuli

We constructed an experiment where participants were presented n number of dots, where $1 \leq n \leq 100$. These dots were randomly distributed across a 250 x 250 pixel window using the open source software Psychopy (?, ?). The participant was presented with these dots for a three-second period, after which, they were presented with a screen to estimate how many dots they saw using their keyboard. After the user provided their feedback, the learning strategy had the opportunity to manipulate certain experiment parameters for every trial, depending on the type of strategy and hyper-parameters defined, with the ultimate goal of learning the relationship between experiment manipulations and participant predictions.

Throughout each trial, the BAMS method used a set of defined kernels to mathematically model how each of the tested dimensions effected the participant's prediction of n number of dots; updating the posterior probability, or confidence, of each model as the experiment progressed. BAMS used each participant's input as feedback to update every model, allowing the learner to optimize each decision for the subsequent trial based on previously seen data. Models were constructed using predefined base kernels, or a combination of these base kernels. These combinations, known as kernel grammars, used the summation and/or multiplication of the base kernels as the foundation for learning (?, ?).

Learner hyper-parameters

We tested two different learning strategies, Bayesian active learning by disagreement (BALD) and a random strategy learner (?, ?). We defined a certain set of hyper-parameters which were used for both strategies, and these remained the same hyper-parameters in every experiment. These included the budget, pool size, base kernels, and depth. The budget, or resource limitation on iterations was set to 50, which defined the number of trials (?, ?). The learner also contained a predefined set of base kernels used to construct the kernel grammar which included the exponential (SE), periodic (PER), and linear (LIN) kernels. The learner was also given a depth of 2, which was used to specify how deep or shallow the manipulations to the kernel grammar extended, while the pool size was set to 200 to define the width of the exploration space.

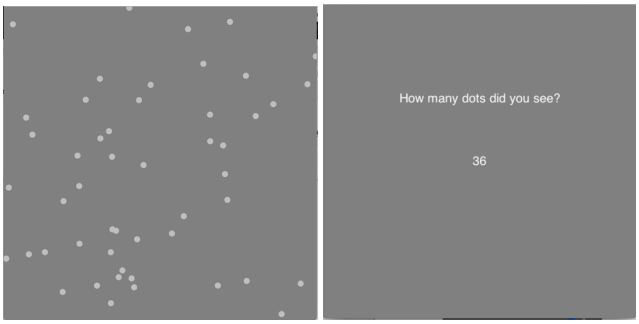

Figure 1: Experiment stimuli

Experiment matrix

We set up a 2 x 4 experiment matrix using the two strategies across four separate dimension manipulations. The BALD and random strategies were used in four types of experiments that manipulated either (1) the n number of dots, (2) the dot contrast, (3) a dummy variable, or (4) all three aforementioned manipulations simultaneously. Each experiment consisted of a total of 50 trials where the participant was iteratively given the numerosity task. The total experiment runs were 8. The BALD strategy's performance was recorded against our control, the random strategy.

In figures 2, 3, and 4 we analyze each separate dimension, where we manipulate the number of dots in every trial, the contrast density, and a dummy variable, respectively. We see that both the BAMS and Random Strategy learners converge to a Linear Kernel on contrast, and the contrast density. On the dummy variable, we see convergence to the Constant Kernel.

Acknowledgments

This work was supported in part by DARPA NGS2 cooperative agreement D17AC00004.

References