Effects of generative and discriminative learning on use of category variability

Anne S. Hsu (ahsu@gatsby.ucl.ac.uk)
Department of Cognitive, Perceptual and Brain Sciences, University College London, London, UK

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720 USA

Abstract

Rational models of category learning can take two different approaches to representing the relationship between objects and categories. The generative approach solves the categorization problem by building a probabilistic model of each category and using Bayes' rule to infer category labels. In contrast, the discriminative approach directly learns a mapping between inputs and category labels. With this distinction in mind, we revisit a previously studied categorization experiment that showed people are biased towards categorizing objects into a category with higher variability. Modelling results predict that generative learners should be more greatly affected by category variability than discriminative learners. We show that humans can be prompted to adopt either a generative or discriminative approach to learning the same input, resulting in the predicted effect on use of category variability.

Keywords: human category learning; generative models; discriminative models; rational models; Bayesian models

Introduction

Is the plant poisonous or nutritious? Was that the sound of a friend or foe? The ability to categorize objects and events in the world is crucial for survival, and is one of the most heavily researched areas of cognitive psychology. Categorization is also a focus of research in machine learning and statistics, which has provided a source of connections between these disciplines in the form of rational models of cognition (e.g., J. R. Anderson, 1990; Griffiths, Canini, Sanborn, & Navarro, 2007). However, machine learning research makes a distinction between two different approaches to category learning that has not been explored in the psychological literature: the distinction between *generative* and *discriminative* models (e.g., Ng & Jordan, 2001).

Generative and discriminative models represent two distinct strategies for estimating the probability that a particular object belongs to a category. Generative learners solve this problem by building a probabilistic model of each category, and then using Bayes' rule to identify which category was most likely to have generated the object. Discriminative learners estimate the probability distribution over category labels given objects directly. These different strategies have implications for the performance of these models. Theoretical and empirical analyses have shown that generative and discriminative models differ in their generalization behavior, as well as the speed and accuracy of learning (Efron, 1975; Ng & Jordan, 2001; Xue & Titterington, 2008).

While the generative/discriminative distinction has been studied extensively in machine learning and statistics, it has been little examined in human behavior. A recent study has shown humans can adopt these two different strategies while learning an artificial language (Hsu & Griffiths, 2009). In this paper, we explore whether people can adopt these two strategies in category learning. We revisit a previously studied paradigm that showed people are sensitive to category variability, being more likely to assign an object equidistant from the mean of two categories to the category with higher variance (Cohen, Nosofsky, & Zaki, 2001; Rips, 1989; Smith & Sloman, 1994). Modelling results show that a generative model exhibits greater sensitivity to category variability than a discriminative model. We use this analysis as the basis for an empirical investigation of whether human learners can be prompted to take these two distinct learning approaches. Our results support the idea that humans adopt generative and discriminative approaches when appropriate. This provides new insight into the factors affecting human category learning.

Generative and discriminative models

Rational models of categorization identify the underlying problem as one of estimating the probability of a given object x belonging to a category c, as expressed by the distribution $p(c|x)$. The difference between generative and discriminative approaches to categorization comes down to how this probability distribution is estimated. Generative models build a probabilistic model of the input by learning the probability that an object x is generated given that the category is c, $p(x|c)$, and then solving the categorization problem by applying Bayes' rule. Discriminative models estimate $p(c|x)$ directly. Generative models thus assume that observed objects are sampled in a way that reflects $p(x|c)$, while discriminative models do not make any assumptions about the distribution from which the input is sampled. These two approaches to categorization are illustrated schematically in Figure 1.

Comparison of generative and discriminative approaches to category learning has been done in the machine learning and statistics literature, where the classic *generative-discriminative pair* being compared is usually (generative) naïve Bayes vs. (discriminative) logistic regression (Efron, 1975; Ng & Jordan, 2001; Xue & Titterington, 2008). Under certain conditions, these two models are identical in the asymptotic form of the function $p(c|x)$ that they produce, differing only in how that function is estimated. Such generative-discriminative pairs can thus be used to explore the consequences of adopting these different strategies through mathematical analysis and simulations. For example, if the training data consist of two normally distributed samples, generative models learn categories more quickly (Efron, 1975; Ng & Jordan, 2001). However, when the train-

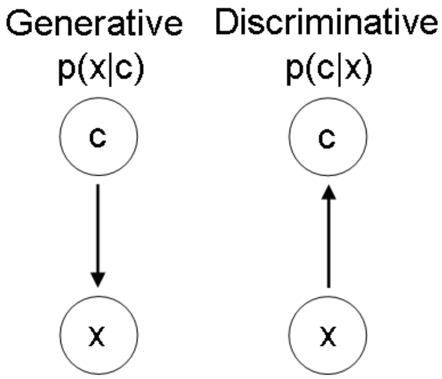

Figure 1: Generative and discriminative models. Generative models aim to estimate the probability distribution over the input given the category label. Discriminative models find a direct mapping between inputs and category labels.

ing data come from other distributions, discriminative models are asymptotically more accurate (Xue & Titterington, 2008), though in some cases generative models may perform better initially and arrive at their (higher) asymptotic error more quickly (Ng & Jordan, 2001).

Relationship to previous categorization research

Previous models of categorization have used both generative and discriminative strategies, without necessarily recognizing that the significance of the distinction. Rational models of categorization (e.g., J. R. Anderson, 1990; Griffiths et al., 2007) have tended to take a generative approach, explicitly estimating $p(x|c)$. Connectionist models (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004) typically take a discriminative approach, learning a function that takes x as an input and produces a distribution over categories as an output, and thus estimating $p(c|x)$ directly. Generative-discriminative pairs also exist among previous models of categorization. In addition to the relationship between naïve Bayes and logistic regression, which is roughly analogous to the relationship between prototype (e.g., Reed, 1972) and decision-bound (e.g., Maddox & Ashby, 1993) models, exemplar models such as the generalized context model (Nosofsky, 1986) can be interpreted either in terms of approximating $p(x|c)$ using a kernel density estimator (Ashby & Alfonso-Reese, 1995), or as a way to estimate $p(c|x)$ using a radial basis function neural network (Kruschke, 1992).

Related distinctions

The generative/discriminative distinction has not previously been explored in human learning. However, it is related to known effects of causal direction, classification vs. inference learning, and observational vs. feedback learning.

Causal direction The effect of the direction of the learned relationship between a set of variables (which is at the heart of the generative/discriminative distinction, as shown in Figure 1) has been examined extensively in the literature on causal learning. Studies have examined the difference between two ways in which people could acquire knowledge about the joint distribution of a set of variables: *predictive* learning, where relationships are learned from causes to effects, and *diagnostic* learning, where relationships are learned from effects to causes (Reips & Waldman, 2008; Waldman, 2000, 2001). In one such study, different substances in the blood were presented as cues that were indicators of a disease. In the predictive condition these measures were characterized as causes of the disease whereas in the diagnostic condition, these measures were characterized as effects of the disease. A second factor was then crossed with this manipulation: In one condition only one cue was correlated with the disease, while in the other condition two cues were correlated with disease. Other than the reversal of cause and effect, participants in both conditions received training on the exact same data. It was found that participants with the predictive instructions gave significantly lower predictive ratings for the two cue condition relative to the one cue condition compared with participants in the diagnostic condition. Thus, lower ratings are only produced when redundant cues are causes and not when they are effects (Waldman, 2001). A further study showed that only diagnostic learners utilized base rate information when asked to rate the probability of low-occurrence diseases given presented symptoms, although this effect occurred only for more complex causal structure between diseases and symptoms (Reips & Waldman, 2008).

Classification vs. inference learning Another line of experiments has shown that human category learning can also be influenced by using different tasks to teach people about the relationship between categories and features. The effect of using these two different tasks is similar to that of changing the direction of a learned causal relationship. (A. L. Anderson, Ross, & Chin-Parker, 2002; Markman & Ross, 2003; Ross & Murphy, 1996). In these experiments, all participants were presented with exactly the same training stimuli, consisting of the features and category membership of a set of objects. In one condition, learning took place via through *classification*: Participants were provided with the values for (some of) the features of an object asked to predict category membership. In the other condition, learning was based on making a predictive *inference*: The category membership and/or values of some of the features were provided and participants were asked to predict the value of another feature. Because participants in both conditions were given feedback, they were both ultimately provided with exactly the same information about categories and features. However, learning results differed in terms of performance accuracy and generalizations made. For example, inference learners performed better than classification learners on single-feature classification tasks but more poorly when all of the features were provided (A. L. Anderson et al., 2002). While this study was not motivated by the difference between generative and discriminative learning, people may have adopted these different strategies in the different conditions: Classification learning can be done using a discriminative model, while inference learning requires a generative model.

Observation vs. feedback training Another study, by Ashby, Maddox, and Bohill (2002), has also examined how learning of the exact same input was affected by presentation style. Here they compared what they called *feedback* training (where the category label appears after the object) with *observation* training (where the category label appears before the object). Their results showed that participants in the feedback condition performed significantly better than those in the observation condition for information-integration categories, where category membership could not be expressed in terms of a rule using a single feature. These two forms of training might encourage learners to adopt generative and discriminative strategies. Feedback training gives an error signal that can be used to adapt a discriminative model. Observation training is more relevant for learning object features based on the category label, which is the approach of a generative model.

Summary

Generative and discriminative models use different approaches to solve the problem of categorizing objects. Existing models of human category learning differ in which of these approaches they use. Previous work has not explored whether people are able to switch the approach they take in learning categories, although the effects of different training regimes that might encourage one approach over the other have been investigated. In the remainder of the paper, we explicitly test whether people can adopt these two approaches to learning categories, using a phenomenon that is diagnostic for one generative-discriminative pair of models.

Differential use of category variability

Several experiments have shown an effect of category variability on human categorization judgments. In these experiments, the stimuli belong to one of two categories with different means and variances. The key question is how stimuli with features lying (perceptually) in between the two categories are categorized. The results of these experiments all showed that there was a bias towards categorizing stimuli into the high-variance category (Cohen et al., 2001; Rips, 1989; Smith & Sloman, 1994). Here we propose that the degree of preference for the high variance category may be affected by whether the learner is adopting a generative or discriminative approach.

Intuitively, we expect category variability to have a greater effect on generative learners because estimating $p(x|c)$ for each category requires being sensitive to the variance of that category. In contrast, one need not consider the variance of the stimuli in simply learning a function from x to c, $p(c|x)$. Indeed many discriminative models used in machine learning, such as support vector machines (Schölkopf & Smola, 2002), focus just on the location of the most extreme members of each category. We are not claiming that all generative models are sensitive to category variance, or that all discriminative models are insensitive, but that these approaches differ in the extent to which they are sensitive to this property of the stimuli. To illustrate this, we will explore the predictions of one generative-discriminative pair of models.

We follow previous work exploring the difference between generative and discriminative models (e.g., Ng & Jordan, 2001) and focus on the generative-discriminative pair of naïve Bayes and logistic regression. Since we will focus on continuous stimuli, we assume a Gaussian generative model, with

$$p(x|c=i) = N(\mu_i, \sigma_i) \tag{1}$$

where μ_i and σ_i are the mean and variance of the ith category with $i \in \{1,2\}$. The parameters μ_i and σ_i can be estimated by maximizing the likelihood $\sum_{j=1}^{n} \log p(x_j|c_j,\mu,\sigma)$, where c_j and x_j are the category membership and features of the jth stimulus respectively. The probability a novel stimulus belongs to a category, $p(c|x)$, is then computed by applying Bayes' rule, with the prior probability of each category being proportional to the number of observed stimuli from that category. The discriminative model uses logistic regression to estimate $p(c|x)$ directly, with

$$p(c=1|x,w,b) = 1/(1+\exp\{-w^T x) - b\}) \tag{2}$$

where w and b are the parameters of the model and x is a vector of feature values. The parameters w and b are estimated by maximizing the log likelihood $\sum_{j=1}^{n} \log p(c_j|x_j,w,b)$. In general, w and b are vectors of length equal to the number of stimulus features. However, we will be using one-dimensional stimuli (x_j is scalar), so w and b will be scalars in our case.

To examine the predictions of these models, we used stimuli based largely on those of Cohen et al. (2001). Stimuli consisted of vertical lines of varying lengths. Training stimuli belonged to one of two categories, A and B. Category A is the low variance category. Category A contained lines of length 110, 120, 130, 140 and 150 pixels. Category B was the high variance category. Category B contained lines of length 300, 375, 450, 525 and 600 pixels. All stimuli were equally likely within each category (categories had a flat distribution of stimuli). We also included novel transfer stimuli in the test stimuli. There were eight transfer stimuli, equally spaced between the highest value of A and the lowest value of B (see Figure 2). A range of intermediate transfer stimuli were used in case the middle stimulus in psychological space differed from the numerical middle stimulus. The precise location of the middle stimulus is not important for our purposes, as the difference in results between generative and discriminative models is the question of interest.

To compare the outcomes of the two models, we analysed categorization predictions for our transfer stimuli using the generative and discriminative models summarized above. The generative model predicts intermediate transfer stimuli will be classified to the high-variance category more often

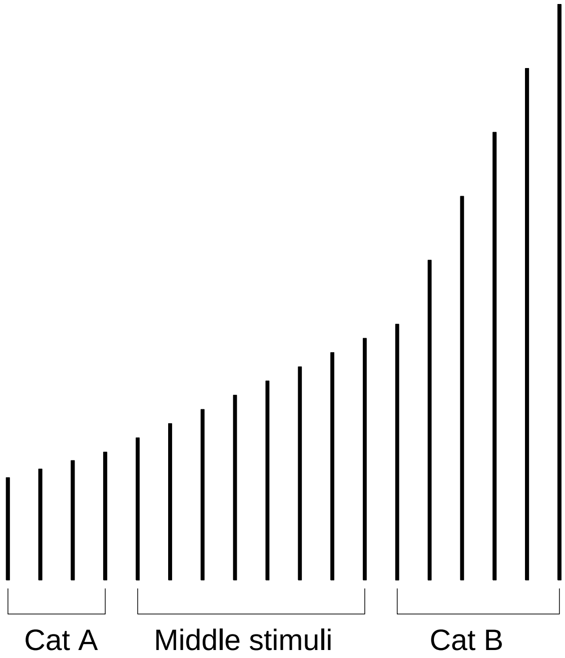

Figure 2: Stimuli used in the experiment. Category A and B were the low and high variance categories respectively

than the discriminative model (see Figure 3). This is because it is more likely that intermediate stimuli are extreme values from the high-variance category than the low-variance category. These results illustrate that sensitivity to category variability may be a diagnostic indicator of whether learners are using a generative or a discriminative strategy. In the next section we present an experiment that uses this indicator to determine whether human learners switch between these strategies depending on the way in which a categorization task is presented.

Human generative and discriminative learning

Method

Participants We collected data from 24 participants (12 in each condition). Participants were recruited from the community at the University of California, Berkeley and received course credit for participation.

Stimuli Stimuli was the same training and transfer stimuli used in the model simulations described in the previous section. In the experiment, these stimuli were presented as white vertical lines in a black circle.

Procedure Participants in both learning conditions were trained under the same randomized sequence of signs and associated tribes. In order to prompt generative vs. discriminative approaches, the two conditions differed in the instructions, cateogory-stimulus presentation order and question presented during testing blocks. Participants in both conditions were told they will see “signs” from an alien tribe. Participants in the *generative* condition were told that two aliens, one from each tribe A and B will appear and produce signs from their respective tribes. A picture of two aliens, who were identical except for the letter on their chest, was shown alongside the instructions). These instructions were intended to make it clear that the observed stimuli were generated from a probability distribution associated with the target category, consistent with the assumptions of a generative model. Participants in the *discriminative* condition were told that there are signs from two alien tribes and they would be shown a single alien translator who can report which tribe a sign was from. A single alien was shown alongside these instructions with a question mark on its chest. These instructions were intended to establish a situation in which participants learned a function from stimuli to category membership, consistent with a discriminative model.

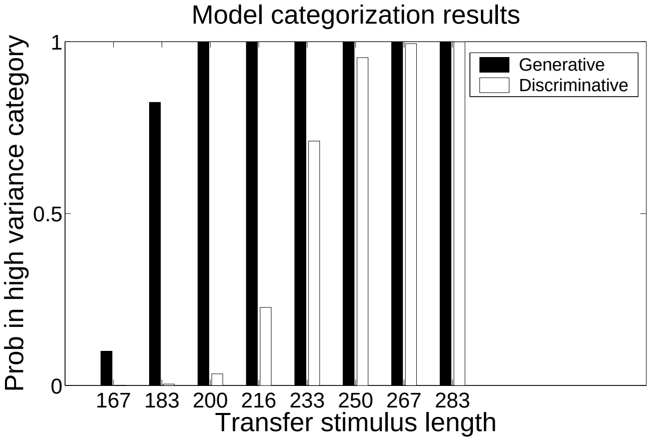

Figure 3: Generative and discriminative model predictions for the probability of categorization stimuli into the high variance category. The model predictions are that a generative learner is more likely to categorize in between stimuli in the high variance category

For all participants, the experiment contained 10 blocks of 20 trials with training blocks (odd blocks) interleaved with testing blocks (even blocks). During training trials, participants were shown a black circular background on which the “sign” appears as a white vertical line, next to an alien with either A or B written on its chest. In the *generative* condition, the alien appeared 500 ms before the sign during training and the alien disappeared between trials to simulate different aliens appearing. In the *discriminative* condition, the sign appeared 500 ms before the alien and the alien did not disappear between trials to simulate one constant alien interpreter. In both conditions,once both stimulus and letter had appeared, both remained simultaneously on the screen for 1.5 s (see Figure 4). The total length of each training trial was 2 s and there were 700 ms between each trial.

During test trials, participants were shown a sign (white vertical line) on the black circular background. Participants in the *generative* condition were asked “Which alien was more likely to have produced this sign?”. Participants in the *dis-*

criminative condition were asked “Which alien tribe does this sign belong to?”. Stimuli during each test block consisted of every example stimulus in A and B, along with the eight transfer stimuli that were equally spaced between and highest value of A and the lowest value of B. (The highest value of A and lowest value of B were seen twice during each test block to make up the 20 trials.)

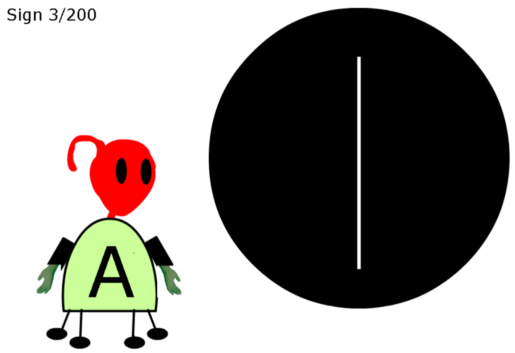

Figure 4: Screen shot of the experiment

Results

The human learning results correspond to the predictions of the models: Generative learners are more likely to categorize transfer stimuli that lie in between the two categories in the high-variance category relative to discriminative learners (see Figure 5). A two-way within-between ANOVA revealed statistically significant effects of test stimulus ($F(9,198) = 76.88$, $MSE = 0.036$, $p < .001$) and condition ($F(1,22) = 5.43$, $MSE = 0.216$, $p < .05$) and a marginally significant interaction ($F(9,198) = 1.90$, $MSE = 0.036$, $p = .054$). Planned comparisons using two-sample t-tests showed statistically significant effects of condition for stimuli 216 ($t(22) = 2.57$, $p < .05$) and 233 ($t(22) = 2.46$, $p < .05$). These statistics are calculated under the most conservative assumption, under which the multiple responses from each participant for each stimulus are averaged together and treated as a single response (ie. assumed to be completely dependent).

The “middle stimulus” that lies midway between the two categories in human perceptual space (i.e. equally likely to be categorized in both categories in the discriminative condition) is of length around 200 pixels. This is smaller than the numerical middle (225 pixels). This is approximately the same value as the perceptual “middle stimulus” that was found in previous work (Cohen et al., 2001). Accounting for this shift, the discriminative model predictions match fairly well with the discriminative human results. The generative model predictions are significantly shifted to the left compared with our generative human results, meaning the generative model predicted an even stronger tendency to categorize the in-between stimuli in the high variance category. This difference in degree between model predictions and human judgments could be explained in many possible ways. For example, it is possible that people are not making the Gaussian assumption that was made by our model. This is highly likely as our stimuli were very non-Gaussian. In this case, it is possible that the probability of belonging in the high variance category under a Gaussian assumption is greater than the probability estimates that generative participants might have made for our actual stimuli. Another possibility is that participants may not be behaving fully generatively, or that the instructions resulted in a mixed population of generative and discriminative learners in this condition.

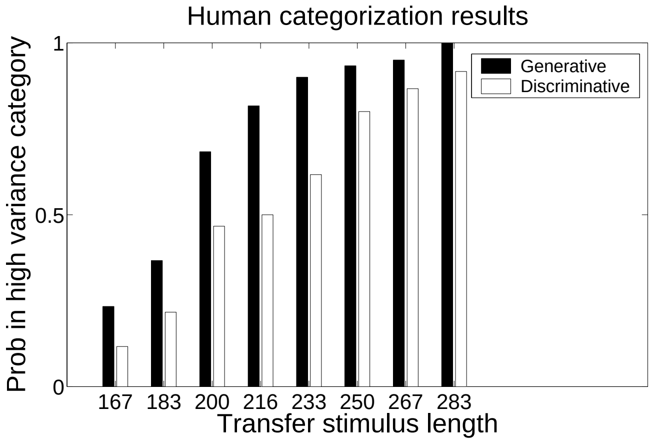

Figure 5: Probability of categorizing transfer stimuli in high variance category for participants in the generative and discriminative learning conditions. Total values are the average of all probabilities for individual stimulus lengths.

Discussion

The distinction between generative and discriminative approaches to categorization has played an important role in machine learning research, but has not previously been explored in cognitive psychology. Our results show that people can be cued to take these two different approaches to category learning through the way in which a categorization task is presented. These results have implications for understanding human category learning, and for establishing links between the communities studying human and machine learning.

The finding that people behave differently when encouraged to adopt these two different approaches to category learning may shed light on previous empirical results in cognitive psychology. For example, some previous experiments have shown effects that may be partly due to learning paradigms that encouraged participants to adopt generative or discriminative learning approaches (e.g., Ashby et al., 2002). The generative/discriminative distinction also has potential implications for previously proposed models of categorization. For example, it seems appropriate that connectionist models (Kruschke, 1992; Love et al., 2004) will best characterize behavior when humans adopt a discriminative learning approach whereas rational models (J. R. Anderson, 1990; Griffiths et al., 2007) will best describe behavior when humans adopt a generative learning approach. Developing a deeper understanding of how this distinction plays out in human learning may provide additional insights into longstanding debates on category learning.

Showing that people can adopt both generative and discriminative learning strategies establishes a new connection between human and machine learning. While many of the goals of machine learning are inspired by human capabilities (e.g., the ability to recognize and categorize complex structures quickly and efficiently), the principal issues that are topical in machine and human learning seldom coincide. By showing that a key distinction long studied in machine learning research is also significant to human learning, this work begins to build an important bridge between machine learning and human learning communities. This will encourage collaboration between the two research communities where computational models of learning provide insight into human learning and human learning, in turn, inspires computational modelling. It also establishes a way to know how advances in specific aspects of machine learning, such as improved discriminative models, might be relevant to predicting aspects of human learning.

Identifying the relevance of the generative/discriminative distinction in human categorization also opens up many new avenues of research questions. For the neuroscience community, one can ask: What neural mechanisms are implementing these two very different learning strategies? Are the neural circuits involved similar or different? This research also provokes many questions about learning more generally: When does human learning tend to be generative or discriminative? How flexible are learners in alternating between generative and discriminative learning approaches? Can learning approaches be retrospectively altered? (i.e. if input is learned with a discriminative perspective and learners were later made to understand that the data was generated from a probability distribution, would they switch their categorization judgments?) Since much of human learning in everyday life consists of a mix of scenarios in which one or the other of these strategies is more appropriate, clarifying when people use generative and discriminative approaches will help us understand differences in learning among individuals and across situations. We anticipate that exploring these questions will result in improved models of human category learning, and a tighter coupling between research on human and machine learning.

Acknowledgments. We thank Ky Merritt for assistance in data collection. This research was funded by the Economics and Social Research Council grant RES-000-22-3275 and National Science Foundation grants SES-0631518 and IIS-0845410.

References

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Journal of Experimental Psychology: Learning, Memory, Cognition*, *30*, 119–128.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Ashby, F. G., Maddox, W. T., & Bohill, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory and Cognition*, *80*, 666–677.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory and Cognition*, *29*, 1165-1175.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, *70*, 892–898.

Griffiths, T. L., Canini, K., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hsu, A. S., & Griffiths, T. L. (2009). Differencial use of implicit negative evidence in generative and discriminative language learning. In *Advances in Neural Information Processing Systems 22.*

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, *53*, 49–70.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592–613.

Ng, A. Y., & Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 17.* Cambridge, MA: MIT Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393–407.

Reips, U. D., & Waldman, M. R. (2008). When learning order affects sensitivity to base rates: Challenges for theories of causal learning. *Experimental Psychology*, *55*, 9–22.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S.Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* New York: Cambridge University Press.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, sCognition*, *22*, 736–753.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Smith, E. E., & Sloman, S. A. (1994). Similarity-versus rule-based categorization. *Memory and Cognition*, *22*, 377–386.

Waldman, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, Cognition*, *26*, 53–76.

Waldman, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. *Psychonomic Bulletin and Review*, *8*, 600–608.

Xue, J., & Titterington, D. M. (2008). Comment on "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes". *Neural Processing Letters*, *28*, 169–187.