Do additional features help or harm during category learning? An exploration of the curse of dimensionality in human learners

Wai Keen Vong (waikeen.vong@adelaide.edu.au)
Andrew T. Hendrickson (drew.hendrickson@adelaide.edu.au)
Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Daniel J. Navarro (dan.navarro@unsw.edu.au)
School of Psychology, University of New South Wales

Abstract

How does the number of features impact category learning? One view suggests that additional features creates a "curse of dimensionality" - where having more features causes the size of the search space to grow so quickly that discovering good classification rules becomes increasingly challenging. The opposing view suggests that additional features provide a wealth of additional information which learners should be able to use to improve their classification performance. Previous research exploring this issue appears to have produced conflicting results: some find that learning improves with additional features (Hoffman & Murphy, 2006) while others find that it does not (Minda & Smith, 2001; Edgell et al., 1996). Here we investigate the possibility that category structure may explain this apparent discrepancy – that more features are useful in categories with family resemblance structure, but are not (and may even be harmful) in more rule-based categories. We find while the impact of having many features does indeed depend on category structure, the results can be explained by a single unified model: one that attends to a single feature on any given trial and uses information learned from that particular feature to make classification judgments.

Keywords: Category learning; supervised learning; curse of dimensionality

Introduction

Category learning can become increasingly difficult as the number of object features increases. This "curse of dimensionality" occurs because the learner must in some way search over the large number of features in order to determine how to weight the importance of each during classification (Sutton & Barto, 1998). Despite this difficulty, people – even small children – easily learn natural categories composed of objects with a very large number of features (Rosch, 1973). How do people overcome the curse of dimensionality when they learn high-dimensional categories such as these? In this paper we present simulations and empirical results that show that susceptibility to the curse depends on what is being learned: whether the categories involved follow a family-resemblance structure or are more rule-based.

Category learning experiments have traditionally avoided the curse of dimensionality by using stimuli that consist of only a few highly salient features, generally between two and four (e.g., Medin & Schaffer, 1978; Minda & Smith, 2001; Nosofsky, 1986; Shepard, Hovland, & Jenkins, 1961). Although these experiments have substantially contributed to our understanding of category learning, it remains largely an open question how learning changes (if at all) when there are a large number of features. The few studies that have investigated this empirically have yielded conflicting results. Increasing the number of features has been variously found to impair learning (Edgell et al., 1996), to facilitate learning (Hoffman & Murphy, 2006), or to not impact learning at all (Minda & Smith, 2001).

How can we resolve this apparent discrepancy? One possibility is that the studies differ in the kinds of categories being learned. After all, the curse of dimensionality stems from having so many possible stimuli configurations in a high dimensional space that it is difficult to figure out which features are the most important ones. This should lead to the greatest inefficiency when most of the possible features are not diagnostic of category membership and only one or a few matter, as in Edgell et al. (1996). By contrast, if all features are diagnostic to some degree – especially if they are not perfectly correlated with each other – then additional features should be beneficial, or at least not hurtful (Hoffman & Murphy, 2006; Minda & Smith, 2001).

This reasoning is sensible, but no studies to date have tested it by manipulating category structure and number of features while holding other factors constant. The goal of the current paper is to do this. Our results suggest that people's ability to evade the curse of dimensionality in natural categories occurs because of the family resemblance structure of natural categories – but that in rule-based categories the curse defeats us. We also show that although people's performance qualitatively varies depending on the nature of the categories to be learned, it can be accounted for by a single unified model with limited attentional abilities.

Experiment

Our experimental design involves systematically manipulating the number of features and the category structure in a simple categorization task. We were interested in how performance changed with increasing numbers of features, and how this depended on the nature of the categories being learned (family resemblance, intermediate or rule-based). As predicted, learning decreased when there were additional features when category structure was rule-based, but did not when it was more of a family resemblance structure.

Method

Participants 442 participants (238 male) were recruited via Amazon Mechanical Turk. Participants ranged in age 19 to 76 (mean 34.2). They were paid US$2.00 for completion of the experiment, which took roughly 12 minutes to complete. 14 participants failed to complete the task, and 5 participants had participated in a pilot version of this study; these data were excluded from further analyses.

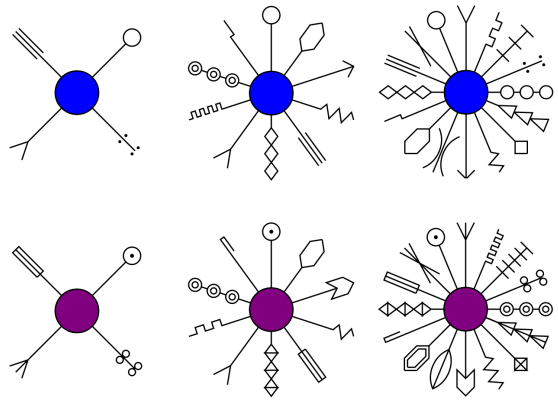

Figure 1: Six example stimuli, displaying two examples from each of the three possible *Dimensionality* conditions (left: 4, middle: 10, right: 16). Features were binary and correspond to the legs of the amoebas. Together, the two examples from the 16-FEATURE condition show all possible feature values.

Design The task was a supervised category learning experiment in which people learned to classify amoeba as either bivimias or lorifens. Each amoeba consisted of a circular base with a set of distinct binary features (legs). The full set of 16 unique pairs of features are shown on the two stimuli in the right column of Figure 1.

The nine experimental conditions were created by manipulating two factors (the *Dimensionality* of the stimuli and the category *Structure*), each with three levels, in a between-participant design. The three levels of *Dimensionality* reflect the number of binary features present on the stimuli (4-FEATURE, 10-FEATURE, or 16-FEATURE conditions). For the lower-dimensionality conditions, the set of displayed features chosen were randomly selected from a subset of the features used in the 16-FEATURE condition. The position of features on the amoeba and the mapping from feature values to category labels were randomized differently for each participant.

The three category *Structures* defined the relationship between feature values and category labels. For all three structures, one dimension was 90% predictive of the correct category label (meaning that on 90% of trials, categorizing according to that dimension would lead to a correct label). The different category structures were defined by the diagnosticity of the other features in the category. In one condition, the other features were 50% predictive of the category label;[1] we call this the RULE condition because maximum performance can be achieved by finding the one highly-diagnostic feature and ignoring all of the rest. In another condition, all of the features were 90% predictive of the category label (though all were generated independently, so none were perfectly correlated with each other). We called this the FAMILY condition because this imposes a family resemblance structure with many highly coherent and predictive features. Finally, in the INTERMEDIATE condition the other features were 70% diagnostic: thus, one feature was most diagnostic but it would be theoretically possible to achieve better performance by using all of the features in concert.

Procedure The experiment consisted of five blocks of 20 learning trials each, for a total of 100 trials. On each trial people were presented with an amoeba[2] and were asked to classify it as either a bivimia or a lorifen. They received points both for answering correctly and quickly. To make it more game-like each trial was associated with a countdown bar that decreased in size over time. After responding, feedback was given by displaying the true category label for three seconds. In addition, the circular base of the amoeba lit up with the appropriate category colour (blue for bivimias and purple for lorifens). There was a one second delay after the feedback before the presentation of the next stimulus. At the end of each block, participants were presented with a short summary of their performance across each completed block.

Results

Before addressing the main question of how learning is influenced by *Dimensionality* and *Structure*, it important to verify that learning in fact took place. As Figure 2 illustrates, participants in all nine conditions showed evidence of learning. We evaluate this quantitatively using a Bayesian mixed effects model with block as a continuous variable, and *Dimensionality* and *Structure* as discrete variables.[3] Across all conditions, accuracy increased during training: the model that included block was strongly preferred over a model that only included a random effect for each participant ($BF > 10^{44} : 1$).

How did the number of features and the structure of the categories affect learning? As is evident from Figure 2, learning was fastest in the FAMILY condition, slowest in the RULE condition, and intermediate in the INTERMEDIATE condition; the corresponding main effect of *Structure* yielded a Bayes factor of more than 700:1 in favor of a difference. It is also evident that performance was affected by the number of features, with the main effect of *Dimensionality* yielding $BF > 47 : 1$.

The main effects are sensible, but the main prediction was that we expected the effect of *Dimensionality* to be different for different *Structures*. Was such an interaction observed? Figure 2 suggests there was one, with performance decreasing with additional features in the INTERMEDIATE and RULE conditions, but not in the FAMILY conditions. Supporting this, a Bayesian mixed effects model containing block, structure, dimensionality and the interaction between structure and dimensionality was strongly preferred over a model without the interaction term ($BF > 10^3 : 1$). Overall, these results suggest that high dimensionality is only a curse as categories grow more rule-based; if they are not, the high informativeness of every feature renders the search problem less of an issue.

[1] Since there were two categories, this means they were not predictive at all.

[2] The stimuli for each person was generated randomly according to the appropriate category structure, rather than pre-generating 100 specific stimuli and showing the same ones to everybody.

[3] All mixed effects models in this paper assume a random intercept for each subject. Bayes factors were calculated using the BayesFactor package 0.9.12-2 (Morey & Rouder, 2015) in R 3.2.3. Because it is typical to obtain a range of possible factors within a confidence interval, for simplicity we report the approximate factor.

	Category Structure		
Number of Features	Family	Intermediate	Rule
4 Features	4 features all 90% predictive	1 feature 90% predictive 3 features 70% predictive	1 feature 90% predictive 3 features 50% predictive
10 Features	10 features all 90% predictive	1 feature 90% predictive 9 features 70% predictive	1 feature 90% predictive 9 features 50% predictive
16 Features	16 features all 90% predictive	1 feature 90% predictive 15 features 70% predictive	1 feature 90% predictive 15 features 50% predictive

Table 1: The nine different conditions tested in the experiment. For all of the Intermediate and Rule conditions where only 1 feature was 90% predictive, this 1 feature was chosen at random.

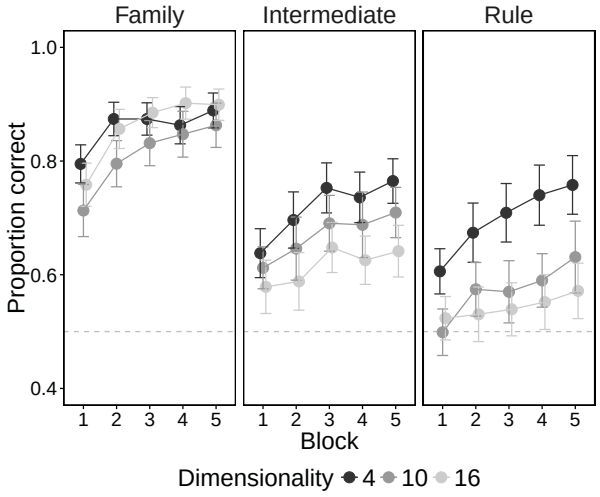

Figure 2: Mean human performance across the three *Dimensionality* and *Structure* conditions. While learning within the FAMILY-resemblance categories was unaffected by the number of features, more features meant that learning was poorer in the RULE-based and INTERMEDIATE categories. Error bars reflect 95% confidence intervals, and the dotted line reflects chance performance.

Two models of human performance

The intuitive reasoning motivating this experiment was based on the insight that if people approach category learning by searching among all possible features, then the curse of dimensionality should hurt performance only when one or few features are useful (as in rule-based categories) but not if most or all of them are (as in family-resemblance categories). The empirical results support our predictions, but another important test is whether a search-based model qualitatively reproduces human performance while a model that uses all of the information from all features does not. In this section we implement such a test by modeling people's behavior with two different learning models. Both models were simulated on nine conditions that exactly paralleled the conditions in the experiment, with three levels of dimensionality (4, 10, and 16) and the same three category structures (FAMILY, INTERMEDIATE and RULE-based).

The structure and notation of the learning environment is identical for both models. The input for each trial is a D-dimensional stimuli vector $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, where D is the dimensionality of the stimulus and each x_i is a binary feature ($x_i \in \{0,1\}$). The predicted category response ($\hat{y} \in \{0,1\}$) for the n^{th} trial is defined by the feature information from the n^{th} trial along with the representation learned by the model based on the previous $N-1$ trials. The two learning models we consider differ according to the representation they learn from experience.

Naive Bayes

The first model we consider is the Naive Bayes classifier, which uses information about every feature to determine category predictions. The model tracks the diagnosticity of each feature ($p(x_i|y)$) across all previous trials to compute an estimated probability of each category label (y) for a given stimulus ($\mathbf{x}$). The model assumes each of these features are independently diagnostic of the category label and combined as in Equation 1. The predicted category response ($\hat{y}$) of the model on each trial is the category with the highest estimated probability (Equation 2).

$$p(y|\mathbf{x}) \propto \prod_{i=1}^{D} p(x_i|y)p(y) \quad (1)$$

$$\hat{y} = \arg\max_{y} p(y|\mathbf{x}) \quad (2)$$

Hypothesis Testing

The second model is a Hypothesis Testing model which assumes that categories are defined by a single binary feature. On each trial, the model maintains a single hypothesis h_{ia} that consists of a simple rule for determining the predicted category response. All rules in the hypothesis space share the same format: `if` $x_i = a$ `then` $\hat{y} = 0$, `otherwise` $\hat{y} = 1$. The space of hypotheses is defined by a, indicating a particular feature value ($a \in 0,1$), and the feature x_i where: ($i \in \{1,\ldots,D\}$). As an example, a particular hypothesis the model might use is: "If the third feature dimension is 0 (i.e. $x_3 = 0$), then respond $\hat{y} = 0$, otherwise respond $\hat{y} = 1$."

The probability of staying with the current hypothesis after each trial is given by the utility u of the current hypothesis. The utility is proportional to prediction accuracy for previous trials on which it was the current hypothesis; it is equal to 1 for those hypotheses that have never been the current hypothesis (Equation 3). On trials in which the current hypothesis is discarded, a new hypothesis is selected from the set of all possible hypotheses. New hypotheses are selected in proportion to their utility, as in Equation 4.

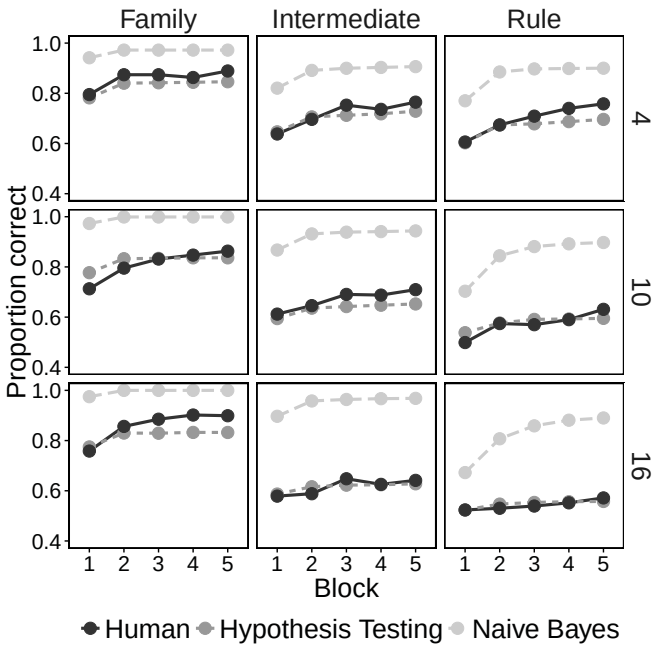

Figure 3: Predicted performance from the Hypothesis Testing and Naive Bayes models across the nine conditions. Each data point is the average of 10,000 simulations. Naive Bayes systematically overestimates performance, while the Hypothesis Testing model provides a much better fit. However, it fails to capture more subtle aspects of human performance, like the gradual learning curves.

$$u(h_{ia}) = \frac{1 + (\text{correct predictions with } h_{ia})}{1 + (\text{trials with } h_{ia})} \quad (3)$$

$$p(h_{ia}) = \frac{u(h_{ia})}{\sum_{x,y} u(h_{xy})} \quad (4)$$

Simulations

Both models were run 10,000 times in each of the experimental conditions. Each simulation mimicked one 100-trial experiment. On each trial a new stimulus was generated as in the experiment, the model made predictions, feedback was provided, and the models were updated.

Figure 3 shows the average prediction accuracy of the two models and the human data, broken down into subplots for each of the nine learning environments. The most striking finding is that Naive Bayes systematically overestimates human performance, especially for rule-based categories. This qualitative effect is mirrored in the root-mean-squared error of each model with the Naive Bayes model producing much larger overall error (0.226) than the Hypothesis Testing model (0.033). That said, both models capture many of the qualitative patterns in the human data, including the advantage for learning the FAMILY category structure relative to the RULE structure. In the FAMILY condition, both find very little effect of increasing feature dimensionality because all of the features are equally very predictive. A small positive effect of additional features does exist for Naive Bayes, because when all information is used, information from enough additional features can improve performance even above the 90% possible from any single one.

The models have similar qualitative outcomes in the RULE condition, doing worse with additional features. In the Hypothesis Testing model this occurs because of the increased difficulty in finding the most useful feature. Naive Bayes shows a very small performance decrement with more features because it does not set the feature weights of the unpredictive features to precisely zero; this effect, however, is tiny and also diminishes with time.

That said, the models make qualitatively different predictions in the INTERMEDIATE category structure: Naive Bayes predicts that additional features should improve prediction accuracy while the Hypothesis Testing model predicts the opposite. This qualitative difference emerges because of the utility of the less-predictive features in each model. The Naive Bayes model combines the information from the additional less-predictive features with the more-predictive feature to make judgments in the INTERMEDIATE condition. As a result, it makes better judgments where there are more features. By contrast, because the Hypothesis Testing model only uses one feature, it does not improve category prediction by adding additional feature information. In fact, as the number of less-useful features grows, the less likely it is for the model to switch to the hypothesis containing the most predictive feature; its performance therefore worsens.

Overall, the Hypothesis Testing model captures human performance much better than Naive Bayes, especially as the structure of the categories become more rule-based. However, the Hypothesis Testing model fails to capture some of the more subtle qualitative effects found in the human data. Most interestingly, the Hypothesis Testing model shows a sharp increase in prediction accuracy after the first block of training but does not continue to improve prediction accuracy beyond the second block. This results in a systematic underestimation of prediction accuracy in the final block across all conditions. These patterns suggest that people might be using information about more than one feature when making decisions or shifting between rules. In the following section we introduce a new model to try to account for these effects.

A hybrid model of category learning

Both the Hypothesis Testing and Naive Bayes models fail to capture all of the qualitative trends in human performance. In this section we propose a hybrid model framework that combines elements from both previous models. Like the Naive Bayes model it represents categories by assigning each feature a diagnosticity value and assumes that features are independent, but it learns in a much more limited way: on any given trial it updates the diagnosticity only for a single feature. The mechanism in the Hybrid model for determining which feature weight to update follows the same switching rule as the Hypothesis Testing model (Equation 4).

We also consider two variants of the model, corresponding to two different ways to incorporate feature information when making decisions about category membership. Both versions

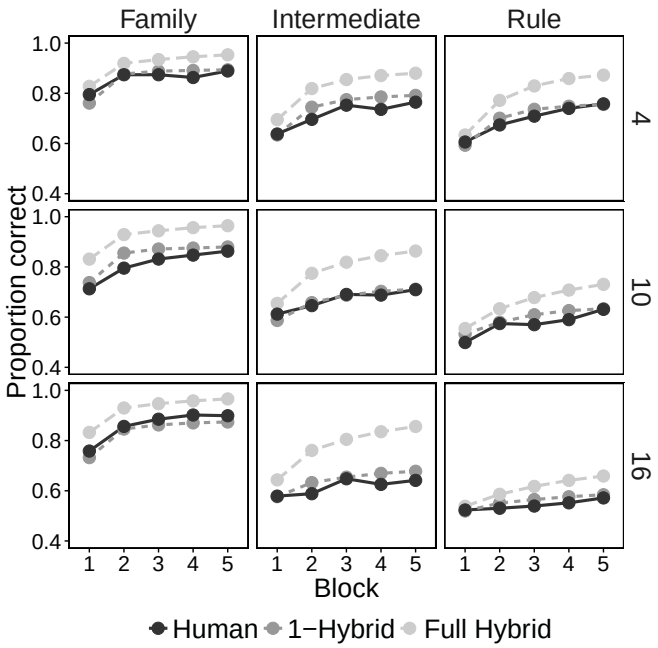

Figure 4: Predicted performance of the 1-Hybrid and Full-Hybrid models, compared against the human data across the nine conditions. Each data point is the average of 10,000 simulations. The 1-Hybrid model captures performance better than all of the other models considered.

implement the same decision rule as the Naive Bayes model (Equation 2), but differ in terms of how many features they include. The Full Hybrid model, like Naive Bayes, incorporates information from all of the features; by contrast, the 1-Hybrid model incorporates information only from the current feature when determining the category assignment.[4]

As Figure 4 shows, both Hybrid models produce learning curves whose shape closely matches that of human performance. However, the 1-Hybrid model captures the overall level and magnitude of performance much better than the Full Hybrid model; indeed, of all of the models we considered, it has the tightest quantitative fits to human data (RMS = 0.026, vs 0.033 for Hypothesis Testing and 0.107 for Full Hybrid). The poor performance of both the Full Hybrid and Naive Bayes models suggests that human learners probably do not make decisions based on combining information from all features. The improved performance of 1-Hybrid over the pure Hypothesis Testing model, however, suggests that people *do* maintain information about the diagnosticity of all features, even if only one feature is used to make category judgments at any given time.

Discussion

This paper demonstrates that the effect of additional features on category learning depends a great deal on the category structure, as reflected by the diagnosticity of the additional features. We found that more features hurt learning when the additional features were less predictive than the best ones, as in rule-based or intermediate categories. This effect was captured best by models that attend to single features for learning and prediction, rather than models that attended to or updated all features at once. These constraints are consistent with known limitations on working memory and attention in humans (e.g., Atkinson & Shiffrin, 1968).

The critical role of category structure and feature diagnosticity may explain the existing disagreements in the literature concerning the impact of feature dimensionality on category learning. Edgell et al. (1996) used a category structure in which additional features were not diagnostic of category membership and found that increasing the number of feature dimensions decreases category learning accuracy. We replicated this effect in our rule-based category structure, and both the Hypothesis Testing and 1-Hybrid model captured it. They do so because in them, increasing the number of less-useful dimensions increases the chance of switching to a less useful hypothesis and thus making a poor prediction.

In contrast, Hoffman and Murphy (2006), who used a category structure in which the new features were predictive of category membership, found that increasing the number of feature dimensions actually improved accuracy. Finally, Minda and Smith (2001) found no effect of number of features in a similar category structure. Performance in our family resemblance condition replicated the results of Minda and Smith (2001), showing no improvement in learning (but also no decrement) with additional features in a family resemblance structure.[5] This is captured by the Hypothesis Testing and 1-Hybrid models because learning about any feature is equally useful, so switching does not hurt performance.

The 1-Hybrid model accounts for the empirical data better than the other three models. The Naive Bayes and Full Hybrid models use a decision rule that produces performance that is much more accurate than human performance across all conditions, suggesting the humans do not make decisions based on information from all of the features. (Other augmentations of the models are possible as well, e.g., using a probabilistic choice rule instead of a maximizing strategy, but it is unlikely that this would change this qualitative aspect of performance). The 1-Hybrid model slightly outperforms the Hypothesis Testing model because it produces accuracy curves that continue to improve over time, while the Hypothesis Testing model underestimates improvement after the second block. This gradual improvement throughout training seems to be due to an advantage from maintaining a representations of the diagnosticity of previous feature hypotheses. In future work we will investigate this issue more precisely.

Another interesting future direction of research is to compare the 1-Hybrid model to other learning mechanisms that have been proposed to address the curse of dimensionality. These methods have focused on reducing the number of dimensions via manifold learning (Tenenbaum, 1998) or structured inference (Kemp & Tenenbaum, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Lake, Salakhutdinov,

[4] We also considered versions that included information from two to four features, but do not include these results for space reasons.

[5] Our results probably did not replicate Hoffman and Murphy (2006), because, like Minda and Smith (2001) but not it, our additional features were not perfectly correlated with existing features.

& Tenenbaum, 2015), rather than preserving the true dimensionality of the stimuli but limiting the learning mechanism by focusing on a reduced set of features on each trial. We can compare our models to such methods in the kinds of learning environments explored in this paper, as well as those that they have already been shown to successfully account for. It is, of course, possible that human learning is versatile enough to incorporate the fundamental insights from both types of models, and apply each appropriately where it is called for. This is all a matter for future work.

Overall, this research suggests that the "curse of dimensionality" negatively impacts category learning mainly in environments in which a single (or a few) features are predictive of the category, but there are many features that are not. Environments that contain many features, in which all or most of them are diagnostic of category membership, do not appear to harm performance. People's behavior can be explained by a computational model that attends to and updates a single feature at a time, shifting between features based on diagnosticity; by contrast, models that integrate information from many features or models that do not learn feature weights at all do more poorly. These results suggest that in the real world, people may be able to overcome the "curse of dimensionality" not because we are optimal learners, but rather because the structure of most natural categories is more similar to family resemblance structures in which most features are predictive of category membership.

Acknowledgments

AP was supported by grants from the Australian Research Council (DP110104949 and DP150103280, with salary support from DE12010378). The salary of AH was supported by DP110104949.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *The Psychology of Learning and Motivation*, *2*, 89–195.

Edgell, S. E., Castellan Jr, N. J., Roe, R. M., Barnes, J. M., Ng, P. C., Bright, R. D., & Ford, L. A. (1996). Irrelevant information in probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1463.

Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 301.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-2)

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). MIT press Cambridge.

Tenenbaum, J. B. (1998). Mapping a manifold of perceptual observations. *Advances in Neural Information Processing Systems*, 682–688.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.