Categorizing animated movies via Bayesian Similarity

Chima Oparaji (co9056@princeton.edu)
Department of Operations Research and Financial Engineering, Princeton University

Abstract

People often categorize complex stimuli without knowing the correct answer, relying instead on partial information and prior expectations. In this project, I examine whether a simple Bayesian similarity model can account for how people assign animated movies to the correct studios (out of Disney, Pixar and DreamWorks movies). Movies are represented as feature vectors, and their studio prototypes are derived from averaging features from randomly selected characteristic movies for each studio. On trials where participants were unfamiliar with the answer, human choices tended to align with model predictions. In addition, both humans and the model showed increased uncertainty for movies with overlapping feature structure. In contrast, known trials showed high accuracy but little relationship to the model. These results suggest that feature-based Bayesian influence captures important aspects of categorization under ambiguity, while memory likely dominates when prior knowledge is available.

Keywords: Bayesian categorization; similarity-based inference; prototype models; uncertainty; feature overlap; animated movies

Introduction

The Humans frequently categorize objects and events without knowing which category they are likely to belong to, using partial information and general expectations. In everyday settings, categories are often. Defined by family resemblances rather than strict rules, and different categories can share many overlapping features. As a result, categorization judgments are frequently prone to error, even when people have general knowledge of something.

This uncertainty is often present in cultural categories, where boundaries are learned implicitly rather than explicitly. Animated movies, for instance, are commonly associated with production studios that are thought to have distinctive visual styles, settings and narrative structures. However, popular studios such as Walt Disney Pictures, Pixar and DreamWorks share many features, including animation technologies and target audience. These overlaps make it difficult to determine whether categorization judgments reflect explicit knowledge or inference based on features.

In this project, I examine whether a simple Bayesian similarity model can account for how people assign animated movies to the studios that produce them. Rather than focusing solely on accuracy, the project emphasizes patterns of confusion and uncertainty as informative signals of underlying category structure. I analyze whether feature overlap predicts systemic misclassification, whether ambiguous movies elicit greater uncertainty, and whether a simple computational model captures these tendencies.

To address these questions, I asked participants to categorize animated movies by studio and report whether they already knew the answer. This allows a distinction between memory base judgments and those that reflect inference based on features. I test whether the model predictions align with human behavior or unfamiliar trials specifically, and whether shared uncertainty and errors provide evidence for rational inference under categories that overlap.

Based on this framework, I test three main hypotheses. First, when participants are unfamiliar with a movie's studio, their categorizations should systematically align with the predictions of a Bayesian model. Second, movies with greater feature overlap across studios should result in greater uncertainty for both humans and the model. Finally, categorization errors should be structured rather than random, with humans and the model tending to confuse the same studios when there is a high degree of feature overlap.

Background

Bayesian Approaches to Categorization

Throughout the history of cognitive science, the question of how people make categorization decisions when the correct answer is uncertain has been pondered. Rather than relying on fixed rules, people often combine partial evidence with general expectations. Bayesian models help formalize this process by treating categories as hypotheses and categorization as probabilistic inference. Based on this, judgments are dependent on how well a stimulus fits a category and how likely that category is overall (Tenenbaum & Griffiths, 2001; Griffiths et al., 2008). Within this framework, uncertainty is not treated as error, but as a natural consequence of having limited information. Bayesian approaches have been used to explain a wide range of topics in cognitive science, including inductive reasoning and concept learning, by showing how graded judgments can arise from rational inference under constraints (Griffiths & Tenenbaum, 2006). This perspective motivates the use of probabilistic models to study categorization behavior in settings where evidence is ambiguous.

Similarity, Prototypes and Overlapping Features

Earlier works on categorization emphasized how similarity shaped category judgments. Prototype theories propose that central tendencies are what categories are organized around, and items that are closer to these properties are categorized more easily and more confidently than atypical items (Rosch, 1975). When stimuli share features with multiple

categories, the process of categorization tends to become less certain. Feature based accounts of similarity further suggest that similarity judgments depend on shared and distinctive features rather than strict metric distance alone. This can produce asymmetric patterns of confusability between categories (Tversky, 1977). In conjunction with this view, psychological similarity has been shown to systematically decrease the further distance there is in a representational space, which provides a quantitative link between feature overlap and evidence for categorization (Shepard, 1987). Thus, the approaches predict that when features overlap, the result should lead to both uncertainty and structured misclassification.

Errors and Uncertainty in Categorization

Categorization research has also shown that errors can be theoretically informative. For instance, models based on "exemplars" demonstrate that misclassification patterns often reflect the underlying similarity structure of the stimulus space rather than random noise (Nosofsky, 1986). From this perspective, confusion between categories reveals which representations are close psychologically. Recent works have emphasized uncertainty as a meaningful component in human cognition. Measures such as entropy provide a way to quantify ambiguity when evidence supports multiple competing category hypotheses (Cover & Thomas, 2006). In Bayesian models, uncertainty naturally comes from the interaction between similarity-based likelihoods and prior expectations, especially when said evidence is weak or overlapping (Griffiths et al., 2008).

Taken together, these perspectives imply that categorization under uncertainty is best understood as probabilistic inference, guided by similarity, with errors and uncertainties providing important signals regarding underlying representations.

Bayesian Similarity Model of Studio Categorization

In the model I created, categorization is treated as probabilistic inference over competing category hypotheses. Each movie is represented as a vector of continuous feature values intended to capture visually and narratively important properties, including animation style, tone and genre. These features are not assumed to be optimal but are chosen to reflect information available to the human observers.
Each studio is represented as a prototype in the same feature space. Prototypes are constructed by averaging feature values across a set of random movies associated with each studio, which yields a sort of "measure of center" that captures the studio's characteristic style. Similarity between an individual movie and a studio is computed as a function of distance in this feature space. Specifically, similarity was modeled using an exponential decay function of squared Euclidean distance, such that the movies closer to the prototype of a studio produced higher values.
Thus, feature based similarity is transformed into a likelihood term, such that movies more like a studio prototype are more likely to be generated by that studio. The likelihood is then combined with prior beliefs over studios using Bayes rule, resulting in a posterior probability distribution over the studios for each movie. Posterior distributions reflect both the strength of feature evidence and any pre-existing expectations about how prevalent the studio is.
Uncertainty in the model is quantified by the entropy term of the posterior distribution. High entropy indicates that a movie shares features with multiple studios and does not strongly favor a single prototype. On the other hand, low entropy reflects a clearer match to one of the categories. Thus, the model's prediction is not just about the most likely studio, but also to what degree of uncertainty is associated with each categorization.
In the absence of definite knowledge, the model is intended to characterize inference based on available features. Rather than committing to a single category, the posterior distribution reflects graded evidence across the different studios, which allows uncertainty to emerge naturally from overlapping features. This makes it possible to compare which studio is most likely, but also how ambiguous a given movie is under the model.

Methods

Participants

Participants were recruited online, and consisted of college age students, mostly from Princeton, but some students who attend other universities as well. The final sample consisted of participants who completed all categorization trials and provided confidence, as well as familiarity judgments for each movie.

Materials

The stimuli consisted of a set of twenty-four animated movies, eight produced by Walt Disney Animation Studios, eight produced by Pixar and 8 produced by DreamWorks. Movies were selected to include both well-known/representative and less familiar/ambiguous items. Each movie was represented in the model as a vector of continuous feature values that ranged from zero to one. Feature dimensions were chosen to reflect properties plausibly available from the image and description, consistent of character type (human or animal), animation style (hand drawn vs CGI), tone (comedy vs drama), setting (fantasy vs realistic) action level and musical emphasis. Feature values were assigned prior to analysis and were held fixed across participants. When assigning feature values, I kept in mind the specific materials shown to the participants as well. From before, there is also a prototype vector for each studio, and these prototype representations were fixed prior to analysis and were used to generate model predictions for all stimuli.

Procedure

For each trial, participants were shown a movie image and description and asked to select which of the three studios they believed produced the movie. After making their choice, they were asked to rank their confidence in their guess on a scale of one to seven, with seven being very confident. Participants then were asked whether they knew what studio produced the movie prior to doing the survey. Trials in which participants reported knowing the studio were classified as known trials, while trials in which participants reported not knowing were classified as unfamiliar trials. This is an important distinction when considering memory-based responses vs feature-based inference. Before this task, participants were also asked to assign priors to each studio, implying what proportion they thought each studio produced in terms of movies overall.

Analysis Implementation

All models and analyses were done using in Python using libraries such as NumPy, Pandas and Matplotlib. Results such as Model Posteriors, entropy measures, similarity computations, figures, etc. were calculated in this environment as well.

Results

All the analysis focuses on trials in which participants marked that they did not already know which studio produced the movie. These trials are the primary target for evaluating inference based on features, as known trials are expected to rely more on memory-based reasoning.

Human-Model Correlation on Unfamiliar trials

Here, I compared participants' studio guesses to the model's predicted studio on unfamiliar trials. This was done to assess whether human studio judgments align with model predictions. Human guesses portrayed a positive association with model predictions (r = .15).

Relationship between Agreement Correctness, and Model Uncertainty

To examine when human and model predictions converged or diverged, movies were grouped according to whether humans and the model agreed, and whether the agreement was correct. Model uncertainty, which as stated before is quantified as posterior entropy, varied systematically across these cases. Movies for which humans and the model agreed and were correct exhibited the lowest model entropy, which indicates clear feature-based evidence for a single studio. In contrast, cases that involved any type of disagreement (whether human or model were incorrect) tended to have higher model entropies. To formally test whether model uncertainty differed between conditions of agreement, I compared mean model entropy for movies where humans and the model both agreed and were correct to movies involving any type of disagreement. Although the "agree and correct" movies had a numerically lower model entropy (.937) than disagreement cases (.983) this difference is not to be statically significant(t = -.91, and p-value = .386).

Human guess entropy showed a related, but slightly different pattern. Movies that humans and the model agreed upon and that were correct tended to produce lower human entropy. In contract, cases that involved disagreement, especially when both humans and the model were incorrect, were associated with higher human entropy. This reflects greater division in participant guesses.

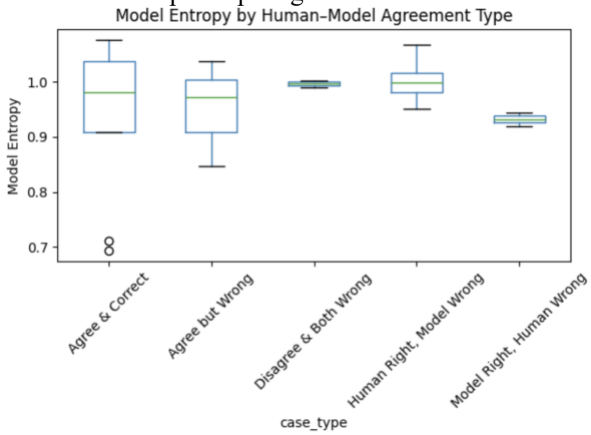

Figure 1. Model human entropy by human–model agreement type (unfamiliar trials).

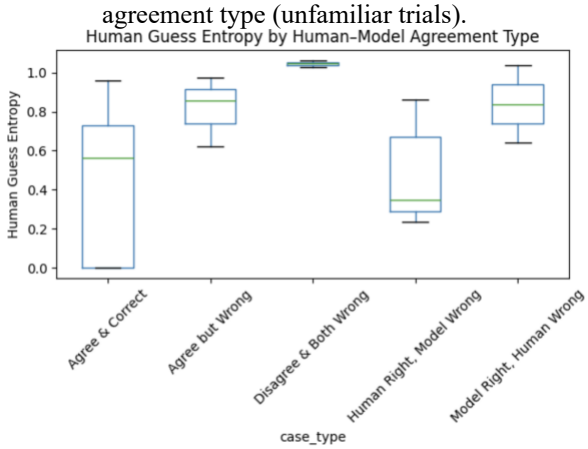

Figure 2: Human entropy by human–model agreement type (unfamiliar trials)

Confidence and Model Uncertainty per trial

Next, I examined whether subjective confidence tracked model uncertainty on individual, unfamiliar trials. At the trial level, participant confidence ratings were not reliably correlated with model entropy (with r being .01).

Known vs Unfamiliar trials

To establish a baseline and confirm that known trials reflect a different decision process, I compared accuracy across both known and unfamiliar trials. Results showed that participants were way more accurate on known trials (85%) than on unfamiliar trials (60%). On known trials, accuracy

showed a weak relationship with model posterior probability for the studio chosen (r = .20).

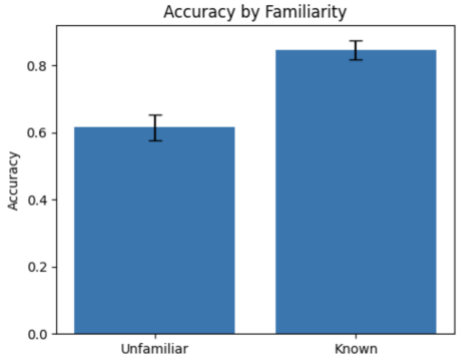

Figure 3: Mean categorization accuracy for known and unfamiliar trials.

Confusion Patterns and Similarity

To test whether human errors reflect the same feature structure as the model, I examined misclassifications on unfamiliar trials. Results showed that errors were not random; participants systematically confused studios that were closer in feature space according to the model. Prototype distances indicated that Pixar and DreamWorks were the most similar studios and this pair showed the highest rates of mutual confusion, whereas Disney, which had more of a distinct feature space, showed less symmetric confusion patterns.

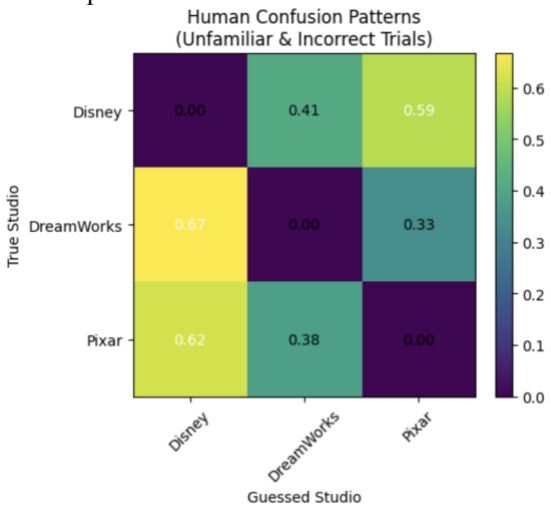

Figure 4: Normalized human confusion matrix for unfamiliar and incorrect trials

Movie-Level Confidence and Model ambiguity

Lastly, I examined uncertainty at the level of individual movies. For each movie, the mean model entropy was compared to mean participant confidence across unfamiliar trials. Movies with higher model entropy elicited lower than average confidence ratings (r = -.38).

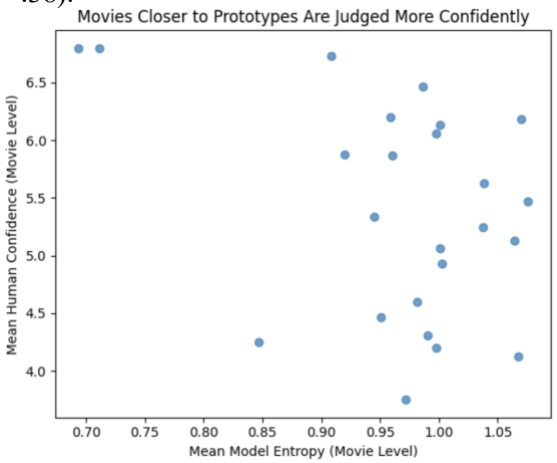

Figure 5: Mean model entropy versus mean human confidence across movies.

Discussion

The goal of this project was to test whether a Bayesian similarity model accounts for how people categorize animated movies to studios, especially when the correct answer is not known. In separating known and unfamiliar trials, this study aimed to separate judgments based on memory from judgments based on features of a given object, which allows for a clearer test of similarity-based categorization under uncertainty. Human studio guesses showed a positive association with the model's predictions on unfamiliar trials, which indicates that both humans and the model relied on overlapping information as they made their judgments. Although this relationship was modest, it is quite meaningful considering this is a three-way-categorization task and the nature of the stimuli. In addition, agreement between humans and the model varied across movies suggesting that some movies were easier to categorize than others. These findings directly support the hypothesis that feature based similarity guides categorization when explicit knowledge is unknown.

Uncertainty and Structured Errors

Uncertainty played a central role in explaining when humans and the model agreed or disagreed. The movies the model identified as more ambiguous, synonymous with higher entropy, were the same movies that produced greater disagreement between participants. When both humans and the model were correct, the entropy of humans and model guesses tended to be low. In contrast, disagreement was more common for movies with overlapping feature structures, where no single studio prototype was favored strongly. This suggests that categorization errors reflect ambiguity in the stimulus rather than noise in responding. Patterns of misclassification further support this view. Human errors were not evenly distributed across

studios and instead were clustered between studios that were closer in the model's feature space. Particularly, Pixar and DreamWorks had the most similar prototypes, and were most frequently confused with one another, compared to Disney, which showed less symmetric confusion patterns. These structured errors are consistent with both prototype based and similarity-based theories of categorization, in which items closer to multiple category centers tend to be harder to classify.

Confidence and Subjective Uncertainty

Along with predicting choices, the model also captured aspects of subjective uncertainty. When looking at individual movies, higher model entropy was associated with lower average confidence ratings, suggesting that movies closer to multiple prototypes not only led to more disagreement, but led to participants feeling less certain. It's worth noting that this relationship did not appear when looking at individual trials, which suggests that uncertainty is related to the properties of the stimuli, rather than moment by moment confidence fluctuations. Considering all of this, the model provides context as to why certain movies might feel more ambiguous than others.

Known vs Unfamiliar Trials

Known trials showed a different pattern than unfamiliar trials. Accuracy on average was way higher, and the relationships between model predictions and human categorization were weaker. This is consistent with the idea that memory-based responding dominates when participants already know the correct answer. While occasionally, the model aligned with performance on known trials, this more likely reflects the fact that movies that are well known tend to be more prototypical for their studio, rather than evidence to suggest that similarity-based inference guides these judgments. This dissociation supports the decision to focus on unfamiliar trials when evaluating the model, where feature-based inference is more relevant.

Limitations and Improvements

While the model captures several important patterns in how people categorize movies when they are uncertain, it also has clear limitations. First, the feature representations were hand-coded and only consisted of 6, which means that some stylistic or narrative cues people may rely on were likely missed. Although this was intentional, future work could incorporate richer feature representations, such as automatically extracted visual or linguistic features. Second, the model treats each studio as having a fixed prototype. This could somewhat be a problem as one's mental representation may be more flexible or based on specific examples rather than a single prototype. Finally, the number of movies and participants was relatively small, which limits how the results could be generalized. For the future, expanding the stimulus set and exploring alternative similarity functions would help clarify how robust these findings are across different kinds of stimuli.

Conclusion

Overall, these findings suggest that a Bayesian similarity model captures important aspects of how people categorize animated movies when they are uncertain of the correct answer. Rather than treating errors as noise, this project highlights how analysis of confusion and uncertainty can reveal structures that underlie category representations. Even more broadly, this project supports the idea that probabilistic models based on similarity are useful for understanding categorization in more naturalistic domains where categories overlap and boundaries are not clearly defined, meaning this type of analysis can be used in the domains of social structures like race, sexuality, etc.

References

Cover, T. M., & Thomas, J. A. (2006).
Elements of information theory (2nd ed.).
Wiley.

Griffiths, T. L., & Tenenbaum, J. B. (2006).
Optimal predictions in everyday cognition.
Psychological Science, 17(9), 767–773.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2008).
Probabilistic models of cognition: Exploring representations and inductive biases.
Trends in Cognitive Sciences, 12(4), 150–157.

Nosofsky, R. M. (1986).
Attention, similarity, and the identification–categorization relationship.
Journal of Experimental Psychology: General, 115(1), 39–57.

Rosch, E. (1975).
Cognitive representations of semantic categories.
Journal of Experimental Psychology: General, 104(3), 192–233.

Shepard, R. N. (1987).
Toward a universal law of generalization for psychological science.
Science, 237(4820), 1317–1323.

Tenenbaum, J. B., & Griffiths, T. L. (2001).
Generalization, similarity, and Bayesian inference.
Behavioral and Brain Sciences, 24(4), 629–640.

Tversky, A. (1977).
Features of similarity.
Psychological Review, 84(4), 327–352.