B

Bayesian Approaches to Color Category Learning

Thomas L. Griffiths[1] and Noga Zaslavsky[2]
[1]Departments of Psychology and Computer Science, Princeton University, Princeton, NJ, USA
[2]Department of Brain and Cognitive Sciences and Center for Brains Minds and Machines, MIT, Cambridge, MA, USA

Synonyms

Ideal observer models of color category learning; Rational models of color category learning

Definition

Bayesian approaches to color category learning formalize learning as a problem of Bayesian inference, requiring the learner to form generalizations that go beyond observed examples of members of a category. This formal framework can be used to make predictions about both individual judgments and how populations form color categories.

Color Category Learning

One of the challenges that children face as they acquire a language is discovering how words are used to refer to different colors. While human languages demonstrate variation in how they partition the space of colors, there are also clear regularities in the kinds of systems of color categories that are used [1, 2]. This raises two important questions: How might color categories be learned? And how might regularities in systems of color categories across languages be explained?

Learning color categories is an inductive problem, requiring learners to make an inference from labeled examples of colors to a full system of color categories. As in other domains of perception [3], an "ideal observer" model can be used to explore the optimal solution to this problem. Let h denote a hypothesis about a possible system of color categories and d the observed data – a set of labeled examples (such as "This color is blue, and this color is yellow"). If learners represent the degree of belief in the truth of each hypothesis with a probability, $P(h)$, then the ideal solution to the problem of updating these beliefs in light of the data d is provided by Bayes' rule:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} P(d|h')P(h')}, \qquad (1)$$

where $P(h|d)$ (known as the posterior probability, in contrast to the prior probability $P(h)$) indicates the degree of belief assigned to h after observing d and $P(d|h)$ (known as the likelihood) indicates

R. Shamey (ed.), *Encyclopedia of Color Science and Technology*,
https://doi.org/10.1007/978-3-642-27851-8_60-9

the probability of observing d if hypothesis h were true.

The sum in the denominator of Bayes' rule ranges over all possible hypotheses and ensures that $P(h|d)$ is a valid probability distribution, summing to 1. The key idea behind Bayes' rule can be obtained by ignoring this constant and simply inspecting the numerator: The new beliefs of the learner result from combining the previous beliefs, captured in the prior distribution $P(h)$, with the probability of the observed data under each hypothesis, expressed by the likelihood $P(d|h)$. The prior distribution captures the expectations of the learner, but also indicates which hypotheses are easy or hard to learn. A hypothesis that has low prior probability requires stronger evidence (in the form of a higher likelihood) to end up with a high posterior probability and so will be harder to learn. The prior distribution thus provides a way of encoding the perceptual or learning biases of the learner, favoring some hypotheses over others.

The Bayesian approach to modeling learning has proven successful in accounting for human behavior in a wide range of tasks [4]. In particular, Bayesian models have been used to account for how people learn new concepts and new words. Tenenbaum and Griffiths [5] presented an account of how people form generalizations from examples, such as inferring what other numbers might belong to a set when told that the set contains 2, 8, and 64. Under this account, hypotheses correspond to possible sets of numbers, and the likelihood is obtained by calculating how likely it is that the examples would be observed if they were sampled at random from this set. Xu and Tenenbaum [6] showed that a closely related model captured how children learned nouns corresponding to sets of objects, such as determining the appropriate referent of words corresponding to "Dalmatian" or "dog." These results suggest that a Bayesian approach might also be fruitful for explaining the acquisition of terms for color categories.

Dowman [7] presented a model that took exactly this approach, providing a Bayesian account of color category learning. In this model, the space of colors is reduced to a one-dimensional ring of hues. Each color category then corresponds to an interval on this ring, picking out a set of adjacent colors. Labeled examples of color categories are assumed to be sampled from the categories at random, with a small probability of an error taking place. This makes it possible to calculate the probability of any observed set of labeled examples for each candidate interval from which they might be drawn, providing the likelihood $P(d|h)$. Bayes' rule can then be used to compute a posterior distribution over possible intervals for each color category. The probability that a color that has not previously been labeled belongs to that category is then obtained by summing the probability of all intervals that contain that color under the posterior distribution. Dowman demonstrated that this model made reasonable inferences for simplified versions of the systems of color categories from real languages, such as Urdu.

The predictions that Dowman's model makes about learning of color categories have not been directly tested with human learners, but results in other domains and with other species provide support for this approach. As noted above, Xu and Tenenbaum [6] found that a very similar model accounted well for the generalizations that children made in learning novel words describing sets of objects. In addition, Jones, Osorio, and Baddeley [8] found that poultry chicks form generalizations about colors in a conditioning task that is consistent with a model based on that of Tenenbaum and Griffiths [5].

Cultural Transmission of Color Category Systems

The results summarized so far indicate how Bayesian models might be used to explain learning of color categories. The same models also have the potential to provide insight into why regularities exist in the systems of color categories that appear across human languages. Dowman [7] explicitly had this goal in mind in defining his Bayesian learning model, which was used as a component in a simulation of the cultural transmission of systems of color categories. Dowman's aim was to investigate the consequences of cultural transmission of systems of color categories

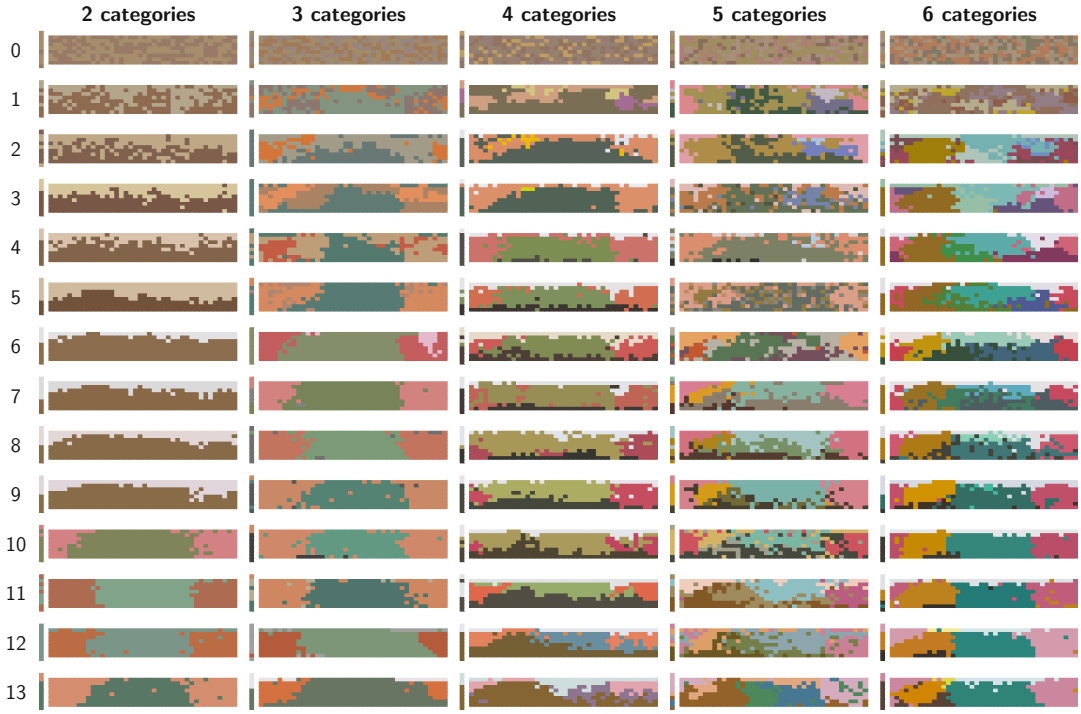
among a set of agents that used a realistic approximation to human learning. He found that cultural transmission by Bayesian agents produced systems of color categories with properties similar to those seen across human languages, providing a potential explanation for the source of those regularities.

Dowman's simulation of cultural transmission was an instance of a more general approach to exploring the origins of different kinds of structure in human languages, known as iterated learning [9]. In an iterated learning model, a set of agents each learn from some observed data and generate data that is observed by other agents. The simplest case is where the agents form a chain, with each agent learning from data generated by the previous agent and generating data that is provided to the next agent. Griffiths and Kalish [10] showed that when this form of iterated learning is carried out by Bayesian agents who all have the same prior, the hypotheses considered by those agents eventually converge to a distribution that matches the prior distribution. More precisely, the probability that an agent selects a hypothesis h converges to the prior probability of that hypothesis $P(h)$ as the chain gets longer.

The results of Dowman [7] and Griffiths and Kalish [10] raise an interesting question: Can the regularities seen in systems of color categories across human languages be accounted for by cultural transmission producing convergence on a shared prior distribution? To explore this question, Xu, Dowman, and Griffiths [11] conducted an experiment in which human participants simulated cultural transmission by iterated learning. Each participant was given some examples of colors from novel categories and then asked to generalize to a larger set of colors. The generalization responses were then recorded and used to generate the examples that were seen by the next participant. Figure 1 shows examples of color category systems produced by chains of participants in this experiment. Each chain was initialized with data generated from a random partition of color space into labeled categories. The number of allowed categories varied across chains, ranging from two to six categories. Xu et al. [11] found that over time, the systems of color categories converged to forms that were consistent with the regularities seen across human languages [2].

These results support the idea that cultural transmission and perceptual or learning biases of the kind that might be captured by a prior distribution in a Bayesian model may be sufficient to explain the origins of cross-linguistic regularities in systems of color categories. However, much remains unknown about the precise characterization of the human prior distribution over systems of color categories and whether the structure of that distribution can capture cross-linguistic variation. One approach to this question was suggested by Carstensen, Xu, Smith, and Regier [12], building on the idea that languages are shaped by pressure for informative communication. Carstensen et al. [12] showed that the iterated learning experiment of Xu et al. [11] produced systems of color categories that become highly informative, in addition to becoming more similar to actual color category systems of human languages [11]. They concluded that cultural transmission with a bias toward informative communication may explain the constrained cross-linguistic variation seen in color category systems.

Carr, Smith, Culbertson, and Kirby [13] proposed an alternative explanation for these results. They argued that a human prior distribution over category systems is better characterized by a bias toward simplicity for learning rather than informativeness. Carr et al. [13] suggested formal definitions for a simplicity prior and an informativeness prior and then used these prior distributions for model simulations of iterated learning with Bayesian agents. They demonstrated that the simplicity model can produce structured category systems that are also highly informative, which may give an impression of a bias toward informativeness even in cases where it is not actually present. In addition, Carr et al. [13] tested these models on human iterated learning data and found stronger empirical support for a bias toward simplicity. Their results, however, are based only on synthetic domains that are qualitatively different than color. Therefore, it remains unclear to what extent simplicity, informativeness, or possibly both, may characterize



Bayesian Approaches to Color Category Learning, Fig. 1 Examples of color category systems produced by chains of human learners in the iterated learning experiment of Xu et al. [11] (figure adapted from [11]). Each column corresponds to one cultural transmission chain with 13 participants (rows) and 2–6 categories. All systems are plotted against the World Color Survey stimulus palette [2], which contains 10 achromatic color chips and 320 maximally saturated color chips arranged by hue (horizontal axis) and lightness (vertical axis). Each category is color-coded by its color centroid in CIELAB space and covers all the chips in the palette that were assigned to it. The chains were initialized with a random system, shown at iteration 0. Over time, the systems produced by the learners become more similar to actual systems across languages [11], capturing observed cross-linguistic regularities.

human biases involved in learning systems of color categories.

Bayesian approaches to color category learning can be used to explore questions about how children might learn how their language partitions the space of colors, why regularities are seen in systems of color categories across languages, and how color categories may evolve over time. However, many important questions remain open for future research. One fundamental question is how well Bayesian models can capture the generalizations that real human children make when learning color categories. Another fundamental question for future research is how communication between two or more agents within a generation may influence the patterns that emerge in culturally transmitted systems of color categories. Communication is believed to play an important role in the cultural evolution of language [14], in addition to learning. Furthermore, it has been argued that systems of color categories evolve along an information-theoretically optimal trajectory of communication systems involving Bayesian agents [15]. This information-theoretic approach captures much of the regularities and variation seen in systems of color categories across languages, suggesting a potential theoretical link between communication and learning in Bayesian models of color categories.

Cross-References

▶ Berlin and Kay Theory
▶ Color Categorical Perception
▶ World Color Survey

References

1. Kay, P., Maffi, L.: Color appearance and the emergence and evolution of basic color lexicons. Am. Anthropol. **101**(4), 743–760 (1999)
2. Kay, P., Berlin, B., Maffi, L., Merrifield, W.R., Cook, R.: The World Color Survey. Center for the Study of Language and Information, Stanford (2009)
3. Kersten, D., Mamassian, P., Yuille, A.: Object perception as Bayesian inference. Annu. Rev. Psychol. **55**(1), 271–304 (2004). https://doi.org/10.1146/annurev.psych.55.090902.142005
4. Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D.: How to grow a mind: Statistics, structure, and abstraction. Science. **331**(6022), 1279–1285 (2011). https://doi.org/10.1126/science.1192788
5. Tenenbaum, J.B., Griffiths, T.L.: Generalization, similarity, and Bayesian inference. Behav. Brain Sci. **24**(4), 629–640 (2001). https://doi.org/10.1017/S0140525X01000061
6. Xu, F., Tenenbaum, J.B.: Word learning as Bayesian inference. Psychol. Rev. **114**(2), 245–272 (2007). https://doi.org/10.1037/0033-295X.114.2.245
7. Dowman, M.: Explaining color term typology with an evolutionary model. Cogn. Sci. **31**(1), 99–132 (2007). https://doi.org/10.1080/03640210709336986
8. Jones, C.D., Osorio, D., Baddeley, R.J.: Colour categorization by domestic chicks. Proc. R. Soc. Lond. Ser. B Biol. Sci. **268**(1481), 2077–2084 (2001). https://doi.org/10.1098/rspb.2001.1734
9. Kirby, S.: Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. IEEE Trans. Evol. Comput. **5**(2), 102–110 (2001)
10. Griffiths, T.L., Kalish, M.L.: Language evolution by iterated learning with Bayesian agents. Cogn. Sci. **31**(3), 441–480 (2007). https://doi.org/10.1080/15326900701326576
11. Xu, J., Dowman, M., Griffiths, T.L.: Cultural transmission results in convergence towards colour term universals. Proc. Roy. Soc. B: Biol. Sci. **280**(1758), 20123073 (2013). https://doi.org/10.1098/rspb.2012.3073
12. Carstensen, A., Xu, J., Smith, C., Regier, T.: Language evolution in the lab tends toward informative communication. In: Noelle, D.C., Dale, R., Warlaumont, A.S., Yoshimi, J., Matlock, T., Jennings, C.D., Maglio, P.P. (eds.) Proceedings of the 37th annual meeting of the Cognitive Science Society. Cognitive Science Society, Austin (2015)
13. Carr, J.W., Smith, K., Culbertson, J., Kirby, S.: Simplicity and informativeness in semantic category systems. Cognition. **202**, 104289 (2020). https://doi.org/10.1016/j.cognition.2020.104289
14. Tamariz, M.: Experimental studies on the cultural evolution of language. Ann. Rev. Linguist. **3**(1), 389–407 (2017). https://doi.org/10.1146/annurev-linguistics-011516-033807
15. Zaslavsky, N., Kemp, C., Regier, T., Tishby, N.: Efficient compression in color naming and its evolution. Proc. Natl. Acad. Sci. **115**(31), 7937–7942 (2018). https://doi.org/10.1073/pnas.1800521115