Naive Bayesian Accounts of Base Rate Effects in Human Categorization

Lewis J. Frey (Frey@VUSE.Vanderbilt.Edu)
Computer Science Department; Village at Vanderbilt
Nashville, TN 37240 USA

Douglas H. Fisher, Jr. (DFisher@VUSE.Vanderbilt.Edu)
Computer Science Department; Village at Vanderbilt
Nashville, TN 37240 USA

Abstract

This paper examines the naive Bayesian model and extensions of it to account for the effects of base rate neglect and inverse base rates. These are human categorization phenomena in which base rate information appears to be ignored. The naive Bayesian classifier accounts for a subset of the phenomena observed in base rate experiments. An extension to the model is examined that uses structure in the data sets resulting from features shared between categories.

Introduction

The base rate of a category is the probability of occurrence of an instance of that category. Humans appear to be sensitive to the base rates of categories in training and testing data sets. In some circumstances the more times a category appears, the more likely humans are to predict its occurrence. However, in other experimental settings, categories with smaller base rates appear to be preferred to categories with larger base rates.

Base rate neglect refers to a categorization phenomenon in which a feature that occurs proportionally in two categories appears to be associated with the less probable category (i.e., lower base rate). Human categorization performance suggests that the higher base rates of the more probable categories are being ignored. Gluck and Bower (1988) showed apparent base rate neglect in a medical categorization task. The participants were trained to predict a common and a rare disease given a symptom set of four symptoms s1 to s4.

The probability of the rare disease occurring was 0.25 and the probability of the common disease occurring was 0.75. The symptom probabilities given the rare disease were 0.69, 0.46, 0.35, and 0.23 for symptoms s1 to s4, respectively. The probabilities given the common disease are in the reverse order. Since the probability of s1 is 0.345, using Bayes formula the probability of the rare disease given s1 is 0.5. However, when asked to predict disease given a cue of s1, participants predicted the rare disease 0.67 of the time. Collectively, the pool of participants tend to over-estimate the probability of the rare disease given the symptom.

The inverse base rate phenomenon can be described as follows. Suppose one feature is only identified with a high base rate category and another feature is identified with a low base rate category. When a cue is given in which both features are together and the participant is asked to categorize the cue, the participant will tend to respond with the lower base rate category.

Base rate neglect and inverse base rate phenomena have apparently struck investigators as surprising because these phenomena seem counter to a tacit, appropriate decision procedure. Thus, this paper investigates the ability of one such procedure, the naive Bayesian classifier and extensions to fit experimental data by Kruschke (1996).

Bayesian Models

The naive Bayesian classifier is a popular machine learning technique, which often outperforms competing learning strategies Langley, Iba, & Thompson (1992). Bayesian classifiers have also been used to account for many categorization phenomena (Anderson, 1991). Thus, it seems natural to examine the ability of these classifiers to fit experimental data on base rate neglect and inverse base rate effects.

Naive Bayesian Model

The Bayesian model is a probabilistic classifier, which assigns a probability to an object's membership in each of a set of contrast categories. Assuming the categories partition the instance space, Bayes' theorem (Eq. 1) is used to assign the probability that an instance, represented as a feature vector, $F_{1..n}$, is a member of class C_i:

$$P(C_i|F_{1..n}) = \frac{P(C_i)P(F_{1..n}|C_i)}{\sum_k P(C_k)P(F_{1..n}|C_k)} \quad (1)$$

where $P(C_i)$ is the base rate of class C_i. In the naive Bayesian classifier, the features of an instance are assumed to be independent for each category, which gives the simplification expressed in Equation 2.

$$P(F_{1..n}|C_k) = \prod_j P(F_j|C_k) \quad (2)$$

Thus, the naive Bayesian classifier assigns probabilities using Equation 3.

$$P(C_i|F_{1..n}) = \frac{P(C_i)\prod_j P(F_j|C_i)}{\sum_k P(C_k)\prod_j P(F_j|C_k)} \quad (3)$$

The base rate phenomena discussed above pose a problem for a naive Bayesian model of human categorization. The model uses base rates in its calculation and in light of the inverse base rate effect an obvious modification would be to

remove the base rate term from the model. A difficulty with this approach is that base rates appear to be used for some of the cues in the test set. Removing base rates would cause a misfit for such cues. Also, some cues seem to use base rate in a more biased fashion than a naive Bayesian classifier. Modifying the naive Bayesian classifier to use even more base rate information has the difficulty of not accounting for the inverse base rate effect. These difficulties suggest the need for a model that finds a middle ground between ignoring and over using base rates.

Cue-Validity-Weighted Bayesian Model

A model that finds middle ground is the cue-validity-weighted Bayesian model. It uses the structure between categories to influence categorization. The cue validity measure, $P(C_k|F_i)$ is used to express the structure in the data set. It does this by relating categories that share features. When a shared feature occurs in a test instance, cue validity expresses the feature's relative weighting between the categories in which it occurs. When novel instances occur, this relative weighting may be used to aid in classification.

$$\frac{P(C_i)\prod_j P(C_i|F_j)P(F_j|C_i)}{\sum_k P(C_k)\prod_j P(C_k|F_j)P(F_j|C_k)} \quad (4)$$

The cue-validity-weighted Bayesian classifier (Eq. 4) weights features according to cue validity. This results in shared features biasing the classification in the direction of the category in which it occurred most frequently. If the feature value F_i occurs in the test instance and in the category k, then the probability of that feature given the class is multiplied by the cue validity. Otherwise, the cue-validity-weighted Bayesian classifier behaves like a naive Bayesian classifier. See discussions on cue validity in Rosch & Mervis (1975) and Hampton (1979).

There are different kinds of features being used by the classifier. Features that have a high cue validity are predictive features. If $P(C_k|F_i)=1.0$, then the feature is perfectly predictive. Features that have a high probability given the category are predictable features. If $P(F_i|C_k)=1.0$, then the feature is perfectly predictable. The weighting of the model has the effect of fully using perfectly predictive features and making shared features more strongly predict the category in which they occurred more frequently. The cue-validity-weighted Bayesian formula is a way of mathematically formalizing the impact of structure in the data set due to shared features.

Experiment 1

Kruschke's (1996) Experiment 1 consisted of two phases: a training phase and a testing phase. The participants were given eight training trials per block for fifteen blocks for a total of 120 trials. After each training trial, they were given accuracy feedback. The training data set, presented in Table 1, consisted of four disease categories (C_{ac}, C_{ar}, C_{bc}, and C_{br}) and six symptoms or features (i.e., f_{ac}, f_{ar}, f_{bc}, f_{br}, p_a, and p_b). Two of the disease categories (i.e., C_{ac} and C_{bc} have a higher frequency of occurrence than the other two categories (i.e., C_{ar} and C_{br}), thus the "c" versus "r" subscript for common and rare, respectively.

Each category can be predicted using one of four "perfect" features: f_{ac}, f_{ar}, f_{bc}, and f_{br}, respectively. For instance, skin-rash is the perfect common predictor, f_{ac}, for category, C_{ac} since $P(C_{ac}|f_{ac})=1.0$. Two of the features, p_a and p_b, are called "imperfect" predictor features. They are imperfect because they occur in two disease categories, but they still inform categorization (e.g., $P(C_{ac}|p_a) > P(C_{ac})$). The imperfect features can be used to distinguish the two category groups a and b. For example, ear-ache occurs in group a's rare and common disease categories, but Ear-ache does not occur at all in C_{bc} or C_{br}.

Table 1. Training stimuli used in Exp. 1 (Kruschke, 1996)

Dis.	Symptoms/Features	Freq.
C_{ac}	Skin-rash (f_{ac}) Ear-ache (p_a)	45
C_{ar}	Backpain (f_{ar}), Ear-ache (p_a)	15
C_{bc}	Sore-muscles (f_{bc}), Dizziness (p_b)	45
C_{br}	Stuff-nose (f_{br}), Dizziness (p_b)	15

During the testing phase participants were required to diagnose nine novel combinations of six symptoms. The nine symptom-combinations were repeated four times for a total of 36 test items. Table 2 presents the diseases chosen by the subjects for these novel combinations (i.e., cues).

Table 2. Observed choice proportion for Exp. 1

Example Symptom(s)	Cue	C_{ac}	C_{ar}	C_{bc}	C_{br}
Ear-ache	p_a	.75	.17	.05	.03
Skin-rash	f_{ac}	.93	.03	.03	.00
Backpain	f_{ar}	.04	.91	.02	.03
Skin-rash + Backpain	$f_{ac}+f_{ar}$	.35	.61	.02	.01
Ear-ache + Skin-rash + Backpain	$p_a+f_{ac}+f_{ar}$	.58	.40	.01	.00
Ear-ache + Sore-muscles	p_a+f_{bc}	.41	.08	.47	.05
Ear-ache + Stuffy-nose	p_a+f_{br}	.22	.09	.03	.67
Skin-rash + Stuffy-nose	$f_{ac}+f_{br}$	.35	.03	.06	.56
Ear-ache + Skin-rash + Stuffy-nose	$p_a+f_{ac}+f_{br}$	.72	.04	.04	.21

These results indicate that there are inverse base rate effects. Kruschke's (1996) inverse base rate effects replicate the phenomena reported by Medin & Edelson (1988). The inverse base rate effect is observed when the cue $f_{ac}+f_{ar}$ is presented and participants collectively favor the rare disease, C_{ar}, (61%) over the common disease, C_{ac}, (35%). This is the inverse of the 25% to 75% base rates of these rare and common categories.

The inverse base rate effect diminishes with the addition of the imperfect feature in the cue $p_a+f_{ac}+f_{ar}$. This cue is placed in common category, C_{ac}, 58% of the time and in rare category, C_{ar}, 40% of the time. The presence of the

imperfect feature makes the cue's choice proportion closer to the base rates of these categories, 75% and 25%, respectively.

The imperfect symptom (p_a) by itself gives choice proportions that are consistent with base rates, with it being placed in common category, C_{ac}, 75% of the time and in rare category, C_{ar}, 17% of the time.

In cues pitting the imperfect predictors against the perfect predictors, the rare perfect features influenced the choice more than common perfect predictors. For example, the rare disease, C_{br}, is predicted more often by the cue, $p_a + f_{br}$, instead of diseases C_{ac} *or* C_{ar}. An inverse base rate effect also occurs when the cue consists of a group "*a*" imperfect predictor, paired with a group "*b*" perfect predictor for a rare category, $p_a + f_{br}$. In this case, it is placed in common category, C_{ac}, 22% of the time and in rare category, C_{br}, 67% of the time. Hence, it seems that some cues are predicting diseases in inverse proportions of what base rates by themselves would predict, and other cues are predicting disease in proportion to the base rates.

Modeling Phenomena in Experiment 1

Bayesian models are "trained" from the 120 instances used to train subjects in Experiment 1. The instances are represented as a six dimensional vector of values along binary dimensions (i.e., "present" or "absent") corresponding to each feature. The probabilities for the features are approximated from frequencies of the features in the training data. The frequency of a feature given a category is one plus the number of times a feature occurs in a category over one plus the number of instances in a category. The addition of one to the numerator and the denominator is used to avoid a probability of zero if the feature has never occurred in the category.

Naive Bayesian Model

Overall, the naive Bayesian model offers a reasonable fit to the results from Experiment 1 (r^2 = 0.76, and root mean squared deviation (RMSD) = 0.16). Comparing Tables 2 and 3, it becomes evident that the naive Bayesian classifier fits a subset of the effects, and it is the performance involving the effects with imperfect features that the model fails to capture.

Table 3. Modeled choice proportion for Experiment 1

Model	Bayesian				Cue-Validity-Weight Bayesian			
Cue	C_{ac}	C_{ar}	C_{bc}	C_{br}	C_{ac}	C_{ar}	C_{bc}	C_{br}
p_a	.51	.49	.00	.00	.75	.24	.00	.00
f_{ac}	.99	.00	.00	.00	.99	.00	.00	.00
f_{ar}	.00	1.0	.00	.00	.00	1.0	.00	.00
$f_{ac}+f_{ar}$	.27	.73	.00	.00	.27	.73	.00	.00
$p_a+f_{ac}+f_{ar}$	.51	.49	.00	.00	.76	.24	.00	.00
p_a+f_{bc}	.21	.58	.21	.00	.31	.28	.41	.00
p_a+f_{br}	.15	.42	.00	.42	.18	.16	.00	.66
$f_{ac}+f_{br}$	.27	.00	.00	.73	.27	.00	.00	.73
$p_a+f_{ac}+f_{br}$	.99	.00	.00	.00	.99	.00	.00	.00

The naive Bayesian classifier performs in similar ways to the participants on the main effect of inverse base rates. The ambiguous cue $f_{ac} + f_{ar}$ is predicted to be associated with the rare disease. This occurs because there are two "mismatching" features for the common category: p_a which is missing and f_{ar} which never occurred in the common category. The rare category also has two mismatching features, but the category is smaller so the mismatches do not count as much as in the common category (Anderson, 1990). The following are numerators for the naive Bayesian classifier for the categories C_{ac} and C_{ar} (C_{bc} and C_{br} have small numerators due to four "mismatches" each):

Common Category

$$P(C_{ac})P(\neg p_a|C_{ac})P(f_{ac}|C_{ac})P(f_{ar}|C_{ac})P(\neg p_b|C_{ac})P(\neg f_{bc}|C_{ac})P(\neg f_{br}|C_{ac})$$
$$\left(\frac{45}{120}\right)\left(\frac{1}{46}\right)\left(\frac{46}{46}\right)\left(\frac{1}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right) \approx \frac{1}{120*46}$$

versus Rare Category

$$P(C_{ar})P(\neg p_a|C_{ar})P(f_{ac}|C_{ar})P(f_{ar}|C_{ar})P(\neg p_b|C_{ar})P(\neg f_{bc}|C_{ar})P(\neg f_{br}|C_{ar})$$
$$\left(\frac{15}{120}\right)\left(\frac{1}{16}\right)\left(\frac{1}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right) \approx \frac{1}{120*16}$$

The 45/120 is the base rate for the common category and the 15/120 is the base rate for the rare category. 1/46 represents a mismatch while 46/46 or 16/16 represents a match of a feature that occurs in every instance of a category. The difference in the numerators, due to the mismatches, causes the naive Bayesian classifier to choose the common, C_{ac}, and the rare, C_{ar}, 27% and 73%, respectively.

When the cue contains the imperfect feature $p_a + f_{ac} + f_{ar}$ the naive Bayesian classifier predicts the common disease and rare disease equally. The reason for this is that each category mismatches on only one feature, (i.e., f_{ar} for C_{ac} and f_{ac} for C_{ar}). With only one mismatch the numerators are equal. The computations for these two cues are similar to those that Anderson (1990) made in accounting for inverse base rates in Medin & Edelson's data set.

For the imperfect feature alone, the naive Bayesian model predicts the common disease and rare disease equally; although, humans predict in proportion to the base rates of the two diseases. This occurs in the model because both categories mismatch on only one cue (i.e., f_{ac} for C_{ac} and f_{ar} for C_{ar}). The match for the imperfect feature and the mismatch for the perfect feature for both categories gives the following numerators in the naive Bayesian classifier:

Common Category

$$P(C_{ac})P(p_a|C_{ac})P(\neg f_{ac}|C_{ac})P(\neg f_{ar}|C_{ac})P(\neg p_b|C_{ac})P(\neg f_{bc}|C_{ac})P(\neg f_{br}|C_{ac})$$
$$\left(\frac{45}{120}\right)\left(\frac{46}{46}\right)\left(\frac{1}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right) \approx \frac{1}{120}$$

versus Rare Category

$$P(C_{ar})P(p_a|C_{ar})P(\neg f_{ac}|C_{ar})P(\neg f_{ar}|C_{ar})P(\neg p_b|C_{ar})P(\neg f_{bc}|C_{ar})P(\neg f_{br}|C_{ar})$$
$$\left(\frac{15}{120}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{1}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right) \approx \frac{1}{120}$$

This causes the rare and common diseases to be predicted equally which is not what the participants predicted. They chose the common 75% of the time and the rare 17% of the time for the imperfect predictor alone.

The naive Bayesian classifier predicts cue $p_a + f_{bc}$ to be put into C_{ar} 58%, C_{ac} 21%, and C_{bc} 21%. This occurs because the cue matches on one feature and mismatches on

two features for each of these categories. The mismatches effect the common categories more than the rare categories. The rare category gets predicted more because of the mismatches. This effect is not observed in the participant's choice proportions. The participants choose C_{ar} 8%, C_{ac} 41%, and C_{bc} 47%.

For the cue $p_a + f_{br}$ the naive Bayesian classifier equally predicts the rare category, C_{ar}, 42%, the rare category, C_{br}, 42%, and the common category, C_{ac}, 15%. This occurs because the cue matches on one feature and mismatches on two feature on C_{ac}, C_{ar} and C_{br}. The mismatch effects the common category the most and splits the choice proportion between the two rare categories.

An obvious modification to model inverse base rates with a variant of the naive Bayesian model is to remove the contribution due to base rate. This model was examined and over compensates for the inverse base rate effects. It causes too many of the cues to exhibit inverse base rate behavior.

Cue-Validity-Weighted Bayesian Model

The cue-validity-weighted Bayesian model accounts for 93% of the variance in the data set (RMSD = 0.09). It behaves as a naive Bayesian classifier when there are no imperfect features in the cue. This can be observed in Table 3 for any cues that do not have imperfect features. Consistent with the inverse base rate effect and the naive Bayesian classifier, the cue-validity-weighted Bayesian classifier predicts the rare disease with higher probability for the ambiguous cue, $f_{ac} + f_{ar}$.

If an imperfect feature is in the cue, then for categories with that feature the cue validity is multiplied into the equation. This moves the choice proportion from the smaller categories to the larger categories. This predicts the common disease, C_{ac}, with 76% of the choice proportion and the rare disease, C_{ar}, with 24% for the cue $p_a + f_{ac} + f_{ar}$. The choice proportion shifts from equally predicting both in the naive Bayesian classifier to predicting the common disease because the imperfect predictor was associated 3 out of 4 times with the common disease. This result is more consistent with participant choice proportions than the naive Bayesian classifier.

Given the imperfect cue, p_a, by itself the cue-validity-weighted Bayesian model predicts the common disease, C_{ac}, 75% compared to the rare disease, C_{ar}, 25%. This occurs in much the same way as the above cue with the imperfect predictor. The 45/60 is the common category's cue validity for the imperfect predictor and 15/60 is the rare category's cue validity.

Common Category

$$P(C_{ac})P(C_{ac}|p_a)P(p_a|C_{ac})P(\neg f_{ac}|C_{ac})P(\neg f_{ar}|C_{ac})P(\neg p_b|C_{ac})P(\neg f_{bc}|C_{ac})P(\neg f_{br}|C_{ac})$$
$$\left(\frac{45}{120}\right)\left(\frac{45}{60}\right)\left(\frac{46}{46}\right)\left(\frac{1}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right) \approx \frac{3}{120*4}$$

versus Rare Category

$$P(C_{ar})P(C_{ar}|p_a)P(p_a|C_{ar})P(\neg f_{ac}|C_{ar})P(\neg f_{ar}|C_{ar})P(\neg p_b|C_{ar})P(\neg f_{bc}|C_{ar})P(\neg f_{br}|C_{ar})$$
$$\left(\frac{15}{120}\right)\left(\frac{15}{60}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{1}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right) \approx \frac{1}{120*4}$$

The cue-validity-weighted Bayesian model predicts for the cue $p_a + f_{bc}$ the categories C_{ar} 28%, C_{ac} 31%, and C_{bc} 41%. This occurs because the cue validity value associated with the imperfect feature shifts some of the choice proportion from C_{ar} to C_{ac}.

For the cue $p_a + f_{br}$ the cue-validity-weighted Bayesian model predicts the rare category, C_{ar}, and the rare category, C_{br}. The shift from the category C_{ar} related to the imperfect feature causing the choice proportion to move to the C_{br}.

In summary, both Bayesian models exhibit inverse base rate effects. The cue-validity-weighted Bayesian classifier is equivalent to the naive Bayesian classifier for cues without imperfect features (since $P(C|f)=1.0$ for perfect features f relative to class C). When cues have imperfect features, the cue-validity-weighted Bayesian classifier provides a better account of the data.

Experiment 3 (Kruschke, 1996)

Kruschke's Experiment 3 is similar to Experiment 1 except that one of the common diseases, C_{nc}, shares a rare disease's symptom (pf_n; Backpain). The symptom occurs fifteen times in both the rare, C_{nr}, and common, C_{nc}, diseases. This change in the training stimuli increases the complexity in the data set allowing for a larger number of novel cue items and the co-occurrence of the inverse and neglect effects resulting from the same training condition. The n subscript refers to the neglect condition. The diseases with the i subscript refer to the inverse condition.

Table 4. Training stimuli used in Exp. 3 (Kruschke, 1996)

Dis.	Symptoms/Features	Freq.
C_{nc}	Skin-rash (f_{nc}), Ear-ache (p_n)	30
	Skin-rash (f_{nc}), Backpain (pf_{nr}), Ear-ache (p_n)	15
C_{nr}	Backpain (pf_{nr}), Ear-ache (p_n)	15
C_{ic}	Sore-muscles (f_{ic}), Dizziness (p_i)	45
C_{ir}	Stuff-nose (f_{ir}), Dizziness (p_i)	15

Procedures were similar to Experiment 1, the participants were again given eight training trials per block for fifteen blocks for a total of 120 trials. In Experiment 3, the testing phase occurred after every five blocks instead of only at the end of the fifteen blocks. The novel test items are given in the table below as the symptom sets.

As in Experiment 1, the inverse base rate effect was found. In particular, the cue $f_{ic} + f_{ir}$ was categorized 32% of the time in the common disease, C_{ic}, and 64% of the time in the rare disease, C_{ir}.

In the neglect condition, Gluck & Bower's (1988) base rate neglect phenomena is replicated. The cue, pf_{nr}, is classified as the common disease, C_{nc}, 13% of the time and as the rare disease, C_{nr}, 77% of the time. This occurs even though the cue equally predicts both the common and rare disease (i.e., $P(C_{nr}|pf_{nr}) = P(C_{nc}|pf_{nr}) = 0.5$).

Although base rate neglect occurs in the neglect condition, the inverse base rate effect did not occur. The cue, $f_{nc} + pf_{nr}$, is categorized as the common disease, C_{nc}, 54% of the time and as the rare, C_{nr}, 40% of the time.

For the imperfect features alone, p_i and p_n, both conditions are similar to base rates. For the cue p_i the

category, C_{ic}, received 78% and C_{ir} received 13%. For the cue p_n the category C_{nc} received 64% and C_{nr} received 27% of the choice. For the cue $p_n + p_i$ the choice proportion was 36% for the neglect common and 17% for the neglect rare and 37% for the inverse common and 10% for the inverse rare.

For the imperfect feature paired with the common perfect feature of the other condition, for the cue $p_n + f_{ic}$, the category, C_{nc}, received 33% of the choice and C_{ic} receives 50% of the choice. For the cue $p_i + f_{nc}$, the category, C_{ic}, receives 35% of the choice and C_{nc} receives 54% of the choice.

For the imperfect feature paired with the rare perfect feature of the other condition, given the cue, $p_n + f_{ir}$, the category, C_{ir}, received 65% of the choice and the C_{nc} received 23% of the choice. For the cue, $p_i + pf_{nr}$, the category, C_{nr}, received 51% of the choice and C_{ic} received 27% of the choice. These cues exhibit the inverse base rate effect with the rare categories being preferred over the common categories.

Table 5. Observed choice proportion for Exp. 3.

Cue	C_{nc}	C_{nr}	C_{ic}	C_{ir}
p_n	.64	.27	.03	.07
p_i	.04	.06	.78	.13
$p_n + p_i$	.36	.17	.37	.10
f_{nc}	.83	.11	.02	.04
f_{ic}	.04	.03	.89	.04
$f_{nc} + f_{ic}$	.49	.05	.41	.05
pf_{nr}	.13	.77	.04	.05
f_{ir}	.01	.03	.03	.94
$pf_{nr} + f_{ir}$	.02	.29	.04	.65
$f_{nc} + pf_{nr}$	.54	.40	.02	.04
$f_{ic} + f_{ir}$	.01	.02	.32	.64
$p_n + f_{nc} + pf_{nr}$	.88	.11	.01	.01
$p_i + r_{ic} + r_{ir}$	.09	.02	.48	.41
$p_n + f_{ic}$	.33	.16	.50	.02
$p_i + f_{nc}$	.54	.05	.35	.07
$p_n + f_{ir}$	.23	.09	.03	.65
$p_i + pf_{nr}$	.08	.51	.27	.14
$f_{nc} + f_{ir}$	.29	.03	.01	.67
$pf_{nr} + f_{ic}$	.08	.46	.42	.04
$p_n + f_{nc} + f_{ir}$	.70	.04	.00	.27
$p_i + pf_{nr} + f_{ic}$	.14	.17	.64	.05

Modeling Phenomena in Experiment 3

Both Bayesian models use the probabilities of the features in the 120 training instances to model the data. In Experiment 3, the models are tested on twenty-one novel instances. The instances are represented in the same fashion as in Experiment 1.

Naive Bayesian Model

The naive Bayesian model for Experiment 3 again performs in similar ways to the participants in general and on the main effect of inverse base rates ($r^2 = 0.70$, RMSD = 0.18; compare Tables 5 & 6). For the same reasoning as in Experiment 1, the ambiguous cue, $f_{ic} + f_{ir}$, is predicted to be associated with the rare disease, C_{ir} 73% and 27% for the common disease.

The naive Bayesian classifier models the apparent base rate neglect effect for the cue pf_{nr}. The model predicts the rare category in the neglect condition with 99% of the choice proportion. This occurs because there are two mismatches and a partial match for the common category while there is only one mismatch for the rare category (see below numerators).

Common Category

$$P(C_{nc})P(\neg p_n|C_{nc})P(\neg f_{nc}|C_{nc})P(pf_{nr}|C_{nc})P(\neg p_i|C_{nc})P(\neg f_{ic}|C_{nc})P(\neg f_{ir}|C_{nc})$$
$$\left(\frac{45}{120}\right)\left(\frac{1}{46}\right)\left(\frac{1}{46}\right)\left(\frac{16}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right)\left(\frac{46}{46}\right) \approx \frac{1}{360*46}$$

versus Rare Category

$$P(C_{nr})P(\neg p_n|C_{nr})P(\neg f_{nc}|C_{nr})P(pf_{nr}|C_{nr})P(\neg p_i|C_{nr})P(\neg f_{ic}|C_{nr})P(\neg f_{ir}|C_{nr})$$
$$\left(\frac{15}{120}\right)\left(\frac{1}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right)\left(\frac{16}{16}\right) \approx \frac{1}{120}$$

For the imperfect features alone, p_i and p_n, the naive Bayesian model predicts the rare and common categories equally for the inverse condition. For the neglect condition the model predicts the rare category 59% and the common 41%. This is in the wrong direction in relation to the participants' choice proportions. For the cue $p_n + p_i$ the naive Bayesian model predicts the rare neglect category 38% and the rare inverse category 38%. This also is in the wrong direction.

For the imperfect features paired with the perfect features of the other condition, $p_n + f_{ic}$ and $p_i + f_{nc}$, the naive Bayesian classifier predicts the rare class of the imperfect feature condition. This occurs due to the same reasons expressed for Experiment 1.

The cue $p_n + f_{ir}$ predicts the rare neglect category 45% and the rare inverse category 45%. This is over-predicting the rare neglect category and under-predicting the rare inverse category. This is due to mismatched features on the common neglect category which gets 11% of the choice proportion. The cue $p_i + pf_{nr}$ predicts both rare categories equally each with 42% of the choice proportion. This is an over-prediction of the rare inverse category.

Cue-Validity-Weighted Bayesian Model

The cue-validity-weighted Bayesian model performs in a similar way to the participants on inverse base rates and base rate neglect (cf., Tables 5 & 6). The general fit of the model is $r^2 = 0.89$ and RMSD = 0.12. It predicts the same choice proportions as the naive Bayesian classifier when there are no imperfect features in the cue. It predicts the rare disease for the cue, $f_{ic} + f_{ir}$, which is consistent with the inverse base rate, and it predicts the apparent base rate neglect for the cue, pf_{nr}.

For the imperfect features alone, p_i and p_n, the cue-validity-weighted Bayesian model predicts the rare category 24% and common category 75%. This is similar to human performance which is 13% and 78% for the rare and common categories, respectively. For the neglect condition the model predicts the rare category 32% and the common 67%. This is consistent with human performance in the inverse condition which is 27% for the rare and 64% for the

common category. For the cue $p_n + p_i$ the cue-validity-weighted Bayesian model predicts inverse common 28% and the inverse rare 26%. This is more consistent with human performance than the naive Bayesian classifier. Humans choose the common 37% and the rare 10%. The common neglect category is chosen 19% by the cue-validity-weighted model and 26% for the rare neglect category. The human performance is 36% for the common category and 17% for the rare.

Table 6. Modeled choice proportion for Experiment 3.

Model	Bayesian				Cue-Validity-Weighted Bayesian			
Cue	C_{nc}	C_{nr}	C_{ic}	C_{ir}	C_{nc}	C_{nr}	C_{ic}	C_{ir}
p_n	.41	.59	.00	.00	.67	.32	.00	.00
p_i	.00	.00	.51	.49	.00	.00	.75	.24
$p_n + p_i$	.09	.38	.14	.38	.19	.26	.28	.26
f_{nc}	.99	.01	.00	.01	.99	.01	.00	.01
f_{ic}	.00	.00	.99	.00	.00	.00	.99	.00
$f_{nc}+ f_{ic}$	.40	.01	.59	.01	.40	.01	.59	.01
pf_{nr}	.01	.99	.00	.00	.01	.98	.00	.01
f_{ir}	.00	.00	.00	1.0	.00	.00	.00	1.0
$pf_{nr} + f_{ir}$	.00	.50	.00	.50	.00	.33	.00	.67
$f_{nc}+ pf_{nr}$	.85	.15	.00	.00	.85	.15	.00	.00
$f_{ic}+ f_{ir}$	.00	.00	.27	.73	.00	.00	.27	.73
$p_n+f_{nc}+pf_{nr}$	.94	.06	.00	.00	.98	.02	.00	.00
$p_i+r_{ic}+ r_{ir}$	.00	.00	.51	.49	.00	.00	.76	.24
$p_n+ f_{ic}$	.15	.62	.23	.00	.23	.31	.45	.00
$p_i+ f_{nc}$	.15	.00	.23	.62	.32	.01	.35	.32
$p_n+ f_{ir}$	.11	.45	.00	.45	.13	.17	.00	.70
$p_i+ pf_{nr}$	.00	.42	.15	.42	.00	.49	.27	.24
$f_{nc}+ f_{ir}$	.20	.00	.00	.80	.20	.00	.00	.80
$pf_{nr}+f_{ic}$	.00	.73	.27	.00	.00	.58	.42	.00
$p_n+f_{nc}+f_{ir}$	.99	.01	.00	.01	.99	.00	.00	.01
$p_i+pf_{nr}+f_{ic}$	.00	.00	.99	.00	.00	.00	1.0	.00

For the imperfect neglect feature paired with the perfect common inverse feature, $p_n + f_{ic}$, the cue-validity-weighted Bayesian model predicts the inverse common 45%, neglect common 23%, and the neglect rare 31%. This is a better fit than the naive Bayesian classifier. Humans' choice proportion is 50%, 33%, and 16% respectively.

The imperfect inverse feature paired with the perfect common neglect feature, $p_i + f_{nc}$, predicts the common neglect 32%, the common inverse 35%, and the rare inverse 32%. The human performance respectively is 54%, 35%, and 7%.

The cue-validity-weighted Bayesian model predicts for the cue, $p_n + f_{ir}$, the rare neglect category 13% and the rare inverse category 70%. This is consistent with human performance of 23% and 65% respectively. The cue $p_i + pf_{nr}$ predicts the rare neglect category 49%, the common inverse category 27%, and the rare inverse category 24%. This is consistent with human performance of 51%, 27%, and 14% respectively.

The effect of using cue validity to weight the naive Bayesian classifier better fits the inverse base rate and base rate neglect effects of Experiment 3 than the naive Bayesian classifier. If the cues do not have imperfect features, the model achieves similar results to the naive Bayesian classifier. When cues with imperfect features are presented, the cue-validity-weighted Bayesian model provides a better match of the human performance than the naive Bayesian classifier. Although Experiment 3 is more complex, the cue-validity-weighted Bayesian classifier is able to model many of the effects.

Conclusion

This paper has presented a cue-validity-weighted Bayesian account of the psychological phenomena inverse base rates and base rate neglect. The cue-validity-weighted feature Bayesian model accounted for 93% and 89% of the variance in the performance data of Kruschke's Experiment 1 and 3, respectively. The naive Bayesian model accounted for 75% and 70% of the variance, respectively. Kruschke modeled the phenomena with a connectionist model with five parameters. The model accounted for 99% of the variance in Experiment 1 and 97% of the variance in Experiment 3. It posited an early-late learning process to account for the phenomena. Without positing processes or parameterizing the model, the cue-validity-weighted feature Bayesian classifier modeled inverse base rate effects and the base rate neglect phenomena just as well as the naive Bayesian classifier and better for cues including imperfect features. The names of the phenomena "base rate neglect" and "inverse base rate" may be misnomers, and more investigation is required to evaluate the role of base rate when participants are faced with novel composite cues.

References

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

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98 (3), 409-429.

Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General, 117 (3)*, 227-247.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18(4), 441-461.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22 (1)*, 3-26.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. From Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose: AAAI Press.

Medin, D.L. & Edelson, S.M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117 (1)*, 68-85.

Rosch, E. & Mervis, C.B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.