Augmented Naïve Bayesian Model of Classification Learning

Lewis Frey (frey@Vuse.Vanderbilt.Edu)
Computer Science Department, Box 1679 Station B, Vanderbilt University
Nashville, TN 37232 USA
Douglas Fisher (dfisher@Vuse.Vanderbilt.Edu)
Computer Science Department, Box 1679 Station B, Vanderbilt University
Nashville, TN 37232 USA

Abstract

The Naïve Bayesian Classifier and an Augmented Naïve Bayesian Classifier are applied to human classification tasks. The Naïve Bayesian Classifier is augmented with feature construction using a Galois lattice. The best features, measured on their within- and between-category overlap, are added to the category's concept description. The results show that space efficient concept descriptions can predict much of the variance in the classification phenomena.

Introduction

The optimal Bayesian classifier chooses the most likely category given the evidence. Determining the most likely category in "real" environments is complicated due to inadequate evidence and dependencies between events. A way to simplify estimating the most likely category is to assume independence between events. This assumption is an oversimplification and can result in a sub-optimal classifier when dependencies exist.

The Naive Bayesian Classifier (NB), which assumes independence between events (e.g., features used to categorize), has a wide range of optimal behavior even when dependencies exist. Because it only uses the independent events to estimate the ranking of categories, it also has low memory requirements.

Assuming that humans approximate an optimal classifier but with severe memory constraints, this paper addresses the natural question of whether human categorization phenomena can be modeled with a NB.

When the Naive Bayesian Classifier supplies a non-optimal ranking of categories it needs to be augmented with a mechanism that gives an optimal ranking. An Augmented Naive Bayesian Classifier (ANB) that uses feature construction to find relevant conjunctions of features as well as singleton features is proposed as a way of better approximating the optimal classifier with memory constraints (i.e., not all feature conjunctions are used). This is similar in motivation to Anderson's (1991) rational model, except with a different method of finding relevant features and conjunctions.

The Augmented Naive Bayesian Classifier is compared to the Naive Bayesian Classifier on the task of modeling human performance on categorization tasks. The models' ability to provide an account of human categorization is examined. Optimal classifiers, dependence and independence are discussed formally below in relationship to the optimal Bayesian Classifier and the Naïve Bayesian Classifier. The human categorization phenomena to be modeled will be reviewed along with model fits. This is used as motivation for the augmentation of the Naive Bayesian Classifier. The paper ends with a discussion of the findings.

Optimal Classifier

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. Probability of C_i, $P(C_i)$, is the base rate of class C_i.

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

An optimal classifier is a classifier that minimizes the misclassification rate or zero-one loss. If $P(C_i|F_{1..n})$ is the probability that an instance represented by the feature vector $F_{1..n}$ is in class C_i, zero-one loss is minimized if, and only if, $F_{1..n}$ is assigned to the class C_k for which $P(C_k|F_{1..n})$ is maximum (Duda & Hart, 1973; Domingos & Pazzani, 1997). The Bayesian classifier, assigns $F_{1..n}$ to the class with maximal probability given the evidence. The Bayesian classifier does not assume independence between features, and consequently, the amount of information to determine the probabilities of classes becomes impractical as the number of features and categories grow.

In contrast, the Naive Bayesian classifier assumes independence between the features given a class: knowing the probability of one feature gives no information about another feature conditioned on class. That is,

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

Naive Bayesian Classifier

Substituting Equation 2 into Equation 1 yields the Naïve Bayesian Classifier (Equation 3).

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

By definition, the Naive Bayesian classifier is optimal when the independence assumption holds. Additionally it remains optimal when the independence assumption is violated, but the maximal class's relative rank remains the highest. Domingos and Pazzani (1997) show that it is optimal under a greater range of independence violations than previously assumed. Further support for the utility of NB is found in Langley, Iba, and Thompson (1992) and elsewhere; when compared to other machine learning techniques, NB worked well on natural domains. Bayesian approaches have also been used to model human categorization (Anderson, 1991; Frey & Fisher, 1998; Tenenbaum, 1999).

Data Sets & Model Fits

The phenomena of correlated features and learning trends are fit with the Naive Bayesian Classifier. Examining correlated features, the first data set comes from Pavel et al. (1988). This experiment examines human classification when there is competition between single and conjunctive features. The second data set from Nosofsky et al. (1994) replicates Shepard et al. (1961) examination of category complexity. Nosofsky et al. extends Shepard et al. by examining how these categories are learned over multiple trials. This is referred to as learning trend phenomena. This experiment explores different levels of category complexity and how quickly humans learn them.

Data Set 1: Comparison of singleton vs. conjunction features (Pavel et al. 1988)

Pavel et al. (1988) replicates Medin, Altom, Edelson and Freko (1982) who explored how singleton and conjoined features influence the classification of stimuli. The stimuli varied on four binary dimensions and are represented by a four digit number. Each digit corresponds to one of the dimensions' logical values of 1 or 2 (e.g., 2111). In Pavel's experiment the training stimuli are presented with accuracy feedback and then a transfer phase is given with both trained and novel instances. In this experiment, the category structure compares single dimension, dim 1 and dim 2, which have values that are *strongly* associated with each category, against conjoined dimensions, dim 3 and dim 4, which have values pairs that are *perfectly* correlated with each category. The measure used is the probability of choosing category A.

Results: Data Set 1

The NB does not exploit dependencies/correlations in data; whereas apparently human subjects were able to exploit the correlations found in this data (see Table 1, compare observed to NB predicted proportion). The NB accounts for only 18% of variance (SSD=1.7252, RMSD=0.328). There are four training stimuli (i.e., A2, A4, B1 & B3) that the NB does not categorize correctly due to dependencies and are not optimally classified by the Naive Bayesian Classifier. The proportion predicted for each of these stimuli falls directly on fifty-percent giving no preference for either category. This is inconsistent with human performance and suggests that human subjects find and exploit correlations between features. The NB choice proportion for transfer or test stimuli T1-T8 are also inconsistent with human performance.

Table 1: Predicted proportions for the Naive Bayesian Classifier (NB) and the Augmented Naive Bayesian Classifier (ANB) across stimuli (Stim) and categories (Cat). The overall observed proportions (Obs Prop) were obtained from Pavel et al. (1988).

Cat	Stim	Obs Prop	NB Predict	ANB Predict
A1	1111	0.990	0.800	0.99
A2	2111	0.980	0.500	0.99
A3	1122	0.990	0.800	0.99
A4	1222	0.950	0.500	0.99
B1	1212	0.010	0.500	0.01
B2	2212	0.010	0.200	0.01
B3	2121	0.010	0.500	0.01
B4	2221	0.000	0.200	0.01
T1	2222	0.580	0.200	0.50
T2	2211	0.560	0.200	0.50
T3	2122	0.710	0.500	0.50
T4	1211	0.700	0.500	0.50
T5	1112	0.460	0.800	0.50
T6	1121	0.450	0.800	0.50
T7	2112	0.400	0.500	0.50
T8	1221	0.260	0.500	0.50

Data Set 2: Learning Trends (Nosofsky et al. 1994)

Shepard, Hovland, and Jenkins (1961) studied the complexity of categorization tasks. The six tasks, presented in Table 2, consist of studying categories that could be distinguished by examining one dimension (task type I), two dimensions (type II), a dimension with exceptions (types III-V) and all dimensions (type VI). The point of such tasks is to determine what category types are easier to learn.

Each task consists of learning two categories given eight stimuli each with three binary dimensions. The stimuli are split into two four-stimulus categories. There are six different category structures, which are represented by the I-VI columns in Table 2.

Task-type I can be distinguished by one dimension (e.g., A: 1**, for category A the where the first dimension has a value of one and wildcards, asterisks, representing any value assignment). Type II is non-linearly separable and needs two dimensions to distinguish the categories (e.g., B: 11*). One dimension and a single exception can distinguish Task-types III-V. Type VI requires all three dimensions to determine the category. The participants are presented with a stimulus and their task is to put it in one of the two categories. After each trial, the participant

receives feedback. Performance is measured by the number of errors made for each category across all trials.

Table 2. The six category (A or B) assignments for task types I, II, III, IV, V & VI. There are eight stimuli with three binary dimensions. The dimensional values are logical 1 or 2.

		Task Type					
#	Stim	I	II	III	IV	V	VI
1	111	A	B	B	B	B	A
2	112	A	B	A	A	A	B
3	121	A	A	A	A	A	B
4	122	A	A	A	A	A	A
5	211	B	A	B	B	B	B
6	212	B	A	B	B	B	A
7	221	B	B	A	B	A	A
8	222	B	B	B	A	B	B

The one-dimensional Task-type I is the easiest to learn, followed by the non-linearly separable Task-type II, followed by tasks III-V. Type VI, which required all dimensions to determine membership, is the most difficult to learn.

Nosofsky, Gluck, Palmeri, McKinley and Glauthier (1994) use a large population of subjects and recorded error rates per block of training instead of only the total number of errors. This provides a representation of category learning over time. The participants were trained until they achieved 32 trials without error. If this level was not reached they maximally went through twenty-five blocks of sixteen trials (two repetitions of each stimulus), except in the first and second blocks which consisted of eight trials. Nosofsky et al.'s results are presented in Figure 1.

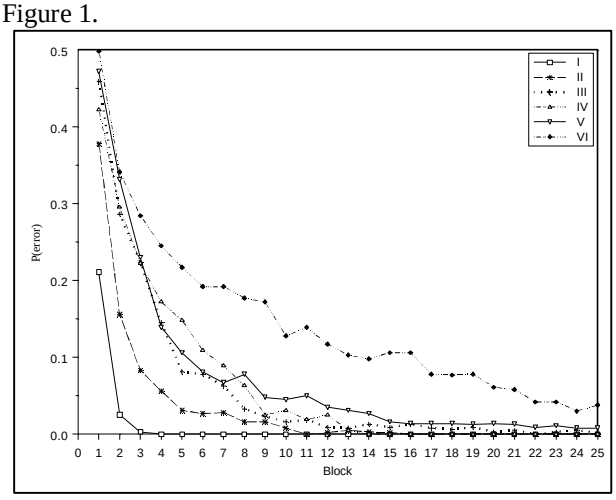

Figure 1. Observed proportion of errors across training blocks and Task-types I - VI (Nosofsky et al.'s, 1994).

Nosofsky et al.'s (1994) results are consistent with Shepard, Hovland, and Jenkins (1961). There is a main effect for type of task. Task-type I has the least errors. Type II has fewer errors than Types III-V. Type VI has the most errors.

Results: Data Set 2

The ordering of task difficulty predicted by the Naive Bayesian Classifier is inconsistent with human performance, and it accounts for only 7% of variance (R^2 = 0.07, SSD=13.139 RMSD=0.2959). Tasks I and IV the Naive Bayesian Classifier can distinguish the class of the stimuli and are easiest to learn. For tasks III and V are harder for it to learn and it predicts 50% for stimuli 1,2, 7 and 8. Tasks II and VI are the hardest for the Naive Bayesian Classifier with all stimuli predictions split at 50% between classes.

Discussion

When there are stimuli in the data set that are ranked sub-optimally by the Naive Bayesian Classifier then the fit to human performance is poor (18% of variance for data set 1). The fit degrades more when there are a larger number of non-optimal stimuli (7% of the variance in data set 2). While humans are memory-constrained, they appear to nonetheless find and exploit (at least simple) correlations/dependencies in data.

Augmented Naïve Bayesian Classifier

Since the Naïve Bayesian Classifier is space efficient and an optimal classifier under a wide range of conditions, it is used as a base classifier to model human categorization. But to account for categorization phenomena where NB does not find optimal classification but human subjects do, the NB model is augmented with feature construction. This results in a model with some similarities to Gluck, Bower and Hee's (1989) configural-cue network model except it is not restricted to pair-wise conjunctions. Instead a Galois Lattice is used to organize the space of conjunctive features, which are scored and selected for a space efficient concept representation. This augmentation gives the model the ability to find conjunctive features that allow stimuli that would be sub-optimally categorized by NB to be optimally categorized by the Augmented Naive Bayesian Classifier.

Feature Construction

The Galois lattice consists of nodes that are each represented by a feature description. With feature construction there can be conjunctive features which consist of more than one feature value assignment plus wildcards: asterisks that represent any feature value assignment. Figure 2 is a Galois lattice initialized with singleton features and the first two stimuli from Table 2 in the learning trend data.

Each node is linked as a *parent* to *lower boundary* nodes that are maximally general in their conjunctive feature description from the set of nodes that are more

specific than the given node (i.e., fewer wildcards). For example given node [*1*], then the set of more specific nodes is {11*, 111, 112}. The lower boundary is [11*] because it is the most general (i.e., most wildcards) from the specific set. Each node is also linked to as a *child* to *upper boundary* nodes that are maximally specific from the set of nodes that are more general than the given node. Given node [111], then the upper boundary is {11*, **1}. The Galois lattice is built incrementally using the algorithm in Godin and Missaoui (1994). A stimulus is intersected with existing nodes in the lattice, and any novel intersection is linked to its parents and children. This is similar to the algorithm used by Carpineto and Romano (1996). The Galois lattice supplies a partial order of the space of conjunctions.

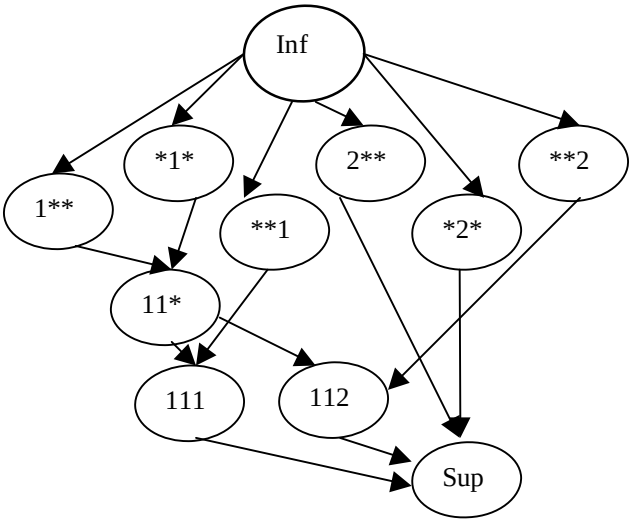

Figure 2: The Galois Lattice initialized with singleton features and stimuli 111 & 112 from table 2.

FScore is used to measure the utility of the features in the Galois lattice. It is a combination of three measures: within-category overlap, between-category overlap and specificity.

$$\text{FScore}(F \mid C_{\max}) = \max_{\forall i}\left[P(F \mid C_i) * \frac{Confirm(F \mid C_i)}{Infirm(F \mid C_i)} * \text{Cardinality}(F)^{specificity} \right] \quad (4)$$

The within-category overlap is the probability of the feature occurring within the category P(F|C). The between-category overlap is the ratio of the number of times the feature occurs in the category (Confirming evidence) over the number of times the feature occurs in other categories (Infirming evidence). Confirming and infirming evidence is similar to a measure used in Schlimmer's (1987) Stagger system. The bias chosen for the scoring is towards features that have high within-category overlap and low between-category overlap.

The *specificity* parameter allows an exploration of specific versus general features. Specific conjunctions have a larger cardinality due to fewer wildcard values (e.g., Cardinality(111*) = 3). With a high specificity setting, larger conjunctions are scored higher. With a large negative value, only singleton valued conjunctions (e.g., 1***) are scored high. If the parameter is set to zero, the cardinality of the conjunctions has no effect. The maximum value of the product of within-category, between-category and specificity is kept for each feature as FScore.

Equation 5 is the heuristic scoring mechanism for the conjunctive features in the lattice. The features are scored as a ratio relative to the highest scored feature. Features with a RScore above the threshold parameter are included in the concept. FScore is calculated via Equation 4.

$$RScore(F) = \frac{FScore(F \mid C_{\max})}{\max_{\forall j}\left(FScore(F_j \mid C_{\max}) \right)} > threshold \quad (5)$$

The *threshold* parameter is used to constrain the number of features of the lattice that are placed into the concept and participate in the choice proportion calculation. A RScore is assigned to the conjunctive features and all conjunctions above the threshold are added to the concept. If the threshold parameter is high then the concept has fewer conjunctions and if it is low then more conjunctions score above threshold and are a part of the concept.

In the Augmented Naïve Bayesian Classifier, the classes are scored by calculating the score for the features in the concept representation. The probability of F given C (i.e., $P(F_j|C_k)$, from Equation 3) has been replaced with $Score(F_j|C_k)$, Equation 6.

$$Score(F_j \mid C_k) = \left[\frac{2^{strength} * \left| F_j \cap C_k \right| + 1}{\left| C_k \right| + 1} \right] \quad (6)$$

When the *Strength* parameter is zero, Score approximates P(F|C) in the sample limit and Equation 7 is the Naïve Bayesian Classifier (Equation 3). Strength is used to scale the model's output with that of the human behavior so that a comparison can be made. It can be thought of as an index of how strongly the evidence is weighted. When strength is negative, the evidence is not strongly weighted. That is it will take more evidence to increase the distinction between categories for stimuli.

$$Score(C_i \mid F_{1..n}) = \frac{P(C_i) \prod_{j \in concept} Score(F_j \mid C_i)}{\sum_k P(C_k) \prod_{j \in concept} Score(F_j \mid C_k)} \quad (7)$$

The category scores for the stimuli's feature vectors are compared against the human categorization probabilities for the categories given the stimuli. Note that this collapses to its base Naïve Bayesian Classifier when specificity biases towards singleton features and the strength parameter is set to zero. When dependencies between features make the Naïve Bayesian Classifier sub-optimal conjunctive features are used to improve the model's fit. The parameters are chosen by the downhill simplex method in multi-dimensions (Press et al, 1994)

that minimizes one minus R^2. This results in parameters that account for the greatest amount of variance.

Results: Data Set 1

The ANB accounts for 92% of the variance (SSD=0.1676, RMSD=0.1023) in Table 1. The ANB provides a better prediction of human behavior than the NB, $F(3, 13) = 40.27$, $p<.05$.

Augmented Naive Bayesian Parameters. Specificity is 1.0. This allows a more specific search of the space. The *threshold* (0.75) compresses the representations using features with low inter-category overlap. For this data set, all the features only occur within one category. This is a case where inter-category overlap is very important for scoring the features. *Strength* (1.02) is positive and thus, the conjunctions are given strong weight.

The result shows that the competition is not strictly single dimensions 1 & 2 versus conjoined dimensions 3 & 4. Large conjunctions using combinations of three dimensions can be used to account for 92% of the variance. The features used by Augmented Naive Bayesian model for classifying the stimuli are 1*22 and *111 for category A and. *212 and 2*21 for category B. The representation used is also compact, using a smaller amount of space resources. The ANB uses four conjunctive features that correctly classify the training stimuli (in categories A & B, see Table 1) but do not occur in any of the transfer (T1-T8) stimuli. This results in all the transfer stimuli having a fifty-percent chance of being classified in category A.

Results: Data Set 2

In order to compare the ANB to the learning trends exhibited in Nosofsky et al. (1994) a mechanism for controlling the rate with which features are added to the concept is needed. The below equation is used to incrementally add features to the concept being learned.

$$TotalFConcept = \left(1-(trial)^{-rate}\right) * LatticeSize \quad (8)$$

As the number of trials increase a larger number of features are added to the concept. Features are no longer added when there are no more features above threshold. This allows a gradual addition of features to the concept and makes the model comparable for the Nosofsky et al. (1994) learning trend data.

The augmented Bayesian (Figure 3) has similar performance to human subjects (Figure 1) on the six task types. The R^2 is 0.92 and rmsd is 0.0401. The same parameter values are used across all six tasks. The fits are made based on the average probability of misclassifying any of the stimuli for a given trial block.

Augmented Naive Bayesian Parameters. Specificity is 0.94 and biases towards larger conjuncts, but still allows for singleton features as in task I. The *threshold* (0.76) is similar to that of Data Set 1 (0.75) and compresses the representations using features with low inter-category overlap (i.e., all features only occur within one category). This is a case where inter-category overlap is very important for scoring the features. *Strength* (1.22) is positive. Thus, the conjunctions are given more weight as in the fit for Data Set 1 (1.02). *Rate* (0.08) is used to compare the model with learning trend. The parameter is set to quickly learn features in first trials and then learn new features more slowly in later trials. This in addition to feature construction and compression accounts for the phenomena.

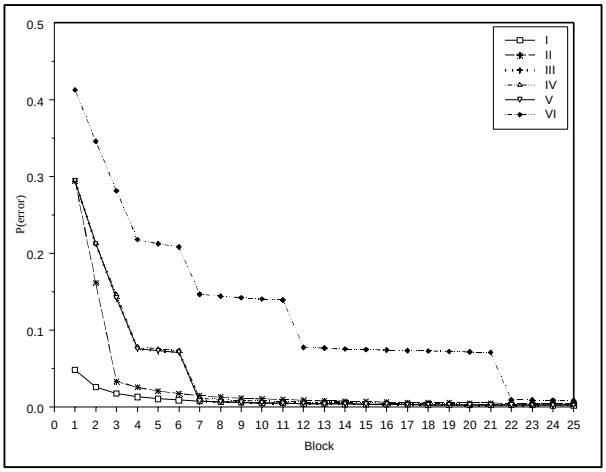

Figure 3. Proportion of errors across training blocks for the Augmented Naïve Bayesian.

The feature sets for the six task types learn the categories accurately. The model is consistent with the complexity of the tasks as measured by human error rates in the learning trend data. The ANB selects conjunctive features that occur exclusively in one category. For this human categorization behavior, the Augmented Naive Bayesian Classifier provides a better prediction than the Naive Bayesian Classifier, $F(4, 120) = 1967.6$, $p<.05$.

Discussion

The ANB gives two features for task I (A: 1**: B: 2**), four features for task II (A: 12*, 21*; B: 11*, 22), six features for task III (A: 12*, 1*2, *21; B: 21*, 2*2, *11), IV (A: 12*, *22, 1*2; B: 21*, *11, 2*1) and V (A: 12*, 1*2, *21; B: 21*, 2*2, *11) and eight features for task VI (all eight stimuli).

The number of features in each task is consistent with the ordering of difficulty for the Shepard et al. results. The ordering of difficulty in the experiment is task I followed by task II. Tasks III-V are the next most difficult and task VI is the most difficult. The number of features in the ANB concept representation also corresponds to the Boolean complexity demonstrated by Feldman (2000).

General Discussion

Because of independence violations, the NB is non-optimal in classifying stimuli across a number of tasks. This is inconsistent with human performance *in the domains studied here*. The ANB finds conjunctions that can classify these stimuli in the tasks. Thus, ANB as

compared to the NB provides a better account of human performance.

The Augmented Naive Bayesian parameters are effective at exploring possible classification representations. The ANB converges on representations that account for the phenomena. For data set 1 conjunctive features win over singleton features. The conjunctive features enabled certain stimuli, which are non-optimal for NB, to be learned in a consistent way to human performance. The compressed representation for data set 1 accounts for much of the variance. For data set 2 conjunctive features along with compressed representations for the tasks allow ANB to follow learning trend data patterns. The different tasks in this experiment have varying degrees of representation complexity (number of features) which maps onto complexity of the task as measured by human performance.

Limitations of ANB Classifier

There are classification tasks in which the Augmented Naive Bayesian model would tend to converge on an inappropriate representation. For example, the competition between singleton and conjunctive feature makes it more difficult to have representations that consist of both. Note that for learning trend data it did have both single and conjunctive representations for different tasks while using the same parameters. It is possible to have tasks that involve both singleton and conjunctions, but the model may include one over the other. With a more relaxed threshold the Augmented Naive Bayesian model places a large number of conjunctions in the representation. When there are a large number of conjunctions, the fit could be good because of each conjunction contributing to the fit. This would argue that the power of the representation is accounting for variance. This does not take place in these data sets because each representation is compressed due to the thresholds biasing towards small representations.

Conclusion

It is encouraging that the Augmented Naive Bayesian model selected simple consistent feature representations for these data sets. It helps to supply a parsimonious account via a smaller feature representation that can be tested experimentally. It also supplies novel views on the data due to the ordered approach of looking at conjunctions.

For the two data sets, the ANB provides a better prediction of human categorization phenomena than the NB. The NB does not model the behavior well when there are violations of the independence assumption in the category structures. The process of space efficient feature construction in the Augmented Naive Bayesian Classifier corrects the Naive Bayesian Classifier's incorrect ranking of the categories while at the same time better modeling human data.

References

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

Carpineto, C. & Romano, G. (1996). A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval. *Machine Learning* 24, 95-122.

Domingos, P. & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 29, 103-130, 1997.

Duda, R. & Hart, P. (1973). *Pattern classification and scene analysis*. New York: John Wiley & Sons.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. Nature, 407, 630-633.

Frey, L. & Fisher, D. (1998). Naive Bayesian Accounts of Base Rate Effects in Human Categorization. In the Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, 380-385.

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.

Godin, R. & Missaoui, R. (1994). An incremental concept formation approach for learning from databases. *Theoretical Computer Science*, 133, 387-419.

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., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 37-50.

Nosofsky, R. M., Gluck, M. A., Palmeri T. J., McKinley, S. C. & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition, 22 (3),* 352-369.

Palmeri, T. J. & Nosofsky, R. M. (1995). Recognition Memory for Exceptions to the Category Rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(3), 548-568.

Pavel, M., Gluck, M.A., & Henkle, V. (1988). Generalization by humans and multi-layer networks. *Proceedings of the 10th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Press, W. H, Tenkolsky, S. A., Vetterling, W. T. & Flannery, B.P. (1994). *Numerical Recipes in C*. Cambridge University Press.

Schlimmer, J. C. (1987). Concept Acquisition Through Representational Adjustment (Technical Report Number 87-19). Irvine CA: University of California, Department of Information and Computer Science.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classification. Psychological Monographs: General and Applied, 75 (13), 1-41.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. Advances in Neural Information Processing Systems, 11.