Midterm Report

Travis Bartley

***Abstract*—This paper provides an empirical demonstration of the optimality of discrete Bayes classifiers by measuring expected gain over idealized data sets. Uniformly random distributions of class-measure pairs are generated from a six-dimensional measurement space and fitted through iterative biasing towards improving classifier performance. This fitting is repeated over variations in class ranges and sampling sizes to examine the relationship between requisite variety and classifier performance. Classifier performance is considered generally invariant over all variations, with a slight trend to under-perform with an increasing range of class values.**

I. Introduction

Bayesian classification is a common methodology in which maximum *a posteriori* estimation is utilized for the creation of decision rules for classification tasks. Along with being trained in linear time [1] and conceptually intuitive, they hold prominence for their demonstrable property of optimality when utilized in tasks evaluated on metrics of expected value. Indeed, by definition the decision rule produced through Bayesian classification is the function that defines the upper bounds of expected value for a decision task, exceeding the cost-gain estimation of any other classifier. That is, for any Bayesian decision rule f:

$$E[e; f] \geq E[e; g] \tag{1}$$

where E is the expected gain estimation function, g is any arbitrary decision rule, and e is the economic cost matrix for the assignment of a class value to a measurement given the measurement's true class value.

However, the conceptual simplicity and ease of calculation that merit the classifier′s use can also be disadvantageous. Given a set of measurement values, the classifier function chooses the class assignment that maximizes the sum of all possible economic gain values for all possible true class assignments, weighted by the conditional probability of each true assignment given the measurement being classified. Such may be expressed as the function that chooses c^k such that:

$$\sum_{c^j \in C} e(c^j, c^k) P_T(c^j|d) \geq \sum_{c^j \in C} e(c^j, c^n) P_T(c^j|d) \tag{2}$$

where C is the set of all class values, d is a member of measurement space M, $P_T(c|d)$ is the probability of class c being the true classification of d given d in the measurement set, and $e_{i,j}$ the economic cost of e for the assignment of a measure to class j when it was of real class i. For derivation of $P_T(c|d)$, Bayes′ theorem is used:

$$P_T(c|d) = P(d|c)P(c)/P(d) \tag{3}$$

or, as $P(d)$ is typically canceled out in the derivation of expected gain:

$$P_T(c|d) \propto P(d|c)P(c) \tag{4}$$

While the calculation of $P(c)$ can be reliably estimated from a suitably robust measurement set of size proportionate to $K = |C|$, the observation of all possible measurement-class pairs for derivation of $P(d|c)$ requires a set-size proportionate to $K \cdot S$. For low probability pairs, this question of sample size becomes more pertinent, as a failure to witness a pair in a training set poses the issue of assigning a zero value to a non-zero $P_T(c|d)$ [1], limiting classifier accuracy. While this issue may be ignored for relatively simple classification task, orientation towards real-world applications poses a demand for sample sizes that are infeasible for many tasks (e.g. sentiment analysis of natural language) [2].

To address the likely poverty of training data, numerous methodologies are typically employed. Accommodating the simple necessity of non-zero probabilities, one may add a smoothing factor Δ to all conditional probability entries and normalize the distribution for each class [3]. Such allows for calculation of *a posteriori* distributions that observe measurement-value pairs that may have not been present in the training, though at the cost of skewing real distributions and removing any authentic zero probabilities from the data. Subspace classifier projection permits reduction of nonessential measurement features, reducing the measurement space and thus the minimum threshold for possible data accuracy. Given suitable knowledge of the measurement space, the conditional priors may also be approximated through a known probability distribution (e.g. Gaussian distribution for parametric variables, Bernoulli distribution for binomial measures). As well, assuming independence among features, a ′naive′ Bayes estimation of conditional values may also be used [1] [2], calculating the *a posteriori* of a set of measurements as a product of the independent conditional distribution of a measurement's features. That is, for measurement d possessing n measurements $(x_1, x_2, ..., x_n)$,

$$P_t(c|d) = P_t(c|x_1, x_2, ..., x_n) \propto P(c) \prod_{i=1}^{n} p(d|x_i) \tag{5}$$

Such reduces the need for training data to exhibit all measure-class pairs and instead simply need to be robust enough for observing all values for the individual features. As such, the minimum observation size would be proportionate to $K \cdot R$, with R being the cardinality of the largest feature space.

Core to the above is the trade-off of providing approximations of the *a posteriori* distributions instead of their true derivations. While such approximations seem to have marginal

effects on the "surprising" performance of Bayesian classifiers [1], they do nonetheless leave open the theoretical concern of lacking opportunities to observe exclusive performance of the classifier, particularly in the case of empirically validating ideal testing set for performance.

To address this, this paper will present experimental findings of several Bayesian classifiers over synthetic data sets specifically generated for ideal performance. Classifiers are constructed for a randomly generated five-dimensional measurement space, varying in size of the size of the training set and range of possible classes for estimation. After establishing a baseline performance for the classifier, the probability distributions use for generating the measurement space and classifier are updated to improve performance. With ideal measurement spaces provided for each classifier, performance is once again evaluated and compared across classifier groups. From this evaluation, it is demonstrated that discrete Bayesian classifiers are largely invariant to variation in measure sizes and only slightly disadvantaged by the increase of class range size, validating their use for supervised learning tasks.

II. Technical

A. The classifier

As detailed in (1-5) above, a Bayesian calssifier is the function that assigns measurement d class c_k fullfilling the expected gain equation:

$$\sum_{c^j \in C} e(c^j, c^k) P_T(c^j|d) \geq \sum_{c^j \in C} e(c^j, c^n) P_T(c^j|d)$$

where $P(c|d)$ is derived from the equation:

$$P_T(c|d) \propto P(d|c)P(c)$$

As such, the classifier requires three features: the distribution of class probabilities $P(c)$, the distribution of class-conditional probabilities $P(c|d)$, and an economic gain matrix of KxK, where each entry e_{ij} refers to the economic gain/cost for assigning class j to a measure of real value i.

For initializing the classifier, a matrix of class conditional values was generated as follows: given the set of classes C, cardinality $K = |C|$, measurement space $M = \{d_1, d_2, ..., d_N\}$ where $d_n = (x_1, x_2, ..., x_N)$ and cardinality $S = |M|$, K arrays of length S were generated. For each value of each array, a uniform number was generated between $[0, 1]$. After each array A_c was filled, all values were normalized so $\sum_{i=1}^{S} a_i = 1$, with a_i being the ith element of A_c. These K matrices were then treated as tables of class conditional probability values, with $P(c_i|d_j)$ being expressed by the jth element in array A_i.

Prior probabilities were generated from the training set, calculating the proportion of each class c'_ks occurrence within a given training set. Meanwhile, economic gain matrices were generated as detailed in the following section.

B. The economic gain matrix

Prior to testing, two economic $K \times K$ gain matrices are generated for a classifier. The first is the identity matrix, a collection of values that assign zero to any incorrect classification and 1 to a correct classification. One notes that the identity matrix alters the classifier function to one maximizing a measure′s conditional *a posteriori*, as:

$$\sum_{c^j \in C} e(c^j, c^k) P_T(c^j|d) \geq \sum_{c^j \in C} e(c^j, c^n) P_T(c^j|d)$$

$$1 \cdot P_T(c^k|d) + 0 \cdot (K-1) \geq 1 \cdot P_T(c^n|d) + 0 \cdot (K-1)$$

$$P_T(c^k|d) \geq P_T(c^n|d) \tag{6}$$

The other matrix used was a penalty matrix that shared reward values for correct identification but substituted a -1 cost for any incorrect classification. Such was chosen for two purposes: the first was allowing evaluation of a punitive metric for evaluation and allow evaluation of differences in philosophies of punitive and reward based classification. The second was for validation of the classifier′s performance. By the expected gain calculation:

$$E(e) = \sum_{i=1}^{K} \sum_{j=1}^{K} e(c^i, c^j) P_{TA}(c^i, c^j) \tag{7}$$

one notes that the identity matrix produces the sum:

$$E(e_{identity}) = \sum_{j=1}^{K} P_{TA}(c^j, c^j) \tag{8}$$

as all entries $e(c^i, c^J) i \neq j$ are 0, producing the accuracy metric for the classifier. Meanwhile, the penalty matrix evaluates to:

$$E(e_{penalty}) = \sum_{i=1}^{K} \sum_{j=1}^{K} e(c^i, c^j) P_{TA}(c^i, c^j)$$

$$E(e_{penalty}) = -\sum_{i=1}^{K} \sum_{j=1}^{K} P_{TA}(c^i, c^j), c^i \neq c^j \tag{9}$$

which is the negative probability of incorrect classification. That is, is is the negative complement of $E(e_i dentity)$ such that

$$E(e_{identity}) - E(e_{penalty}) = 1 \tag{10}$$

Using both the identity and punitive economic gains thus allows validation of accurate performance of the evaluation metric across randomized trials as their difference should sum to $|1|$ (ignoring the possibility of floating point error).

For reference, matrix representations of each form are provided as follows:

$$e_{identity} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{11}$$

$$e_{penalty} = \begin{pmatrix} 0 & -1 & \cdots & -1 \\ -1 & 0 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & 0 \end{pmatrix} \tag{12}$$

C. Measurements

For initial data, measurements were generated at random through a uniform distribution. S values from $[0, 1]$ were generated and normalized so for generated values $\{s_1, s_2, ...s_S\}, \sum_{i=1}^{S} s_i = 1$. Values were then converted into cumulative probability distribution $\{p_1, p_2, ..., p_S\}$ to represent the distribution of all measurements in space M.

A similar procedure was conducted for generating classes for measures. K random samples were taken from a uniform distribution $[0, 1]$. These samples $\{k_1, k_2, ..., k_K\}$ were normalized so $\sum_{j=1}^{K} k_j = 1$ and then transformed into a cumulative probability distribution $\{q_1, q_2, ..., q_K\}$. An array of length S was then generated with S samples from the same uniform distribution. For each sample s, index c was chosen so as to fulfill the inequality:

$$q_{c-1} \leq s \leq q_c \tag{13}$$

As the cumulative probability distribution organizes K elements as $\{q_1 \leq q_2 \leq .... \leq q_K\}$, this index maps each s to a class value mimicking the synthetic probability distribution, effectively creating a mapping function $\{h : M- > C\}$.

These distributions provide for random generation of samples that conform to a synthetic probability distributions of $d \in M$ and $c \in C$. Z samples are taken from the uniform distribution $[0, 1]$ and then assigned an s such that for sample z:

$$p_{s-1} \leq z \leq p_s \tag{14}$$

As with class assignment, the cumulative probability distribution permits the use of d as an index for measurement s_d in the array of S values so that z maps to measurement d of M. This measurement is then mapped to the corresponding c from the class assignment distribution, providing randomly sampled measurement-value pairs.

D. Testing Protocol

1) Synthetic Measures: For each test, Z measure-value pairs are generated from a 5-dimensional space M. Then, K conditional probability distributions of length $S = |M|$ are generated, along with the two $K \times K$ economic gain matrices detailed above.

2) V-fold Testing: Measurements were then shuffled and partitioned for V-folds evaluation with V-factor 10 (chosen empirically). The classifier was trained on 9 partitions of the measurements to estimate $P(c)$ and then tests were performed on the 10th partition. This was repeated for each combination of $\binom{10}{9}$.

3) Evaluation: For each round of V-fold testing, a $K \times K$ confusion matrix Q was produced, with q_{ij} representing the counts of class assignment j for real class value i. After all rounds of V-fold validation were completed, all matrices were summed and then normalized so $\sum_{i=1}^{K} \sum_{j=1}^{K} q_{ij} = 1$, producing probability values for the assignment of class j to measurements of real class i, $P_{TA}(c^i, c^j)$. With the provided economic gain matrix e, expected gains were calculated by:

$$E(e) = \sum_{i=1}^{K} \sum_{j=1}^{K} e(c^i, c^j) P_{TA}(c^i, c^j)$$

providing the evaluation metric for each classification.

4) Fitting: To fit data to improve classifier performance, a two-fold technique was used. After each set of V-fold testing, the distribution of class conditional values $P(d|c)$ were increased by $\Delta = .005$ (value empirically determined) for each measurement-class pair seen in the data. Then, each set of conditional values were normalized so $\sum_{d \in M} P(d|c) = 1$ for each class. Doing so, the classifier biased itself towards pairs observed in the data-set while reducing preference for unobserved pairs, better fitting it for the data.

This updated class-conditional distribution was then used to regenerate measurement-class pairs. The probability distribution of measurement $d \in M$ was recalculated as:

$$P(d) = \sum_{c \in C} P(d|c) P(c). \tag{15}$$

with $P(c)$ generated from the entire set of measures-value pairs first generated and held constant over all rounds of fitting. These measurements were then converted into a cumulative probability distribution and used to generate samples as detailed above.

For class assignment, Bayes$'$ rule was used to calculate $P(c|d)$. For each d, a cumulative probability distribution of $P(c|d)$ across all classes $c \in C$ was generated. A random sample was taken from uniform distribution $[0, 1]$ and then a class index was calculated by using the same inequality from the generation procedure, providing a new set of class assignments for all d.

With the new probability distributions calculated based off the conditional probability used for the generation of the new classifier, Z measurement-class pairs are generated as before, followed by another iteration of V-fold testing for evaluation.

III. Experiments

A. Measure Space

For experiments, all evaluation was performed for 5-dimensional measurement space M, with each dimension having 10 potential measurement values (determined empirically for ease of testing), providing $|M| = 10^5 = 100000$ measurement values. Each tuple of measurement values $(x_1, x_2, ..., x_5)$ was converted to a linear representation L through row conversion formula:

$$L = \sum_{k=1}^{d} (\prod_{q=k+1}^{5} N_l) n_k \tag{16}$$

with n_k being the kth feature of measure d and N_k being the cardinality of kth dimension of M. One notes that this conversion of 5 feature tuples to single linear indexes reduces the dimensionality of M to be a question of quantity instead of complexity, allowing loss of generality of experimental findings for other dimensionalities of M.

B. Variants

To observe the relationship between class complexity, sample size, and choice of economic gain matrix on classifier accuracy, experiments were ran over combinations of the following factors:

- Class ranges: $K = 2, 4, 6$.
- Sample sizes: $Z = S, SK, 2SK$
- Economic Gain Matrix: Identity, Penalty

Class range values were chosen to examine variation as class complexity was increased from a binary classification task. Sample range values were chosen to examine effects of minimum sample sizes for viewing all measurements, $Z = S$ and all measurement-class pairs, $Z = SK$. To provide a more realistic chance of obtaining a representative sample, a larger sample size, $Z = 2SK$ was also examined.

The fitting factor Δ was .005, added over 10 iterations. Increase in expected value was charted over each iteration.

Each experiment is produced by a unique seed for repeat evaluation. The seed for each experiment is listed with its resultant set of data.

IV. Results

For all tests, the difference of expected gain between the identity and penalty matrices summed to 1(within tolerance of floating point error), verifying performance. As such, values presented will be for experiments conducted on the identity matrix.

A. Initial Results

Expected gain values before and after fitting are displayed in Table 1 As expected, all fittings significantly increased

TABLE I
EXPECTED GAIN BEFORE AND AFTER FITTING

Classifier Parameters	Expected Gain	
	Before	***After***
$K = 2, Z = S, seed = 2580$	0.86318	0.99981
$K = 2, Z = KS, seed = 2590$	0.51288	0.99990
$K = 2, Z = 2KS, seed = 2600$	0.51840	0.99986
$K = 4, Z = S, seed = 2550$	0.37808	0.99924
$K = 4, Z = KS, seed = 2560$	0.41811	0.99948
$K = 4, Z = 2KS, seed = 2570$	0.43520	0.99964
$K = 6, Z = S, seed = 2545$	0.25873	0.99922
$K = 6, Z = KS, seed = 2535$	0.20309	0.99926
$K = 6, Z = S, seed = 2525$	0.26554	0.99959

[a]All values have been rounded to five significant digits.

classifier performance. Further, it is observed that, excluding the group $K = 2$, all classifiers performed significantly above chance for randomized samples. In the case of $K = 2$, all classifiers performed either slightly or significantly above chance. For the case of $Z = K$, the proportion above chance was $(0.51228 - 0.5)/0.5 = 0.02456$ while for $Z = 2KS$ was $(0.51840 - 0.50)/0.5 = 0.0368$. While a 2% difference may be considered slightly significant, the high variation of randomness in the synthetic data suggested that this difference should be judged as chance.

For classifier $K = 2, Z = S, seed = 2580$, the high accuracy was concerning, prompting tests with other seeds. For seeds 1,2, 3, and 4, expected values were generated of 0.50032, 0.4955, and 0.68852, 0.72285 respectively. This high variation in expected gain suggests that while performance above chance is likely, it is unreliable and should not be taken as a general property of Bayesian classification.

A slight trend can be observed with the increase of expected gain with sampling size for all classifiers excluding group $K = 2$. However, given the largest difference, $(0.99959 - 0.99926)/0.99926 = 0.0003$, is not significant, this trend may be ignored.

Likewise, a decrease in fitted performance is observed increasing complexity of classes. With sample sizes equal, the increase in K appears proportional to a decrease in fitting accuracy. However, these differences do not appear stable and even the largest ratio of decrease $(0.99981 - 0.99924)/0.99981 = 0.00057$ is barely significant for the amount of randomness core to the testing. However, the trend is noted for later discussion.

B. Fitting

The increase in expected gain for each iteration of fitting is charted as below, with same proportions of Z charted together for comparison. The graphs visualize a few trends. Fitting

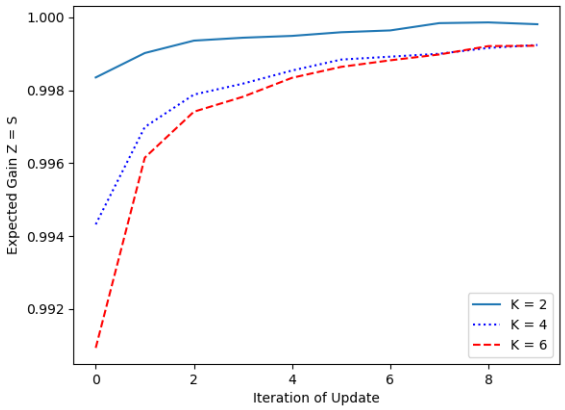

Fig. 1. Iterative Improvements for Z=S

has greater efficacy for less complex measurement spaces regardless of sample size, with $K = 2$ having higher gain for all groups and $K = 6$ having lowest gain. This supports the earlier trend noting the increase in class size being correlated with decrease in expected gain.

Also worth noting is the tendency of classifiers for $K = 4$ and $K = 6$ to maintain closer proximity to each other than

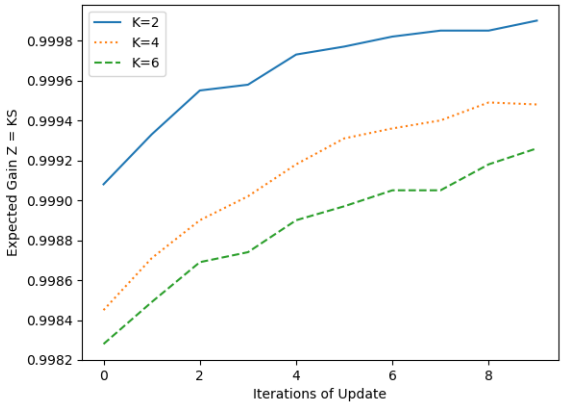

Fig. 2. Iterative Improvements for Z=KS

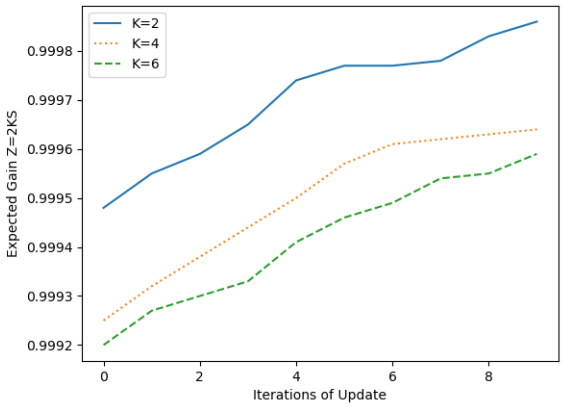

Fig. 3. Iterative Improvements for Z=2KS

those of $K = 2$. This suggests that increasing class complexity and decreasing expected gain are not in a linear relationship and have a lower rate with each additional class.

V. Discussion and Conclusion

A. Discussion

The baseline performance for the classifiers provides a rather apt demonstration of the robustness of Bayesian classification. Despite likely erroneous conditional values $P(d|c)$, the estimation of class probability distributions provides the classifier enough robustness to perform above chance - provided a suitably representative training sample. While this observation may be banal (obviously making classification decisions informed by likely distribution allows performance beyond chance), it deserves mention if only to reflect on how such features allow Bayesian classifiers to mitigate possible limitations regarding any poverty of class-conditional observations.

That classification over all fitted samples achieved near equal (within three significant digits) expected gains suggests the classifier has relatively little dependence on sample sizes given an idealized testing environment. In such a case, Bayesian classifiers with well-intuited class-conditional distributions may be trusted to perform well in scenarios where samples sizes are sparse or expensive to acquire.

The exhibited tendency for increasing class ranges to slightly limit performance is of debatable significance. On one hand, the increasing complexity of a measurement space logically will limit performance of any classifier, Bayesian or otherwise, simply due to the increase of possible means of classifications that can emerge from the data. However, that this relationship was not linear in the fitted samples and difference in expected gain measurements appeared to greatly reduce among larger ranges suggests a possible lower-bound for the effect. Future empirical observations of several class ranges are required to determine if this suggestion is valid.

B. Conclusion

Bayesian classification is an incredibly robust methodology of predictive classification that can be generally invariant over training size and range of class values in an idealized environment. Though generally simple in comparison to other methodologies, this simplicity appears to be crucial in their performance and ability to mitigate even disadvantageous initialization environments. Provided idealized environments, they can perform at near perfect accuracy, providing an invaluable tool for machine learning applications.

References

[1] K. M. Al-Aidaroos, A. A. Bakar and Z. Othman, "Nave bayes variants in classification learning," 2010 International Conference on Information Retrieval and Knowledge Management (CAMP), Shah Alam, Selangor, 2010, pp. 276-281.

[2] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Lingusitics, and Speech Recognition*. 3rd ed., Draft. 2019.

[3] A. Smola and S. V. N. Vishwanathan, *Introduction to Machine Learning*. Cambridge, U.K: Cambridge University Press, 2008.